GENERATIVE MODEL-ENHANCED HUMAN MOTION PREDICTION
Anthony Bourached,Ryan-Rhys Griffiths,Robert Gray,Ashwani Jha,Parashkev Nachev
The task of predicting human motion is complicated by the natural heterogeneity and compositionality of actions, necessitating robustness to distributional shifts as far as out-of-distribution (OoD). Here we formulate a new OoD benchmark based on the Human3.6M and CMU motion capture datasets, and introduce a hy- brid framework for hardening discriminative architectures to OoD failure by aug- menting them with a generative model. When applied to current state-of-the-art discriminative models, we show that the proposed approach improves OoD ro- bustness without sacrificing in-distribution performance, and can theoretically facilitate model interpretability. We suggest human motion predictors ought to be constructed with OoD challenges in mind, and provide an extensible general framework for hard- ening diverse discriminative architectures to extreme distributional shift.

****

Net-DNF: Effective Deep Modeling of Tabular Data
Liran Katzir,Gal Elidan,Ran El-Yaniv
A challenging open question in deep learning is how to handle tabular data. Unlike domains such as image and natural language processing, where deep architectures prevail, there is still no widely accepted neural architecture that dominates tabular data. As a step toward bridging this gap, we present Net-DNF a novel generic architecture whose inductive bias elicits models whose structure corresponds to logical Boolean formulas in disjunctive normal form (DNF) over affine soft-threshold decision terms. Net-DNFs also promote localized decisions that are taken over small subsets of the features. We present an extensive experiments showing that Net-DNFs significantly and consistently outperform fully connected networks over tabular data. With relatively few hyperparameters, Net-DNFs open the door to practical end-to-end handling of tabular data using neural networks. We present ablation studies, which justify the design choices of Net-DNF including the inductive bias elements, namely, Boolean formulation, locality, and feature selection.

****

Predicting Inductive Biases of Pre-Trained Models
Charles Lovering,Rohan Jha,Tal Linzen,Ellie Pavlick
Most current NLP systems are based on a pre-train-then-fine-tune paradigm, in which a large neural network is first trained in a self-supervised way designed to encourage the network to extract broadly-useful linguistic features, and then fine-tuned for a specific task of interest. Recent work attempts to understand why this recipe works and explain when it fails. Currently, such analyses have produced two sets of apparently-contradictory results. Work that analyzes the representations that result from pre-training (via "probing classifiers") finds evidence that rich features of linguistic structure can be decoded with high accuracy, but work that analyzes model behavior after fine-tuning (via "challenge sets") indicates that decisions are often not based on such structure but rather on spurious heuristics specific to the training set. In this work, we test the hypothesis that the extent to which a feature influences a model's decisions can be predicted using a combination of two factors: The feature's "extractability" after pre-training (measured using information-theoretic probing techniques), and the "evidence" available during fine-tuning (defined as the feature's co-occurrence rate with the label). In experiments with both synthetic and natural language data, we find strong evidence (statistically significant correlations) supporting this hypothesis.

****

FMix: Enhancing Mixed Sample Data Augmentation
Ethan Harris,Antonia Marcu,Matthew Painter,Mahesan Niranjan,Adam Prugel-Bennett, Jonathon Hare
Mixed Sample Data Augmentation (MSDA) has received increasing attention in recent years, with many successful variants such as MixUp and CutMix. We analyse MSDA from an information theoretic perspective, characterising learned models in ter

ms of how they impact the models' perception of the data. Ultimately, our analyses allow us to decouple two complementary properties of augmentations that are useful for reasoning about MSDA. From insight on the efficacy of CutMix in particular, we subsequently propose FMix, an MSDA that uses binary masks obtained by applying a threshold to low frequency images sampled from Fourier space. FMix improves performance over MixUp and CutMix for a number of models across a range of data sets and problem settings, obtaining new state-of-the-art results on CIFAR-10 and Fashion-MNIST.
****

A Theory of Self-Supervised Framework for Few-Shot Learning
Zhong Cao,Jiang Lu,Jian Liang,Changshui Zhang
Recently, self-supervised learning (SSL) algorithms have been applied to Few-shot learning(FSL). FSL aims at distilling transferable knowledge on existing classes with large-scale labeled data to cope with novel classes for which only a few labeled data are available. Due to the limited number of novel classes, the initial embedding network becomes an essential component and can largely affect the performance in practice. But almost no one analyzes why a pre-trained embedding network with self-supervised training can provide representation for downstream FSL tasks in theory. In this paper, we first summarized the supervised FSL methods and explained why SSL is suitable for FSL. Then we further analyzed the main difference between supervised training and self-supervised training on FSL and obtained the bound for the gap between self-supervised loss and supervised loss. Finally, we proposed potential ways to improve the test accuracy under the setting of self-supervised FSL.
****

Optimism in Reinforcement Learning with Generalized Linear Function Approximation
Yining Wang,Ruosong Wang,Simon Shaolei Du,Akshay Krishnamurthy
We design a new provably efficient algorithm for episodic reinforcement learning with generalized linear function approximation. We analyze the algorithm under a new expressivity assumption that we call ``optimistic closure,'' which is strictly weaker than assumptions from prior analyses for the linear setting. With optimistic closure, we prove that our algorithm enjoys a regret bound of $\widetilde{O}\left(H\sqrt{d^3 T}\right)$ where H is the horizon, d is the dimensionality of the state-action features and T is the number of episodes. This is the first statistically and computationally efficient algorithm for reinforcement learning with generalized linear functions.
****

Empirical Frequentist Coverage of Deep Learning Uncertainty Quantification Procedures
Benjamin Kompa,Jasper Snoek,Andrew Beam
Uncertainty quantification for complex deep learning models is increasingly important as these techniques see growing use in high-stakes, real-world settings. Currently, the quality of a model's uncertainty is evaluated using point-prediction metrics such as negative log-likelihood or the Brier score on heldout data. In this study, we provide the first large scale evaluation of the empirical frequentist coverage properties of well known uncertainty quantification techniques on a suite of regression and classification tasks. We find that, in general, some methods do achieve desirable coverage properties on \emph{in distribution} samples, but that coverage is not maintained on out-of-distribution data. Our results demonstrate the failings of current uncertainty quantification techniques as dataset shift increases and establish coverage as an important metric in developing models for real-world applications.
****

Deep Reinforcement Learning For Wireless Scheduling with Multiclass Services
Apostolos Avranas,Marios Kountouris,Philippe Ciblat
In this paper, we investigate the problem of scheduling and resource allocation over a time varying set of clients with heterogeneous demands. This problem appears when service providers need to serve traffic generated by users with different classes of requirements. We thus have to allocate bandwidth resources over ti

me to efficiently satisfy these demands within a limited time horizon. This is a highly intricate problem and solutions may involve tools stemming from diverse fields like combinatorics and optimization. Recent work has successfully proposed Deep Reinforcement Learning (DRL) solutions, although not yet for heterogeneous user traffic. We propose a deep deterministic policy gradient algorithm combining state of the art techniques, namely Distributional RL and Deep Sets, to train a model for heterogeneous traffic scheduling. We test on diverse number scenarios with different time dependence dynamics, users' requirements, and resources available, demonstrating consistent results. We evaluate the algorithm on a wireless communication setting and show significant gains against state-of-the-art conventional algorithms from combinatorics and optimization (e.g. Knapsack, Integer Linear Programming, Frank-Wolfe).

****

Neural Lyapunov Model Predictive Control

Mayank Mittal,Marco Gallieri,Alessio Quaglino,Seyed Sina Mirrazavi Salehian,Jan Koutnik

With a growing interest in data-driven control techniques, Model Predictive Control (MPC) provides a significant opportunity to exploit the surplus of data reliably, particularly while taking safety and stability into account. In this paper, we aim to infer the terminal cost of an MPC controller from transitions generated by an initial \emph{unknown} demonstrator. We propose an algorithm to alternatively learn the terminal cost and update the MPC parameters according to a stability metric. We design the terminal cost as a Lyapunov function neural network and theoretically show that, under limited approximation error, our proposed approach guarantees that the size of the stability region (region of attraction) is greater than or equal to the one from the initial demonstrator. We also present theorems that characterize the stability and performance of the learned MPC in the presence of model uncertainties and sub-optimality due to function approximation. Empirically, we demonstrate the efficacy of the proposed algorithm on non-linear continuous control tasks with soft constraints. Our results show that the proposed approach can improve upon the initial demonstrator also in practice and achieve better task performance than other learning-based baselines.

****

Connecting Sphere Manifolds Hierarchically for Regularization

Damien Scieur,Youngsung Kim

This paper considers classification problems with hierarchically organized classes. We force the classifier (hyperplane) of each class to belong to a sphere manifold, whose center is the classifier of its super-class. Then, individual sphere manifolds are connected based on their hierarchical relations. Our technique replaces the last layer of a neural network by combining a spherical fully-connected layer with a hierarchical layer. This regularization is shown to improve the performance of widely used deep neural network architectures (ResNet and DenseNet) on publicly available datasets (CIFAR100, CUB200, Stanford dogs, Stanford cars, and Tiny-ImageNet).

****

Model-Free Counterfactual Credit Assignment

Thomas Mesnard,Theophane Weber,Fabio Viola,Shantanu Thakoor,Alaa Saade,Anna Harutyunyan,Will Dabney,Tom Stepleton,Nicolas Heess,Marcus Hutter,Lars Holger Buesing,Remi Munos

Credit assignment in reinforcement learning is the problem of measuring an action's influence on future rewards.

In particular, this requires separating \emph{skill} from \emph{luck}, ie.\ disentangling the effect of an action on rewards from that of external factors and subsequent actions. To achieve this, we adapt the notion of counterfactuals from causality theory to a model-free RL setup.

The key idea is to condition value functions on \emph{future} events, by learning to extract relevant information from a trajectory. We then propose to use these as future-conditional baselines and critics in policy gradient algorithms and we develop a valid, practical variant with provably lower variance, while achieving unbiasedness by constraining the hindsight information not to contain inform

ation about the agent’s actions. We demonstrate the efficacy and validity of our algorithm on a number of illustrative problems.
****
FSV: Learning to Factorize Soft Value Function for Cooperative Multi-Agent Reinforcement Learning
Yueheng Li,Tianhao Zhang,Chen Wang,Jinan Sun,Shikun Zhang,Guangming Xie
We explore energy-based solutions for cooperative multi-agent reinforcement learning (MARL) using the idea of function factorization in centralized training with decentralized execution (CTDE). Existing CTDE based factorization methods are susceptible to the relative overgeneralization, where finding a suboptimal Nash Equilibrium, which is a well-known game-theoretic pathology. To resolve this issue, we propose a novel factorization method for cooperative MARL, named FSV, which learns to factorize the joint soft value function into individual ones for decentralized execution. Theoretical analysis shows that FSV solves a rich class of factorization tasks. Our experiment for the well-known task of the Max of Two Quadratics game shows that FSV fully converges to global optima in the joint action space in the continuous tasks by local searching in the joint action space. We evaluate FSV on a challenging set of StarCraft II micromanagement tasks, and show that FSV significantly outperforms existing factorization multi-agent reinforcement learning methods.
****
Hellinger Distance Constrained Regression
Egor Rotinov
This paper introduces an off-policy reinforcement learning method that uses Hellinger distance between sampling policy (from what samples were collected) and current policy (policy being optimized) as a constraint.
Hellinger distance squared multiplied by two is greater than or equal to total variation distance squared and less than or equal to Kullback-Leibler divergence, therefore a lower bound for expected discounted return for the new policy is improved compared to the lower bound for training with KL.
Also, Hellinger distance is less than or equal to 1, so there is a policy-independent lower bound for expected discounted return.
HDCR is capable of training with Experience Replay, a common setting for distributed RL when collecting trajectories using different policies and learning from this data centralized.
HDCR shows results comparable to or better than Advantage-weighted Behavior Model and Advantage-Weighted Regression on MuJoCo tasks using tiny offline datasets collected by random agents. On bigger datasets (100k timesteps) obtained by pretrained behavioral policy, HDCR outperforms ABM and AWR methods on 3 out of 4 tasks.
****
Collaborative Filtering with Smooth Reconstruction of the Preference Function
Ali Shirali,Reza Kazemi,Arash Amini
The problem of predicting the rating of a set of users to a set of items in a recommender system based on partial knowledge of the ratings is widely known as collaborative filtering. In this paper, we consider a mapping of the items into a vector space and study the prediction problem by assuming an underlying smooth preference function for each user, the quantization at each given vector yields the associated rating. To estimate the preference functions, we implicitly cluster the users with similar ratings to form dominant types. Next, we associate each dominant type with a smooth preference function; i.e., the function values for items with nearby vectors shall be close to each other.
The latter is accomplished by a rich representation learning in a so-called frequency domain. In this framework, we propose two approaches for learning user and item representations. First, we use an alternating optimization method in the spirit of k-means to cluster users and map items. We further make this approach less prone to overfitting by a boosting technique.
Second, we present a feedforward neural network architecture consisting of interpretable layers which implicitely clusters the users. The performance of the method is evaluated on two benchmark datasets (ML-100k and ML-1M). Albeit the metho

d benefits from simplicity, it shows a remarkable performance and opens a venue for future research. All codes are publicly available on the GitLab.
****
A Real-time Contribution Measurement Method for Participants in Federated Learning
Bingjie Yan,Yize Zhou,Boyi Liu,Jun Wang,Yuhan Zhang,Li Liu,Xiaolan Nie,Zhiwei Fan,Zhixuan Liang
Federated learning is a framework for protecting distributed data privacy and has participated in commercial activities. However, there is a lack of a sufficiently reasonable contribution measurement mechanism to distribute the reward for each agent. In the commercial union, if there is no mechanism like this, every agent will get the same reward. This is unfair to agents that provide better data, so such a mechanism is needed. To address this issue, this work proposes a real-time contribution measurement method. Firstly, the method defines the impact of each agent. Furthermore, we comprehensively consider the current round and the previous round to obtain the contribution rate of each agent. To verify effectiveness of the proposed method, the work conducts pseudo-distributed training and an experiment on the Penn Treebank dataset. Comparing the Shapley Value in game theory, the comparative experiment result shows that the proposed method is more sensitive to both data quantity and data quality under the premise of maintaining real-time.
****
Learned Threshold Pruning
Kambiz Azarian,Yash Sanjay Bhalgat,Jinwon Lee,Tijmen Blankevoort
This paper presents a novel differentiable method for unstructured weight pruning of deep neural networks. Our learned-threshold pruning (LTP) method learns per-layer thresholds via gradient descent, unlike conventional methods where they are set as input. Making thresholds trainable also makes LTP computationally efficient, hence scalable to deeper networks. For example, it takes 30 epochs for LTP to prune ResNet50 on ImageNet by a factor of 9.1. This is in contrast to other methods that search for per-layer thresholds via a computationally intensive iterative pruning and fine-tuning process. Additionally, with a novel differentiable L_0 regularization, LTP is able to operate effectively on architectures with batch-normalization. This is important since L_1 and L_2 penalties lose their regularizing effect in networks with batch-normalization. Finally, LTP generates a trail of progressively sparser networks from which the desired pruned network can be picked based on sparsity and performance requirements. These features allow LTP to achieve competitive compression rates on ImageNet networks such as AlexNet ($26.4\times$ compression with 79.1% Top-5 accuracy) and ResNet50 ($9.1\times$ compression with 92.0% Top-5 accuracy). We also show that LTP effectively prunes modern \textit{compact} architectures, such as EfficientNet, MobileNetV2 and MixNet.
****
SCoRe: Pre-Training for Context Representation in Conversational Semantic Parsing
Tao Yu,Rui Zhang,Alex Polozov,Christopher Meek,Ahmed Hassan Awadallah
Conversational Semantic Parsing (CSP) is the task of converting a sequence of natural language queries to formal language (e.g., SQL, SPARQL) that can be executed against a structured ontology (e.g. databases, knowledge bases). To accomplish this task, a CSP system needs to model the relation between the unstructured language utterance and the structured ontology while representing the multi-turn dynamics of the dialog. Pre-trained language models (LMs) are the state-of-the-art for various natural language processing tasks. However, existing pre-trained LMs that use language modeling training objectives over free-form text have limited ability to represent natural language references to contextual structural data. In this work, we present SCORE, a new pre-training approach for CSP tasks designed to induce representations that capture the alignment between the dialogue flow and the structural context. We demonstrate the broad applicability of SCORE to CSP tasks by combining SCORE with strong base systems on four different tasks (SPARC, COSQL, MWOZ, and SQA). We show that SCORE can improve t

he performance over all these base systems by a significant margin and achieves state-of-the-art results on three of them.
****
Expressive yet Tractable Bayesian Deep Learning via Subnetwork Inference
Erik Daxberger,Eric Nalisnick,James Allingham,Javier Antoran,José Miguel Hernández-Lobato
The Bayesian paradigm has the potential to solve some of the core issues in modern deep learning, such as poor calibration, data inefficiency, and catastrophic forgetting. However, scaling Bayesian inference to the high-dimensional parameter spaces of deep neural networks requires restrictive approximations. In this paper, we propose performing inference over only a small subset of the model parameters while keeping all others as point estimates. This enables us to use expressive posterior approximations that would otherwise be intractable for the full model. In particular, we develop a practical and scalable Bayesian deep learning method that first trains a point estimate, and then infers a full covariance Gaussian posterior approximation over a subnetwork. We propose a subnetwork selection procedure which aims to maximally preserve posterior uncertainty. We empirically demonstrate the effectiveness of our approach compared to point-estimated networks and methods that use less expressive posterior approximations over the full network.
****
High-Likelihood Area Matters --- Rewarding Correct,Rare Predictions Under Imbalanced Distributions
guangxiang zhao,Lei Li,Xuancheng Ren,Xu Sun,Bin He
Learning from natural datasets poses significant challenges for traditional classification methods based on the cross-entropy objective due to imbalanced class distributions. It is intuitive to assume that the examples from rare classes are harder to learn so that the classifier is uncertain of the prediction, which establishes the low-likelihood area. Based on this, existing approaches drive the classifier actively to correctly predict those incorrect, rare examples. However, this assumption is one-sided and could be misleading. We find in practice that the high-likelihood area contains correct predictions for rare class examples and it plays a vital role in learning imbalanced class distributions. In light of this finding, we propose the Eureka Loss, which rewards the classifier when examples belong to rare classes in the high-likelihood area are correctly predicted. Experiments on the large-scale long-tailed iNaturalist 2018 classification dataset and the ImageNet-LT benchmark both validate the proposed approach. We further analyze the influence of the Eureka Loss in detail on diverse data distributions.
****
Composite Adversarial Training for Multiple Adversarial Perturbations and Beyond
Xinyang Zhang,Zheng Zhang,Ting Wang
One intriguing property of deep neural networks (DNNs) is their vulnerability to adversarial perturbations. Despite the plethora of work on defending against individual perturbation models, improving DNN robustness against the combinations of multiple perturbations is still fairly under-studied. In this paper, we propose \underline{c}omposite \underline{a}dversarial \underline{t}raining (CAT), a novel training method that flexibly integrates and optimizes multiple adversarial losses, leading to significant robustness improvement with respect to individual perturbations as well as their ``compositions''. Through empirical evaluation on benchmark datasets and models, we show that CAT outperforms existing adversarial training methods by large margins in defending against the compositions of pixel perturbations and spatial transformations, two major classes of adversarial perturbation models, while incurring limited impact on clean inputs.
****
Learning Representation in Colour Conversion
Arash Akbarinia,Raquel Gil-Rodriguez,Alban Flachot,Matteo Toscani
Colours can be represented in an infinite set of spaces highlighting distinct features. Here, we investigated the impact of colour spaces on the encoding capacity of a visual system that is subject to information compression, specifically v

ariational autoencoders (VAEs) where bottlenecks are imposed. To this end, we pr
opose a novel unsupervised task: colour space conversion (ColourConvNets). We tr
ained several instances of VAEs whose input and output are in different colour s
paces, e.g. from RGB to CIE L*a*b* (in total five colour spaces were examined).
This allowed us to systematically study the influence of input-output colour spa
ces on the encoding efficiency and learnt representation. Our evaluations demons
trate that ColourConvNets with decorrelated output colour spaces produce higher
quality images, also evident in pixel-wise low-level metrics such as colour diff
erence (ΔE), peak signal-to-noise ratio (PSNR) and structural similarity
index measure (SSIM). We also assessed the ColourConvNets' capacity to reconstr
uct the global content in two downstream tasks: image classification (ImageNet)
and scene segmentation (COCO). Our results show a 5-10% performance boost for de
correlating ColourConvNets with respect to the baseline network (whose input and
output are RGB). Furthermore, we thoroughly analysed the finite embedding space
of Vector Quantised VAEs with three different methods (single feature, hue shif
t and linear transformation). The interpretations reached with these techniques
are in agreement suggesting that (i) luminance and chromatic information are enc
oded in separate embedding vectors, and (ii) the structure of the network's embe
dding space is determined by the output colour space.
****

Gradient-based tuning of Hamiltonian Monte Carlo hyperparameters
Andrew Campbell,Wenlong Chen,Vincent Stimper,José Miguel Hernández-Lobato,Yichua
n Zhang
Hamiltonian Monte Carlo (HMC) is one of the most successful sampling methods in
machine learning. However, its performance is significantly affected by the choi
ce of hyperparameter values, which require careful tuning. Existing approaches f
or automating this task either optimise a proxy for mixing speed or consider the
HMC chain as an implicit variational distribution and optimize a tractable lowe
r bound that is too loose to be useful in practice. Instead, we propose to optim
ize an objective that quantifies directly the speed of convergence to the target
distribution. Our objective can be easily optimized using stochastic gradient d
escent. We evaluate our proposed method and compare to baselines on a variety o
f problems including synthetic 2D distributions, the posteriors of variational a
utoencoders and the Boltzmann distribution for molecular configurations of a 22
atom molecule. We find our method is competitive with or improves upon alternati
ve baselines on all problems we consider.
****

Representation and Bias in Multilingual NLP: Insights from Controlled Experiment
s on Conditional Language Modeling
Ada Wan
Inspired by the phenomenon of performance disparity between languages in machine
translation, we investigate whether and to what extent languages are equally ha
rd to "conditional-language-model". Our goal is to improve our understanding and
expectation of the relationship between language, data representation, size, an
d performance. We study one-to-one, bilingual conditional language modeling thro
ugh a series of systematically controlled experiments with the Transformer and t
he 6 languages from the United Nations Parallel Corpus. We examine character, by
te, and word models in 30 language directions and 5 data sizes, and observe indi
cations suggesting a script bias on the character level, a length bias on the by
te level, and a word bias that gives rise to a hierarchy in performance across l
anguages. We also identify two types of sample-wise non-monotonicity --- while w
ord-based representations are prone to exhibit Double Descent, length can induce
unstable performance across the size range studied in a novel meta phenomenon w
hich we term "erraticity". By eliminating statistically significant performance
disparity on the character and byte levels by normalizing length and vocabulary
in the data, we show that, in the context of computing with the Transformer, the
re is no complexity intrinsic to languages other than that related to their stat
istical attributes and that performance disparity is not a necessary condition b
ut a byproduct of word segmentation. Our application of statistical comparisons
as a fairness measure also serves as a novel rigorous method for the intrinsic e

valuation of languages, resolving a decades-long debate on language complexity. While these quantitative biases leading to disparity are mitigable through a shallower network, we find room for a human bias to be reflected upon. We hope our work helps open up new directions in the area of language and computing that would be fairer and more flexible and foster a new transdisciplinary perspective for DL-inspired scientific progress.
****

A teacher-student framework to distill future trajectories
Alexander Neitz,Giambattista Parascandolo,Bernhard Schölkopf
By learning to predict trajectories of dynamical systems, model-based methods can make extensive use of all observations from past experience. However, due to partial observability, stochasticity, compounding errors, and irrelevant dynamics, training to predict observations explicitly often results in poor models. Model-free techniques try to side-step the problem by learning to predict values directly. While breaking the explicit dependency on future observations can result in strong performance, this usually comes at the cost of low sample efficiency, as the abundant information about the dynamics contained in future observations goes unused. Here we take a step back from both approaches: Instead of hand-designing how trajectories should be incorporated, a teacher network learns to interpret the trajectories and to provide target activations which guide a student model that can only observe the present. The teacher is trained with meta-gradients to maximize the student's performance on a validation set. We show that our approach performs well on tasks that are difficult for model-free and model-based methods, and we study the role of every component through ablation studies.
****

Episodic Memory for Learning Subjective-Timescale Models
Alexey Zakharov,Matthew Crosby,Zafeirios Fountas
In model-based learning, an agent’s model is commonly defined over transitions between consecutive states of an environment even though planning often requires reasoning over multi-step timescales, with intermediate states either unnecessary, or worse, accumulating prediction error. In contrast, intelligent behaviour in biological organisms is characterised by the ability to plan over varying temporal scales depending on the context. Inspired by the recent works on human time perception, we devise a novel approach to learning a transition dynamics model, based on the sequences of episodic memories that define the agent's subjective timescale – over which it learns world dynamics and over which future planning is performed. We implement this in the framework of active inference and demonstrate that the resulting subjective-timescale model (STM) can systematically vary the temporal extent of its predictions while preserving the same computational efficiency. Additionally, we show that STM predictions are more likely to introduce future salient events (for example new objects coming into view), incentivising exploration of new areas of the environment. As a result, STM produces more informative action-conditioned roll-outs that assist the agent in making better decisions. We validate significant improvement in our STM agent's performance in the Animal-AI environment against a baseline system, trained using the environment's objective-timescale dynamics.
****

Certify or Predict: Boosting Certified Robustness with Compositional Architectures
Mark Niklas Mueller,Mislav Balunovic,Martin Vechev
A core challenge with existing certified defense mechanisms is that while they improve certified robustness, they also tend to drastically decrease natural accuracy, making it difficult to use these methods in practice. In this work, we propose a new architecture which addresses this challenge and enables one to boost the certified robustness of any state-of-the-art deep network, while controlling the overall accuracy loss, without requiring retraining. The key idea is to combine this model with a (smaller) certified network where at inference time, an adaptive selection mechanism decides on the network to process the input sample. The approach is compositional: one can combine any pair of state-of-the-art (e.g., EfficientNet or ResNet) and certified networks, without restriction. The resu

lting architecture enables much higher natural accuracy than previously possible with certified defenses alone, while substantially boosting the certified robustness of deep networks. We demonstrate the effectiveness of this adaptive approach on a variety of datasets and architectures. For instance, on CIFAR-10 with an ℓ_∞ perturbation of 2/255, we are the first to obtain a high natural accuracy (90.1%) with non-trivial certified robustness (27.5%). Notably, prior state-of-the-art methods incur a substantial drop in accuracy for a similar certified robustness.
****
Network-Agnostic Knowledge Transfer for Medical Image Segmentation
Shuhang Wang,Eugene Cheah,Elham Yousef Kalafi,Mercy Asiedu,Alex Benjamin,Vivek Kumar Singh,Ge Zhang,Viksit Kumar,Anthony Edward Samir
Conventional transfer learning leverages weights of pre-trained networks, but mandates the need for similar neural architectures. Alternatively, knowledge distillation can transfer knowledge between heterogeneous networks but often requires access to the original training data or additional generative networks. Knowledge transfer between networks can be improved by being agnostic to the choice of network architecture and reducing the dependence on original training data. We propose a knowledge transfer approach from a teacher to a student network wherein we train the student on an independent transferal dataset, whose annotations are generated by the teacher. Experiments were conducted on five state-of-the-art networks for semantic segmentation and seven datasets across three imaging modalities. We studied knowledge transfer from a single teacher, combination of knowledge transfer and fine-tuning, and knowledge transfer from multiple teachers. The student model with a single teacher achieved similar performance as the teacher; and the student model with multiple teachers achieved better performance than the teachers. The salient features of our algorithm include: 1) no need for original training data or generative networks, 2) knowledge transfer between different architectures, 3) ease of implementation for downstream tasks by using the downstream task dataset as the transferal dataset, 4) knowledge transfer of an ensemble of models, trained independently, into one student model. Extensive experiments demonstrate that the proposed algorithm is effective for knowledge transfer and easily tunable.
****
On the Transfer of Disentangled Representations in Realistic Settings
Andrea Dittadi,Frederik Träuble,Francesco Locatello,Manuel Wuthrich,Vaibhav Agrawal,Ole Winther,Stefan Bauer,Bernhard Schölkopf
Learning meaningful representations that disentangle the underlying structure of the data generating process is considered to be of key importance in machine learning. While disentangled representations were found to be useful for diverse tasks such as abstract reasoning and fair classification, their scalability and real-world impact remain questionable.
We introduce a new high-resolution dataset with 1M simulated images and over 1,800 annotated real-world images of the same setup. In contrast to previous work, this new dataset exhibits correlations, a complex underlying structure, and allows to evaluate transfer to unseen simulated and real-world settings where the encoder i) remains in distribution or ii) is out of distribution.
We propose new architectures in order to scale disentangled representation learning to realistic high-resolution settings and conduct a large-scale empirical study of disentangled representations on this dataset. We observe that disentanglement is a good predictor for out-of-distribution (OOD) task performance.
****
Robust Reinforcement Learning on State Observations with Learned Optimal Adversary
Huan Zhang,Hongge Chen,Duane S Boning,Cho-Jui Hsieh
We study the robustness of reinforcement learning (RL) with adversarially perturbed state observations, which aligns with the setting of many adversarial attacks to deep reinforcement learning (DRL) and is also important for rolling out real-world RL agent under unpredictable sensing noise. With a fixed agent policy, we demonstrate that an optimal adversary to perturb state observations can be fou

nd, which is guaranteed to obtain the worst case agent reward. For DRL settings, this leads to a novel empirical adversarial attack to RL agents via a learned adversary that is much stronger than previous ones. To enhance the robustness of an agent, we propose a framework of alternating training with learned adversaries (ATLA), which trains an adversary online together with the agent using policy gradient following the optimal adversarial attack framework. Additionally, inspired by the analysis of state-adversarial Markov decision process (SA-MDP), we show that past states and actions (history) can be useful for learning a robust agent, and we empirically find a LSTM based policy can be more robust under adversaries. Empirical evaluations on a few continuous control environments show that ATLA achieves state-of-the-art performance under strong adversaries. Our code is available at https://github.com/huanzhang12/ATLA_robust_RL.
****

Interpretability Through Invertibility: A Deep Convolutional Network With Ideal Counterfactuals And Isosurfaces
Leon Sixt,Martin Schuessler,Philipp Weiß,Tim Landgraf
Current state of the art computer vision applications rely on highly complex models. Their interpretability is mostly limited to post-hoc methods which are not guaranteed to be faithful to the model. To elucidate a model’s decision, we present a novel interpretable model based on an invertible deep convolutional network. Our model generates meaningful, faithful, and ideal counterfactuals. Using PCA on the classifier’s input, we can also create “isofactuals”– image interpolations with the same outcome but visually meaningful different features. Counter- and isofactuals can be used to identify positive and negative evidence in an image. This can also be visualized with heatmaps. We evaluate our approach against gradient-based attribution methods, which we find to produce meaningless adversarial perturbations. Using our method, we reveal biases in three different datasets. In a human subject experiment, we test whether non-experts find our method useful to spot spurious correlations learned by a model. Our work is a step towards more trustworthy explanations for computer vision.
****

Active Tuning
Sebastian Otte,Matthias Karlbauer,Martin V. Butz
We introduce Active Tuning, a novel paradigm for optimizing the internal dynamics of recurrent neural networks (RNNs) on the fly. In contrast to the conventional sequence-to-sequence mapping scheme, Active Tuning decouples the RNN's recurrent neural activities from the input stream, using the unfolding temporal gradient signal to tune the internal dynamics into the data stream. As a consequence, the model output depends only on its internal hidden dynamics and the closed-loop feedback of its own predictions; its hidden state is continuously adapted by means of the temporal gradient resulting from backpropagating the discrepancy between the signal observations and the model outputs through time. In this way, Active Tuning infers the signal actively but indirectly based on the originally learned temporal patterns, fitting the most plausible hidden state sequence into the observations. We demonstrate the effectiveness of Active Tuning on several time series prediction benchmarks, including multiple super-imposed sine waves, a chaotic double pendulum, and spatiotemporal wave dynamics. Active Tuning consistently improves the robustness, accuracy, and generalization abilities of all evaluated models. Moreover, networks trained for signal prediction and denoising can be successfully applied to a much larger range of noise conditions with the help of Active Tuning. Thus, given a capable time series predictor, Active Tuning enhances its online signal filtering, denoising, and reconstruction abilities without the need for additional training.
****

PERIL: Probabilistic Embeddings for hybrid Meta-Reinforcement and Imitation Learning
Alvaro Prat,Edward Johns
Imitation learning is a natural way for a human to describe a task to an agent, and it can be combined with reinforcement learning to enable the agent to solve that task through exploration. However, traditional methods which combine imitat

ion learning and reinforcement learning require a very large amount of interacti on data to learn each new task, even when bootstrapping from a demonstration. On e solution to this is to use meta reinforcement learning (meta-RL) to enable an agent to quickly adapt to new tasks at test time. In this work, we introduce a n ew method to combine imitation learning with meta reinforcement learning, Probab ilistic Embeddings for hybrid meta-Reinforcement and Imitation Learning (PERIL). Dual inference strategies allow PERIL to precondition exploration policies on d emonstrations, which greatly improves adaptation rates in unseen tasks. In contr ast to pure imitation learning, our approach is capable of exploring beyond the demonstration, making it robust to task alterations and uncertainties. By exploi ting the flexibility of meta-RL, we show how PERIL is capable of interpolating f rom within previously learnt dynamics to adapt to unseen tasks, as well as unsee n task families, within a set of meta-RL benchmarks under sparse rewards.

****

Practical Real Time Recurrent Learning with a Sparse Approximation

Jacob Menick,Erich Elsen,Utku Evci,Simon Osindero,Karen Simonyan,Alex Graves

Recurrent neural networks are usually trained with backpropagation through time, which requires storing a complete history of network states, and prohibits upda ting the weights "online" (after every timestep). Real Time Recurrent Learning (RTRL) eliminates the need for history storage and allows for online weight updat es, but does so at the expense of computational costs that are quartic in the st ate size. This renders RTRL training intractable for all but the smallest networ ks, even ones that are made highly sparse.

We introduce the Sparse n-step Approximation (SnAp) to the RTRL influence matrix . SnAp only tracks the influence of a parameter on hidden units that are reached by the computation graph within n timesteps of the recurrent core. SnAp with $n=1$ is no more expensive than backpropagation but allows training on arbitrari ly long sequences. We find that it substantially outperforms other RTRL approxim ations with comparable costs such as Unbiased Online Recurrent Optimization. For highly sparse networks, SnAp with $n=2$ remains tractable and can outperform ba ckpropagation through time in terms of learning speed when updates are done onli ne.

****

Exchanging Lessons Between Algorithmic Fairness and Domain Generalization

Elliot Creager,Joern-Henrik Jacobsen,Richard Zemel

Standard learning approaches are designed to perform well on average for the dat a distribution available at training time. Developing learning approaches that a re not overly sensitive to the training distribution is central to research on d omain- or out-of-distribution generalization, robust optimization and fairness. In this work we focus on links between research on domain generalization and alg orithmic fairness---where performance under a distinct but related test distribu tions is studied---and show how the two fields can be mutually beneficial. While domain generalization methods typically rely on knowledge of disjoint "domains" or "environments", "sensitive" label information indicating which demographic g roups are at risk of discrimination is often used in the fairness literature. Dr awing inspiration from recent fairness approaches that improve worst-case perfor mance without knowledge of sensitive groups, we propose a novel domain generaliz ation method that handles the more realistic scenario where environment partitio ns are not provided. We then show theoretically and empirically how different pa rtitioning schemes can lead to increased or decreased generalization performance , enabling us to outperform Invariant Risk Minimization with handcrafted environ ments in multiple cases. We also show how a re-interpretation of IRMv1 allows us for the first time to directly optimize a common fairness criterion, group-suff iciency, and thereby improve performance on a fair prediction task.

****

Monotonic Robust Policy Optimization with Model Discrepancy

Yuankun Jiang,Chenglin Li,Junni Zou,Wenrui Dai,Hongkai Xiong

State-of-the-art deep reinforcement learning (DRL) algorithms tend to overfit in some specific environments due to the lack of data diversity in training. To mi

tigate the model discrepancy between training and target (testing) environments, domain randomization (DR) can generate plenty of environments with a sufficient diversity by randomly sampling environment parameters in simulator. Though standard DR using a uniform distribution improves the average performance on the whole range of environments, the worst-case environment is usually neglected without any performance guarantee. Since the average and worst-case performance are equally important for the generalization in RL, in this paper, we propose a policy optimization approach for concurrently improving the policy's performance in the average case (i.e., over all possible environments) and the worst-case environment. We theoretically derive a lower bound for the worst-case performance of a given policy over all environments. Guided by this lower bound, we formulate an optimization problem which aims to optimize the policy and sampling distribution together, such that the constrained expected performance of all environments is maximized. We prove that the worst-case performance is monotonically improved by iteratively solving this optimization problem. Based on the proposed lower bound, we develop a practical algorithm, named monotonic robust policy optimization (MRPO), and validate MRPO on several robot control tasks. By modifying the environment parameters in simulation, we obtain environments for the same task but with different transition dynamics for training and testing. We demonstrate that MRPO can improve both the average and worst-case performance in the training environments, and facilitate the learned policy with a better generalization capability in unseen testing environments.

****

Brain-like approaches to unsupervised learning of hidden representations - a comparative study

Naresh Balaji,Anders Lansner,Pawel Herman

Unsupervised learning of hidden representations has been one of the most vibrant research directions in machine learning in recent years. In this work we study the brain-like Bayesian Confidence Propagating Neural Network (BCPNN) model, recently extended to extract sparse distributed high-dimensional representations. The saliency and separability of the hidden representations when trained on MNIST dataset is studied using an external linear classifier and compared with other unsupervised learning methods that include restricted Boltzmann machines and autoencoders.

****

A new framework for tensor PCA based on trace invariants

Mohamed Ouerfelli,mohamed Tamaazousti,Vincent Rivasseau

We consider the Principal Component Analysis (PCA) problem for tensors $T \in (\mathbb{R}^n)^{\otimes k}$ of large dimension n and of arbitrary order $k\geq 3$. It consists in recovering a spike $v_0^{\otimes k}$ (related to a signal vector $v_0 \in \mathbb{R}^n$) corrupted by a Gaussian noise tensor $Z \in (\mathbb{R}^n)^{\otimes k}$ such that $T=\beta v_0^{\otimes k} + Z$ where β is the signal-to-noise ratio. In this paper, we propose a new framework based on tools developed by the theoretical physics community to address this important problem. They consist in trace invariants of tensors built by judicious contractions (extension of matrix product) of the indices of the tensor T. Inspired by these tools, we introduce a new process that builds for each invariant a matrix whose top eigenvector is correlated to the signal for β sufficiently large. Then, we give examples of classes of invariants for which we demonstrate that this correlation happens above the best algorithmic threshold ($\beta\geq n^{k/4}$) known so far. This method has many algorithmic advantages: (i) it provides a detection algorithm linear in time and that has only $O(1)$ memory requirements (ii) the algorithms are very suitable for parallel architectures and have a lot of potential of optimization given the simplicity of the mathematical tools involved (iii) experimental results show an improvement of the state of the art for the symmetric tensor PCA. Furthermore, this framework allows more general applications by being able to theoretically study the recovery of a spike in the form of $v_1 \otimes \dots \otimes v_k$ with different dimensions ($T \in \mathbb{R}^{n_1\times n_2\times \dots \times n_k}$ with $n_1,\dots, n_k \in \mathbb{N}$) as well as the recovery of a sum of different orthogonal spikes. We provide experimenta

l results to these different cases that match well with our theoretical findings
.
****
On the Effect of Consensus in Decentralized Deep Learning
Tao Lin,Lingjing Kong,Anastasia Koloskova,Martin Jaggi,Sebastian U Stich
Decentralized training of deep learning models enables on-device learning over n
etworks, as well as efficient scaling to large compute clusters. Experiments in
earlier works revealed that decentralized training often suffers from generaliza
tion issues: the performance of models trained in a decentralized fashion is in
general worse than the performance of models trained in a centralized fashion, a
nd this generalization gap is impacted by parameters such as network size, commu
nication topology, and data partitioning.

We identify the changing consensus distance between devices as a key parameter t
o explain the gap between centralized and decentralized training. We show that w
hen the consensus distance does not grow too large, the performance of centraliz
ed training can be reached and sometimes surpassed. We highlight the intimate in
terplay between network topology and learning rate at the different training pha
ses and discuss the implications for communication efficient training schemes. O
ur insights into the generalization gap in decentralized deep learning allow the
 principled design of better training schemes that mitigate these effects.

****
Structure and randomness in planning and reinforcement learning
Piotr Kozakowski,Piotr Januszewski,Konrad Czechowski,■ukasz Kuci■ski,Piotr Mi■o■
Planning in large state spaces inevitably needs to balance depth and breadth of
the search. It has a crucial impact on planners performance and most manage this
 interplay implicitly. We present a novel method $\textit{Shoot Tree Search (STS
)}$, which makes it possible to control this trade-off more explicitly. Our algo
rithm can be understood as an interpolation between two celebrated search mechan
isms: MCTS and random shooting. It also lets the user control the bias-variance
trade-off, akin to $TD(n)$, but in the tree search context.

In experiments on challenging domains, we show that STS can get the best of both
 worlds consistently achieving higher scores.
****
Retrieval-Augmented Generation for Code Summarization via Hybrid GNN
Shangqing Liu,Yu Chen,Xiaofei Xie,Jing Kai Siow,Yang Liu
Source code summarization aims to generate natural language summaries from struc
tured code snippets for better understanding code functionalities. However, auto
matic code summarization is challenging due to the complexity of the source code
 and the language gap between the source code and natural language summaries. Mo
st previous approaches either rely on retrieval-based (which can take advantage
of similar examples seen from the retrieval database, but have low generalizatio
n performance) or generation-based methods (which have better generalization per
formance, but cannot take advantage of similar examples).
This paper proposes a novel retrieval-augmented mechanism to combine the benefit
s of both worlds.
Furthermore, to mitigate the limitation of Graph Neural Networks (GNNs) on captu
ring global graph structure information of source code, we propose a novel atten
tion-based dynamic graph to complement the static graph representation of the so
urce code, and design a hybrid message passing GNN for capturing both the local
and global structural information. To evaluate the proposed approach, we release
 a new challenging benchmark, crawled from diversified large-scale open-source C
 projects (total 95k+ unique functions in the dataset). Our method achieves the
state-of-the-art performance, improving existing methods by 1.42, 2.44 and 1.29
in terms of BLEU-4, ROUGE-L and METEOR.
****
Learning from others' mistakes: Avoiding dataset biases without modeling them
Victor Sanh,Thomas Wolf,Yonatan Belinkov,Alexander M Rush

State-of-the-art natural language processing (NLP) models often learn to model d
ataset biases and surface form correlations instead of features that target the
intended underlying task. Previous work has demonstrated effective methods to ci
rcumvent these issues when knowledge of the bias is available. We consider cases
 where the bias issues may not be explicitly identified, and show a method for t
raining models that learn to ignore these problematic correlations. Our approach
 relies on the observation that models with limited capacity primarily learn to
exploit biases in the dataset. We can leverage the errors of such limited capaci
ty models to train a more robust model in a product of experts, thus bypassing t
he need to hand-craft a biased model. We show the effectiveness of this method t
o retain improvements in out-of-distribution settings even if no particular bias
 is targeted by the biased model.
****

Fast and Complete: Enabling Complete Neural Network Verification with Rapid and Massively Parallel Incomplete Verifiers
Kaidi Xu,Huan Zhang,Shiqi Wang,Yihan Wang,Suman Jana,Xue Lin,Cho-Jui Hsieh
Formal verification of neural networks (NNs) is a challenging and important prob
lem. Existing efficient complete solvers typically require the branch-and-bound
(BaB) process, which splits the problem domain into sub-domains and solves each
sub-domain using faster but weaker incomplete verifiers, such as Linear Programm
ing (LP) on linearly relaxed sub-domains. In this paper, we propose to use the
backward mode linear relaxation based perturbation analysis (LiRPA) to replace L
P during the BaB process, which can be efficiently implemented on the typical ma
chine learning accelerators such as GPUs and TPUs. However, unlike LP, LiRPA wh
en applied naively can produce much weaker bounds and even cannot check certain
conflicts of sub-domains during splitting, making the entire procedure incomplet
e after BaB. To address these challenges, we apply a fast gradient based bound t
ightening procedure combined with batch splits and the design of minimal usage o
f LP bound procedure, enabling us to effectively use LiRPA on the accelerator ha
rdware for the challenging complete NN verification problem and significantly ou
tperform LP-based approaches. On a single GPU, we demonstrate an order of magnit
ude speedup compared to existing LP-based approaches.
****

Self-supervised Adversarial Robustness for the Low-label, High-data Regime
Sven Gowal,Po-Sen Huang,Aaron van den Oord,Timothy Mann,Pushmeet Kohli
Recent work discovered that training models to be invariant to adversarial pertu
rbations requires substantially larger datasets than those required for standard
 classification. Perhaps more surprisingly, these larger datasets can be "mostly
" unlabeled. Pseudo-labeling, a technique simultaneously pioneered by four separ
ate and simultaneous works in 2019, has been proposed as a competitive alternati
ve to labeled data for training adversarially robust models. However, when the a
mount of labeled data decreases, the performance of pseudo-labeling catastrophic
ally drops, thus questioning the theoretical insights put forward by Uesato et a
l. (2019), which suggest that the sample complexity for learning an adversariall
y robust model from unlabeled data should match the fully supervised case. We in
troduce Bootstrap Your Own Robust Latents (BYORL), a self-supervised learning te
chnique based on BYOL for training adversarially robust models. Our method enabl
es us to train robust representations without any labels (reconciling practice w
ith theory). Most notably, this robust representation can be leveraged by a line
ar classifier to train adversarially robust models, even when the linear classif
ier is not trained adversarially. We evaluate BYORL and pseudo-labeling on CIFAR
-10 and ImageNet and demonstrate that BYORL achieves significantly higher robust
ness (i.e., models resulting from BYORL are up to two times more accurate). Expe
riments on CIFAR-10 against ℓ_2 and ℓ_∞ norm-bounded perturbation
s demonstrate that BYORL achieves near state-of-the-art robustness with as littl
e as 500 labeled examples. We also note that against ℓ_2 norm-bounded pertu
rbations of size $\epsilon = 128/255$, BYORL surpasses the known state-of-the-ar
t with an accuracy under attack of 77.61% (against 72.91% for the prior art).
****

Hippocampal representations emerge when training recurrent neural networks on a

memory dependent maze navigation task
Justin Jude,Matthias Hennig
Can neural networks learn goal-directed behaviour using similar strategies to th e brain, by combining the relationships between the current state of the organis m and the consequences of future actions? Recent work has shown that recurrent n eural networks trained on goal based tasks can develop representations resemblin g those found in the brain, entorhinal cortex grid cells, for instance. Here we explore the evolution of the dynamics of their internal representations and comp are this with experimental data. We observe that once a recurrent network is tra ined to learn the structure of its environment solely based on sensory predictio n, an attractor based landscape forms in the network's representation, which par allels hippocampal place cells in structure and function. Next, we extend the pr edictive objective to include Q-learning for a reward task, where rewarding acti ons are dependent on delayed cue modulation. Mirroring experimental findings in hippocampus recordings in rodents performing the same task, this training paradi gm causes nonlocal neural activity to sweep forward in space at decision points, anticipating the future path to a rewarded location. Moreover, prevalent choice and cue-selective neurons form in this network, again recapitulating experiment al findings. Together, these results indicate that combining predictive, unsuper vised learning of the structure of an environment with reinforcement learning ca n help understand the formation of hippocampus-like representations containing b oth spatial and task-relevant information.
****
Learning to Generate Noise for Multi-Attack Robustness
Divyam Madaan,Jinwoo Shin,Sung Ju Hwang
Adversarial learning has emerged as one of the successful techniques to circumve nt the susceptibility of existing methods against adversarial perturbations. How ever, the majority of existing defense methods are tailored to defend against a single category of adversarial perturbation (e.g. ℓ_∞-attack). In safe ty-critical applications, this makes these methods extraneous as the attacker ca n adopt diverse adversaries to deceive the system. Moreover, training on multipl e perturbations simultaneously significantly increases the computational overhea d during training. To address these challenges, we propose a novel meta-learning framework that explicitly learns to generate noise to improve the model's robus tness against multiple types of attacks. Its key component is Meta Noise Generat or (MNG) that outputs optimal noise to stochastically perturb a given sample, su ch that it helps lower the error on diverse adversarial perturbations. By utiliz ing samples generated by MNG, we train a model by enforcing the label consistenc y across multiple perturbations. We validate the robustness of models trained by our scheme on various datasets and against a wide variety of perturbations, dem onstrating that it significantly outperforms the baselines across multiple pertu rbations with a marginal computational cost.
****
Approximate Probabilistic Inference with Composed Flows
Jay Whang,Erik Lindgren,Alex Dimakis
We study the problem of probabilistic inference on the joint distribution define d by a normalizing flow model. Given a pre-trained flow model $p(\boldsymbol{x}) $, we wish to estimate $p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)$ for some arbi trary partitioning of the variables $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsy mbol{x}_2)$. We first show that this task is computationally hard for a large cl ass of flow models. Motivated by this hardness result, we propose a framework f or $\textit{approximate}$ probabilistic inference. Specifically, our method trai ns a new generative model with the property that its composition with the given model approximates the target conditional distribution. By parametrizing this ne w distribution as another flow model, we can efficiently train it using variatio nal inference and also handle conditioning under arbitrary differentiable transf ormations. Since the resulting approximate posterior remains a flow, it offers e xact likelihood evaluation, inversion, and efficient sampling. We provide an ext ensive empirical evidence showcasing the flexibility of our method on a variety of inference tasks with applications to inverse problems. We also experimentally

demonstrate that our approach is comparable to simple MCMC baselines in terms of sample quality. Further, we explain the failure of naively applying variational inference and show that our method does not suffer from the same issue.
****
Syntactic representations in the human brain: beyond effort-based metrics
Aniketh Janardhan Reddy,Leila Wehbe
We are far from having a complete mechanistic understanding of the brain computations involved in language processing and of the role that syntax plays in those computations. Most language studies do not computationally model syntactic structure, and most studies that do model syntactic processing use effort-based metrics. These metrics capture the effort needed to process the syntactic information given by every word (Brennan et al., 2012; Hale et al., 2018; Brennan et al., 2016). They can reveal where in the brain syntactic processing occurs, but not what features of syntax are processed by different brain regions. Here, we move beyond effort-based metrics and propose explicit features capturing the syntactic structure that is incrementally built while a sentence is being read. Using these features and functional Magnetic Resonance Imaging (fMRI) recordings of participants reading a natural text, we study the brain representation of syntax. We find that our syntactic structure-based features are better than effort-based metrics at predicting brain activity in various parts of the language system. We show evidence of the brain representation of complex syntactic information such as phrase and clause structures. We see that regions well-predicted by syntactic features are distributed in the language system and are not distinguishable from those processing semantics. Our results call for a shift in the approach used for studying syntactic processing.
****
Identifying the Sources of Uncertainty in Object Classification
Luis Armando Pérez Rey,Berk ■■ler,Mike Holenderski,Dmitri Jarnikov
In image-based object classification, the visual appearance of objects determines which class they are assigned to. External variables that are independent of the object, such as the perspective or the lighting conditions, can modify the object's appearance resulting in ambiguous images that lead to misclassifications. Previous work has proposed methods for estimating the uncertainty of predictions and measure their confidence. However, such methods do not indicate which variables are the potential sources that cause uncertainty. In this paper, we propose a method for image-based object classification that uses disentangled representations to indicate which are the external variables that contribute the most to the uncertainty of the predictions. This information can be used to identify the external variables that should be modified to decrease the uncertainty and improve the classification.
****
Reducing the number of neurons of Deep ReLU Networks based on the current theory of Regularization
Jakob Heiss,Alexis Stockinger,Josef Teichmann
We introduce a new Reduction Algorithm which makes use of the properties of ReLU neurons to reduce significantly the number of neurons in a trained Deep Neural Network. This algorithm is based on the recent theory of implicit and explicit regularization in Deep ReLU Networks from (Maennel et al, 2018) and the authors.

We discuss two experiments which illustrate the efficiency of the algorithm to reduce the number of neurons significantly with provably almost no change of the learned function within the training data (and therefore almost no loss in accuracy).
****
MISSO: Minimization by Incremental Stochastic Surrogate Optimization for Large Scale Nonconvex and Nonsmooth Problems
Belhal Karimi,Hoi To Wai,Eric Moulines,Ping Li
Many constrained, nonconvex and nonsmooth optimization problems can be tackled using the majorization-minimization (MM) method which alternates between constructing a surrogate function which upper bounds the objective function, and then mi

nimizing this surrogate. For problems which minimize a finite sum of functions, a stochastic version of the MM method selects a batch of functions at random at each iteration and optimizes the accumulated surrogate.
However, in many cases of interest such as variational inference for latent variable models, the surrogate functions are expressed as an expectation. In this contribution, we propose a doubly stochastic MM method based on Monte Carlo approximation of these stochastic surrogates.
We establish asymptotic and non-asymptotic convergence of our scheme in a constrained, nonconvex, nonsmooth optimization setting. We apply our new framework for inference of logistic regression model with missing data and for variational inference of Bayesian variants of LeNet-5 and Resnet-18 on respectively the MNIST and CIFAR-10 datasets.
****
Modeling the Second Player in Distributionally Robust Optimization
Paul Michel,Tatsunori Hashimoto,Graham Neubig
Distributionally robust optimization (DRO) provides a framework for training machine learning models that are able to perform well on a collection of related data distributions (the "uncertainty set"). This is done by solving a min-max game: the model is trained to minimize its maximum expected loss among all distributions in the uncertainty set. While careful design of the uncertainty set is critical to the success of the DRO procedure, previous work has been limited to relatively simple alternatives that keep the min-max optimization problem exactly tractable, such as f-divergence balls. In this paper, we argue instead for the use of neural generative models to characterize the worst-case distribution, allowing for more flexible and problem-specific selection of the uncertainty set. However, while simple conceptually, this approach poses a number of implementation and optimization challenges. To circumvent these issues, we propose a relaxation of the KL-constrained inner maximization objective that makes the DRO problem more amenable to gradient-based optimization of large scale generative models, and develop model selection heuristics to guide hyper-parameter search. On both toy settings and realistic NLP tasks, we find that the proposed approach yields models that are more robust than comparable baselines.
****
Neural Jump Ordinary Differential Equations: Consistent Continuous-Time Prediction and Filtering
Calypso Herrera,Florian Krach,Josef Teichmann
Combinations of neural ODEs with recurrent neural networks (RNN), like GRU-ODE-Bayes or ODE-RNN are well suited to model irregularly observed time series. While those models outperform existing discrete-time approaches, no theoretical guarantees for their predictive capabilities are available. Assuming that the irregularly-sampled time series data originates from a continuous stochastic process, the L^2-optimal online prediction is the conditional expectation given the currently available information. We introduce the Neural Jump ODE (NJ-ODE) that provides a data-driven approach to learn, continuously in time, the conditional expectation of a stochastic process. Our approach models the conditional expectation between two observations with a neural ODE and jumps whenever a new observation is made. We define a novel training framework, which allows us to prove theoretical guarantees for the first time. In particular, we show that the output of our model converges to the L^2-optimal prediction. This can be interpreted as solution to a special filtering problem. We provide experiments showing that the theoretical results also hold empirically. Moreover, we experimentally show that our model outperforms the baselines in more complex learning tasks and give comparisons on real-world datasets.
****
Gradient Origin Networks
Sam Bond-Taylor,Chris G. Willcocks
This paper proposes a new type of generative model that is able to quickly learn a latent representation without an encoder. This is achieved using empirical Bayes to calculate the expectation of the posterior, which is implemented by initialising a latent vector with zeros, then using the gradient of the log-likelihoo

d of the data with respect to this zero vector as new latent points. The approac h has similar characteristics to autoencoders, but with a simpler architecture, and is demonstrated in a variational autoencoder equivalent that permits samplin g. This also allows implicit representation networks to learn a space of implici t functions without requiring a hypernetwork, retaining their representation adv antages across datasets. The experiments show that the proposed method converges faster, with significantly lower reconstruction error than autoencoders, while requiring half the parameters.
****

On the mapping between Hopfield networks and Restricted Boltzmann Machines
Matthew Smart,Anton Zilman
Hopfield networks (HNs) and Restricted Boltzmann Machines (RBMs) are two importa nt models at the interface of statistical physics, machine learning, and neurosc ience. Recently, there has been interest in the relationship between HNs and RBM s, due to their similarity under the statistical mechanics formalism. An exact m apping between HNs and RBMs has been previously noted for the special case of or thogonal (“uncorrelated”) encoded patterns. We present here an exact mapping in the case of correlated pattern HNs, which are more broadly applicable to existin g datasets. Specifically, we show that any HN with N binary variables and $p<N $ potentially correlated binary patterns can be transformed into an RBM with N binary visible variables and p gaussian hidden variables. We outline the cond itions under which the reverse mapping exists, and conduct experiments on the MN IST dataset which suggest the mapping provides a useful initialization to the RB M weights. We discuss extensions, the potential importance of this correspondenc e for the training of RBMs, and for understanding the performance of feature ext raction methods which utilize RBMs.
****

Efficient Generalized Spherical CNNs
Oliver Cobb,Christopher G. R. Wallis,Augustine N. Mavor-Parker,Augustin Marignie r,Matthew A. Price,Mayeul d'Avezac,Jason McEwen
Many problems across computer vision and the natural sciences require the analys is of spherical data, for which representations may be learned efficiently by en coding equivariance to rotational symmetries. We present a generalized spherica l CNN framework that encompasses various existing approaches and allows them to be leveraged alongside each other. The only existing non-linear spherical CNN l ayer that is strictly equivariant has complexity $\mathcal{O}(C^2L^5)$, where $C $ is a measure of representational capacity and L the spherical harmonic bandl imit. Such a high computational cost often prohibits the use of strictly equiva riant spherical CNNs. We develop two new strictly equivariant layers with reduc ed complexity $\mathcal{O}(CL^4)$ and $\mathcal{O}(CL^3 \log L)$, making larger, more expressive models computationally feasible. Moreover, we adopt efficient sampling theory to achieve further computational savings. We show that these dev elopments allow the construction of more expressive hybrid models that achieve s tate-of-the-art accuracy and parameter efficiency on spherical benchmark problem s.
****

DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION
Pengcheng He,Xiaodong Liu,Jianfeng Gao,Weizhu Chen
Recent progress in pre-trained neural language models has significantly improved the performance of many natural language processing (NLP) tasks. In this paper we propose a new model architecture DeBERTa (Decoding-enhanced BERT with disenta ngled attention) that improves the BERT and RoBERTa models using two novel techn iques. The first is the disentangled attention mechanism, where each word is rep resented using two vectors that encode its content and position, respectively, a nd the attention weights among words are computed using disentangled matrices on their contents and relative positions, respectively. Second, an enhanced mask d ecoder is used to incorporate absolute positions in the decoding layer to predic t the masked tokens in model pre-training. In addition, a new virtual adversaria l training method is used for fine-tuning to improve models’ generalization. We show that these techniques significantly improve the efficie

ncy of model pre-training and the performance of both natural language understan
d(NLU) and natural langauge generation (NLG) downstream tasks. Compared to RoBER
Ta-Large, a DeBERTa model trained on half of the training data performs consiste
ntly better on a wide range of NLP tasks, achieving improvements on MNLI by +0.9
% (90.2% vs. 91.1%), on SQuAD v2.0 by +2.3% (88.4% vs. 90.7%) and RACE by +3.6%
(83.2% vs. 86.8%). Notably, we scale up DeBERTa by training a larger version tha
t consists of 48 Transform layers with 1.5 billion parameters. The significant p
erformance boost makes the single DeBERTa model surpass the human performance on
 the SuperGLUE benchmark (Wang et al., 2019a) for the first time in terms of mac
ro-average score (89.9 versus 89.8), and the ensemble DeBERTa model sits atop th
e SuperGLUE leaderboard as of January 6, 2021, outperforming the human baseline
by a decent margin (90.3 versus
89.8). The pre-trained DeBERTa models and the source code were released at: http
s://github.com/microsoft/DeBERTa.

****
Optimizing Memory Placement using Evolutionary Graph Reinforcement Learning
Shauharda Khadka,Estelle Aflalo,Mattias Marder,Avrech Ben-David,Santiago Miret,S
hie Mannor,Tamir Hazan,Hanlin Tang,Somdeb Majumdar
For deep neural network accelerators, memory movement is both energetically expe
nsive and can bound computation. Therefore, optimal mapping of tensors to memory
 hierarchies is critical to performance. The growing complexity of neural networ
ks calls for automated memory mapping instead of manual heuristic approaches; ye
t the search space of neural network computational graphs have previously been p
rohibitively large. We introduce Evolutionary Graph Reinforcement Learning (EGRL
), a method designed for large search spaces, that combines graph neural network
s, reinforcement learning, and evolutionary search. A set of fast, stateless pol
icies guide the evolutionary search to improve its sample-efficiency. We train a
nd validate our approach directly on the Intel NNP-I chip for inference. EGRL ou
tperforms policy-gradient, evolutionary search and dynamic programming baselines
 on BERT, ResNet-101 and ResNet-50. We additionally achieve 28-78% speed-up comp
ared to the native NNP-I compiler on all three workloads.
****
AutoBayes: Automated Bayesian Graph Exploration for Nuisance-Robust Inference
Andac Demir,Toshiaki Koike-Akino,Ye Wang,Deniz Erdogmus
Learning data representations that capture task-related features, but are invari
ant to nuisance variations remains a key challenge in machine learning. We intro
duce an automated Bayesian inference framework, called AutoBayes, that explores
different graphical models linking classifier, encoder, decoder, estimator and a
dversarial network blocks to optimize nuisance-invariant machine learning pipeli
nes. AutoBayes also enables learning disentangled representations, where the lat
ent variable is split into multiple pieces to impose various relationships with
the nuisance variation and task labels. We benchmark the framework on several pu
blic datasets, and provide analysis of its capability for subject-transfer learn
ing with/without variational modeling and adversarial training. We demonstrate a
 significant performance improvement with ensemble learning across explored grap
hical models.
****
Box-To-Box Transformation for Modeling Joint Hierarchies
Shib Sankar Dasgupta,Xiang Li,Michael Boratko,Dongxu Zhang,Andrew McCallum
Learning representations of entities and relations in knowledge graphs is an act
ive area of research, with much emphasis placed on choosing the appropriate geom
etry to capture tree-like structures. Box embeddings (Vilnis et al., 2018; Li et
 al., 2019; Dasgupta et al., 2020), which represent concepts as n-dimensional hy
perrectangles, are capable of embedding trees by training on a subset of the tra
nsitive closure. In Patel et al. (2020), the authors demonstrate that only the t
ransitive reduction is required, and further extend box embeddings to capture jo
int hierarchies by augmenting the graph with new nodes. While it is possible to
 represent joint hierarchies with this method, the parameters for each hierarchy
 are decoupled, making generalization between hierarchies infeasible. In this wo

rk, we introduce a learned box-to-box transformation which respects the geometri c structure of the box embeddings. We demonstrate that this not only improves t he capability of modeling cross-hierarchy compositional edges but is also capabl e of generalizing from a subset of the transitive reduction.
****
QRGAN: Quantile Regression Generative Adversarial Networks
Sunyeop Lee,Tuan Anh Nguyen,Dugki Min
Learning high-dimensional probability distributions by competitively training ge nerative and discriminative neural networks is a prominent approach of Generativ e Adversarial Networks (GANs) among generative models to model complex real-worl d data. Nevertheless, training GANs likely suffer from non-convergence problem, mode collapse and gradient explosion or vanishing. Least Squares GAN (LSGANs) an d Wasserstein GANs (WGAN) are of representative variants of GANs in literature t hat diminish the inherent problems of GANs by proposing the modification methodo logy of loss functions. However, LSGANs often fall into local minima and cause m ode collapse. While WGANs unexpectedly encounter with inefficient computation an d slow training due to its constraints in Wasserstein distance approximation. In this paper, we propose Quantile Regression GAN (QRGAN) in which quantile regres sion is adopted to minimize 1-Wasserstein distance between real and generated da ta distribution as a novel approach in modification of loss functions for improv ement of GANs. To study the culprits of mode collapse problem, the output space of discriminator and gradients of fake samples are analyzed to see if the discri minator guides the generator well. And we found that the discriminator should no t be bounded to specific numbers. Our proposed QRGAN exposes high robustness aga inst mode collapse problem. Furthermore, QRGAN obtains an apparent improvement i n the evaluation and comparison of Frechet Inception Distance (FID) for generati on performance assessment compared to existing variants of GANs.
****
Run Away From your Teacher: a New Self-Supervised Approach Solving the Puzzle of BYOL
Haizhou Shi,Dongliang Luo,Siliang Tang,Jian Wang,Yueting Zhuang
Recently, a newly proposed self-supervised framework Bootstrap Your Own Latent (BYOL) seriously challenges the necessity of negative samples in contrastive-base d learning frameworks. BYOL works like a charm despite the fact that it discards the negative samples completely and there is no measure to prevent collapse in its training objective. In this paper, we suggest understanding BYOL from the vi ew of our newly proposed interpretable self-supervised learning framework, Run A way From your Teacher (RAFT). RAFT optimizes two objectives at the same time: (i) aligning two views of the same data to similar representations and (ii) runnin g away from the model's Mean Teacher (MT, the exponential moving average of the history models) instead of BYOL's running towards it. The second term of RAFT ex plicitly prevents the representation collapse and thus makes RAFT a more concept ually reliable framework. We provide basic benchmarks of RAFT on CIFAR10 to vali date the effectiveness of our method. Furthermore, we prove that BYOL is equival ent to RAFT under certain conditions, providing solid reasoning for BYOL's count er-intuitive success.
****
The Advantage Regret-Matching Actor-Critic
Audrunas Gruslys,Marc Lanctot,Remi Munos,Finbarr Timbers,Martin Schmid,Julien Pe rolet,Dustin Morrill,Vinicius Zambaldi,Jean-Baptiste Lespiau,John Schultz,Mohamm ad Gheshlaghi Azar,Michael Bowling,Karl Tuyls
Regret minimization has played a key role in online learning, equilibrium comput ation in games, and reinforcement learning (RL). In this paper, we describe a ge neral model-free RL method for no-regret learning based on repeated reconsiderat ion of past behavior: Advantage Regret-Matching Actor-Critic (ARMAC). Rather tha n saving past state-action data, ARMAC saves a buffer of past policies, replayin g through them to reconstruct hindsight assessments of past behavior. These retr ospective value estimates are used to predict conditional advantages which, comb ined with regret matching, produces a new policy. In particular, ARMAC learns fr om sampled trajectories in a centralized training setting, without requiring the

application of importance sampling commonly used in Monte Carlo counterfactual regret (CFR) minimization; hence, it does not suffer from excessive variance in large environments. In the single-agent setting, ARMAC shows an interesting form of exploration by keeping past policies intact. In the multiagent setting, ARMAC in self-play approaches Nash equilibria on some partially-observable zero-sum benchmarks. We provide exploitability estimates in the significantly larger game of betting-abstracted no-limit Texas Hold'em.
****
Simple deductive reasoning tests and numerical data sets for exposing limitation of today's deep neural networks
Kalidas Yeturu,Manish Kumar Srivastava
Learning for Deductive Reasoning is an open problem in the machine learning world today.
Deductive reasoning involves storing facts in memory and generation of newer facts over time.
The concept of memory, processor and code in deduction systems is fundamentally different from the purpose and formulation of weights in a deep neural network.
A majority of the machine learning models are inductive reasoning models including state of the art deep neural networks which are effectively tensor interpolation based models.
A step towards realization of memory is through recurrent neural networks and its variants, however the formal representation is not sufficient enough to capture a complex mapping function between input and output patterns.
Deep neural networks are positioned to do away with feature engineering which is essentially deductive reasoning methodology.
There are existing works in deductive reasoning in neural networks that require learning of syntax, unification and deduction and operate on text data as sequence of tokens.
However the performance of deductive reasoning networks is far from perfection which may be either due to syntax or deduction aspects.
In this context, we have proposed a suite of completely numeric data sets which do not require parsing as with text data.
The 10 data sets are for - (a) selection (3 data sets) - minimum, maximum and top 2nd element in an array of numbers; (b) matching (3 data sets) - duplicate detection, counting and histogram learning; (c) divisibility tests (2 data sets) - divisibility of two numbers and divisibility by 3; (d) representation (2 data sets) - binary representation and parity.
Though extremely simple in terms of feature engineering, in all of these tests, simple deep neural networks, random forest and recurrent neural networks have failed with very low accuracies.
We propose these as numerical test-bed for testing learning models for deductive reasoning.
****
Hybrid and Non-Uniform DNN quantization methods using Retro Synthesis data for efficient inference
TEJPRATAP GVSL,Raja Kumar,Pradeep NS
Existing post-training quantization methods attempt to compensate for the quantization loss by determining the quantized weights and activation ranges with the help of training data. Quantization aware training methods, on the other hand, achieve accuracy near to FP32 models by training the quantized model which consume more time. Both these methods are not effective for privacy constraint applications as they are tightly coupled with training data. In contrast, this paper proposes a data-independent post-training quantization scheme that eliminates the need for training data. This is achieved by generating a faux dataset hereafter called as $\textit{‘Retro-Synthesis Data’}$ from the FP32 model layer statistics and further using it for quantization. This approach outperformed state-of-the-art methods including, but not limited to, ZeroQ and DFQ on models with and without batch-normalization layers for 8, 6 and 4 bit precisions. We also introduced two futuristic variants of post-training quantization methods namely $\textit{‘Hybrid-Quantization’}$ and $\textit{‘Non-Uniform Quantization’}$. The Hybrid-Qua

ntization scheme determines the sensitivity of each layer for per-tensor and per-channel quantization, and thereby generates hybrid quantized models that are $10 - 20\%$ efficient in inference time while achieving same or better accuracy as compared to per-channel quantization. Also this method outperformed FP32 accuracy when applied for models such as ResNet-18, and ResNet-50 onImageNet dataset. In the proposed Non-Uniform quantization scheme, the weights are grouped into different clusters and these clusters are assigned with a varied number of quantization steps depending on the number of weights and their ranges in respective cluster. This method resulted in an accuracy improvement of 1% against state-of-the-art quantization methods on ImageNet dataset.
****

EpidemiOptim: A Toolbox for the Optimization of Control Policies in Epidemiological Models
Cédric Colas,Boris Hejblum,Sébastien Rouillon,Rodolphe Thiebaut,Pierre-Yves Oudeyer,Clément Moulin-Frier,Mélanie Prague
Epidemiologists model the dynamics of epidemics in order to propose control strategies based on pharmaceutical and non-pharmaceutical interventions (contact limitation, lock down, vaccination, etc). Hand-designing such strategies is not trivial because of the number of possible interventions and the difficulty to predict long-term effects. This task can be cast as an optimization problem where state-of-the-art machine learning algorithms such as deep reinforcement learning might bring significant value. However, the specificity of each domain - epidemic modelling or solving optimization problems - requires strong collaborations between researchers from different fields of expertise.
This is why we introduce EpidemiOptim, a Python toolbox that facilitates collaborations between researchers in epidemiology and optimization. EpidemiOptim turns epidemiological models and cost functions into optimization problems via a standard interface commonly used by optimization practitioners (OpenAI Gym). Reinforcement learning algorithms based on Q-Learning with deep neural networks (DQN) and evolutionary algorithms (NSGA-II) are already implemented. We illustrate the use of EpidemiOptim to find optimal policies for dynamical on-off lock-down control under the optimization of death toll and economic recess using a Susceptible-Exposed-Infectious-Removed (SEIR) model for SARS-CoV-2/COVID-19.
Using EpidemiOptim and its interactive visualization platform in Jupyter notebooks, epidemiologists, optimization practitioners and others (e.g. economists) can easily compare epidemiological models, costs functions and optimization algorithms to address important choices to be made by health decision-makers.
****

On the geometry of generalization and memorization in deep neural networks
Cory Stephenson,suchismita padhy,Abhinav Ganesh,Yue Hui,Hanlin Tang,SueYeon Chung
Understanding how large neural networks avoid memorizing training data is key to explaining their high generalization performance. To examine the structure of when and where memorization occurs in a deep network, we use a recently developed replica-based mean field theoretic geometric analysis method. We find that all layers preferentially learn from examples which share features, and link this behavior to generalization performance. Memorization predominately occurs in the deeper layers, due to decreasing object manifolds’ radius and dimension, whereas early layers are minimally affected. This predicts that generalization can be restored by reverting the final few layer weights to earlier epochs before significant memorization occurred, which is confirmed by the experiments. Additionally, by studying generalization under different model sizes, we reveal the connection between the double descent phenomenon and the underlying model geometry. Finally, analytical analysis shows that networks avoid memorization early in training because close to initialization, the gradient contribution from permuted examples are small. These findings provide quantitative evidence for the structure of memorization across layers of a deep neural network, the drivers for such structure, and its connection to manifold geometric properties.

****

Detecting Misclassification Errors in Neural Networks with a Gaussian Process Mo
del
Xin Qiu,Risto Miikkulainen
As neural network classifiers are deployed in real-world applications, it is cru
cial that their predictions are not just accurate, but trustworthy as well. One
practical solution is to assign confidence scores to each prediction, then filte
r out low-confidence predictions. However, existing confidence metrics are not y
et sufficiently reliable for this role. This paper presents a new framework that
produces more reliable confidence scores for detecting misclassification errors
. This framework, RED, calibrates the classifier's inherent confidence indicator
s and estimates uncertainty of the calibrated confidence scores using Gaussian P
rocesses. Empirical comparisons with other confidence estimation methods on 125
UCI datasets demonstrate that this approach is effective. An experiment on a vis
ion task with a large deep learning architecture further confirms that the metho
d can scale up, and a case study involving out-of-distribution and adversarial s
amples shows potential of the proposed method to improve robustness of neural ne
twork classifiers more broadly in the future.
****

Continual learning in recurrent neural networks
Benjamin Ehret,Christian Henning,Maria Cervera,Alexander Meulemans,Johannes Von
Oswald,Benjamin F Grewe
While a diverse collection of continual learning (CL) methods has been proposed
to prevent catastrophic forgetting, a thorough investigation of their effectiven
ess for processing sequential data with recurrent neural networks (RNNs) is lack
ing. Here, we provide the first comprehensive evaluation of established CL metho
ds on a variety of sequential data benchmarks. Specifically, we shed light on th
e particularities that arise when applying weight-importance methods, such as el
astic weight consolidation, to RNNs. In contrast to feedforward networks, RNNs i
teratively reuse a shared set of weights and require working memory to process i
nput samples. We show that the performance of weight-importance methods is not d
irectly affected by the length of the processed sequences, but rather by high wo
rking memory requirements, which lead to an increased need for stability at the
cost of decreased plasticity for learning subsequent tasks. We additionally prov
ide theoretical arguments supporting this interpretation by studying linear RNNs
. Our study shows that established CL methods can be successfully ported to the
recurrent case, and that a recent regularization approach based on hypernetworks
outperforms weight-importance methods, thus emerging as a promising candidate f
or CL in RNNs. Overall, we provide insights on the differences between CL in fee
dforward networks and RNNs, while guiding towards effective solutions to tackle
CL on sequential data.
****

Prediction of Enzyme Specificity using Protein Graph Convolutional Neural Networ
ks
Changpeng Lu,Samuel Z Stentz,Joseph H Lubin,Sijian Wang,Sagar D Khare
Specific molecular recognition by proteins, for example, protease enzymes, is cr
itical for maintaining the robustness of key life processes. The substrate spec
ificity landscape of a protease enzyme comprises the set of all sequence motifs
that are recognized/cut, or just as importantly, not recognized/cut by the enzym
e. Current methods for predicting protease specificity landscapes rely on learni
ng sequence patterns in experimentally derived data with a single enzyme, but ar
e not robust to even small mutational changes. A comprehensive evaluation of spe
cificity requires consideration of the three-dimensional structure and energetic
s of molecular interactions. In this work, we present a protein graph convolutio
nal neural network (PGCN), which uses a physically intuitive, structure-based mo
lecular interaction graph generated using the Rosetta energy function that descr
ibes the topology and energetic features, to determine substrate specificity. We
use the PGCN to recapitulate and predict the specificity of the NS3/4 protease
from the Hepatitic C virus. We compare our PGCN with previously used machine lea
rning models and show that its performance in classification tasks is equivalent
or better. Because PGCN is based on physical interactions, it is inherently mor

e interpretable; determination of feature importance reveals key sub-graph patte
rns responsible for molecular recognition that are biochemically reasonable. The
 PGCN model also readily lends itself to the design of novel enzymes with tailor
ed specificity against disease targets.
****

Joint Learning of Full-structure Noise in Hierarchical Bayesian Regression Model
s

Ali Hashemi,Chang Cai,Klaus Robert Muller,Srikantan Nagarajan,Stefan Haufe

We consider hierarchical Bayesian (type-II maximum likelihood) models for observ
ations with latent variables for source and noise, where both hyperparameters ne
ed to be estimated jointly from data. This problem has application in many domai
ns in imaging including biomagnetic inverse problems. Crucial factors influencin
g accuracy of source estimation are not only the noise level but also its correl
ation structure, but existing approaches have not addressed estimation of noise
covariance matrices with full structure. Here, we consider the reconstruction of
 brain activity from electroencephalography (EEG). This inverse problem can be f
ormulated as a linear regression with independent Gaussian scale mixture priors
for both the source and noise components. As a departure from classical sparse
Bayesan learning (SBL) models where across-sensor observations are assumed to be
 independent and identically distributed, we consider Gaussian noise with full c
ovariance structure. Using Riemannian geometry, we derive an efficient algorith
m for updating both source and noise covariance along the manifold of positive d
efinite matrices. Using the majorization-maximization framework, we demonstrate
that our algorithm has guaranteed and fast convergence. We validate the algorith
m both in simulations and with real data. Our results demonstrate that the novel
 framework significantly improves upon state-of-the-art techniques in the real-w
orld scenario where the noise is indeed non-diagonal and fully-structured.
****

Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching

Jonas Geiping,Liam H Fowl,W. Ronny Huang,Wojciech Czaja,Gavin Taylor,Michael Moe
ller,Tom Goldstein

Data Poisoning attacks modify training data to maliciously control a model train
ed on such data.
In this work, we focus on targeted poisoning attacks which cause a reclassificat
ion of an unmodified test image and as such breach model integrity. We consider
a
particularly malicious poisoning attack that is both ``from scratch" and ``clean
 label", meaning we analyze an attack that successfully works against new, rando
mly initialized models, and is nearly imperceptible to humans, all while perturb
ing only a small fraction of the training data.
Previous poisoning attacks against deep neural networks in this setting have bee
n limited in scope and success, working only in simplified settings or being pro
hibitively expensive for large datasets.
The central mechanism of the new attack is matching the gradient direction of ma
licious examples. We analyze why this works, supplement with practical considera
tions. and show its threat to real-world practitioners, finding that it is the f
irst poisoning method to cause targeted misclassification in modern deep network
s trained from scratch on a full-sized, poisoned ImageNet dataset.
Finally we demonstrate the limitations of existing defensive strategies against
such an attack, concluding that data poisoning is a credible threat, even for la
rge-scale deep learning systems.
****

Overfitting for Fun and Profit: Instance-Adaptive Data Compression

Ties van Rozendaal,Iris AM Huijben,Taco Cohen

Neural data compression has been shown to outperform classical methods in terms
of RD performance, with results still improving rapidly.
At a high level, neural compression is based on an autoencoder that tries to rec
onstruct the input instance from a (quantized) latent representation, coupled wi
th a prior that is used to losslessly compress these latents.
Due to limitations on model capacity and imperfect optimization and generalizati

on, such models will suboptimally compress test data in general.
However, one of the great strengths of learned compression is that if the test-time data distribution is known and relatively low-entropy (e.g. a camera watching a static scene, a dash cam in an autonomous car, etc.), the model can easily be finetuned or adapted to this distribution, leading to improved RD performance.
In this paper we take this concept to the extreme, adapting the full model to a single video, and sending model updates (quantized and compressed using a parameter-space prior) along with the latent representation. Unlike previous work, we finetune not only the encoder/latents but the entire model, and - during finetuning - take into account both the effect of model quantization and the additional costs incurred by sending the model updates. We evaluate an image compression model on I-frames (sampled at 2 fps) from videos of the Xiph dataset, and demonstrate that full-model adaptation improves RD performance by ~1 dB, with respect to encoder-only finetuning.
****

Neural Potts Model
Tom Sercu,Robert Verkuil,Joshua Meier,Brandon Amos,Zeming Lin,Caroline Chen,Jason Liu,Yann LeCun,Alexander Rives
We propose the Neural Potts Model objective as an amortized optimization problem. The objective enables training a single model with shared parameters to explicitly model energy landscapes across multiple protein families. Given a protein sequence as input, the model is trained to predict a pairwise coupling matrix for a Potts model energy function describing the local evolutionary landscape of the sequence. Couplings can be predicted for novel sequences. A controlled ablation experiment assessing unsupervised contact prediction on sets of related protein families finds a gain from amortization for low-depth multiple sequence alignments; the result is then confirmed on a database with broad coverage of protein sequences.
****

GG-GAN: A Geometric Graph Generative Adversarial Network
Igor Krawczuk,Pedro Abranches,Andreas Loukas,Volkan Cevher
We study the fundamental problem of graph generation. Specifically, we treat graph generation from a geometric perspective by associating each node with a position in space and then connecting the edges based on a similarity function. We then provide new solutions to the key challenges that prevent the widespread application of this classical geometric interpretation: (1) modeling complex relations, (2) modeling isomorphic graphs consistently, and (3) fully exploiting the latent distribution.
Our main contribution is dubbed as the geometric graph (GG) generative adversarial network (GAN), which is a Wasserstein GAN that addresses the above challenges. GG-GAN is permutation equivariant and easily scales to generate graphs of tens of thousands of nodes. GG-GAN also strikes a good trade-off between novelty and modeling the distribution statistics, being competitive or surpassing the state-of-the-art methods that are either slower or that are non-equivariant, or that exploit problem-specific knowledge.
****

A Block Minifloat Representation for Training Deep Neural Networks
Sean Fox,Seyedramin Rasoulinezhad,Julian Faraone,david boland,Philip Leong
Training Deep Neural Networks (DNN) with high efficiency can be difficult to achieve with native floating-point representations and commercially available hardware. Specialized arithmetic with custom acceleration offers perhaps the most promising alternative. Ongoing research is trending towards narrow floating-point representations, called minifloats, that pack more operations for a given silicon area and consume less power. In this paper, we introduce Block Minifloat (BM), a new spectrum of minifloat formats capable of training DNNs end-to-end with only 4-8 bit weight, activation and gradient tensors. While standard floating-point representations have two degrees of freedom, via the exponent and mantissa, BM exposes the exponent bias as an additional field for optimization. Crucially, this enables training with fewer exponent bits, yielding dense integer-like hardwa

re for fused multiply-add (FMA) operations. For ResNet trained on ImageNet, 6-bi
t BM achieves almost no degradation in floating-point accuracy with FMA units th
at are $4.1\times(23.9\times)$ smaller and consume $2.3\times(16.1\times)$ less
energy than FP8 (FP32). Furthermore, our 8-bit BM format matches floating-point
accuracy while delivering a higher computational density and faster expected tra
ining times.
****

Training Invertible Linear Layers through Rank-One Perturbations
Andreas Krämer,Jonas Köhler,Frank Noe
Many types of neural network layers rely on matrix properties such as invertibil
ity or orthogonality.
Retaining such properties during optimization with gradient-based stochastic opt
imizers is a challenging task, which is usually addressed by either reparameteri
zation of the affected parameters or by directly optimizing on the manifold.
This work presents a novel approach for training invertible linear layers. In li
eu of directly optimizing
the network parameters, we train rank-one perturbations and add them to the actu
al weight matrices infrequently. This P^{4}Inv update allows keeping track of
inverses and determinants without ever explicitly computing them. We show how su
ch invertible blocks improve the mixing and thus the mode separation of the resu
lting normalizing flows. Furthermore, we outline how the P^4 concept can be ut
ilized to retain properties other than invertibility.
****

Representation Learning via Invariant Causal Mechanisms
Jovana Mitrovic,Brian McWilliams,Jacob C Walker,Lars Holger Buesing,Charles Blun
dell
Self-supervised learning has emerged as a strategy to reduce the reliance on cos
tly supervised signal by pretraining representations only using unlabeled data.
These methods combine heuristic proxy classification tasks with data augmentatio
ns and have achieved significant success, but our theoretical understanding of t
his success remains limited. In this paper we analyze self-supervised representa
tion learning using a causal framework. We show how data augmentations can be m
ore effectively utilized through explicit invariance constraints on the proxy cl
assifiers employed during pretraining. Based on this, we propose a novel self-su
pervised objective, Representation Learning via Invariant Causal Mechanisms (ReL
IC), that enforces invariant prediction of proxy targets across augmentations th
rough an invariance regularizer which yields improved generalization guarantees.
 Further, using causality we generalize contrastive learning, a particular kind
of self-supervised method, and provide an alternative theoretical explanation f
or the success of these methods. Empirically, ReLIC significantly outperform
s competing methods in terms of robustness and out-of-distribution generalizatio
n on ImageNet, while also significantly outperforming these methods on Atari ach
ieving above human-level performance on 51 out of 57 games.
****

On the Role of Pre-training for Meta Few-Shot Learning
Chia-You Chen,Hsuan-Tien Lin,Gang Niu,Masashi Sugiyama
Few-shot learning aims to classify unknown classes of examples with a few new ex
amples per class. There are two key routes for few-shot learning. One is to (pre
-)train a classifier with examples from known classes, and then transfer the pre
-trained classifier to unknown classes using the new examples. The other, called
 meta few-shot learning, is to couple pre-training with episodic training, which
 contains episodes of few-shot learning tasks simulated from the known classes.
Pre-training is known to play a crucial role for the transfer route, but the rol
e of pre-training for the episodic route is less clear. In this work, we study t
he role of pre-training for the episodic route. We find that pre-training serves
 a major role of disentangling representations of known classes, which makes the
 resulting learning tasks easier for episodic training. The finding allows us to
 shift the huge simulation burden of episodic learning to a simpler pre-training
 stage. We justify such a benefit of shift by designing a new disentanglement-ba
sed pre-training model, which helps episodic learning achieve competitive perfor

mance more efficiently.
****
Sparse encoding for more-interpretable feature-selecting representations in probabilistic matrix factorization
Joshua C Chang,Patrick Fletcher,Jungmin Han,Ted L Chang,Shashaank Vattikuti,Bart Desmet,Ayah Zirikly,Carson C Chow
Dimensionality reduction methods for count data are critical to a wide range of applications in medical informatics and other fields where model interpretability is paramount. For such data, hierarchical Poisson matrix factorization (HPF) and other sparse probabilistic non-negative matrix factorization (NMF) methods are considered to be interpretable generative models. They consist of sparse transformations for decoding their learned representations into predictions. However, sparsity in representation decoding does not necessarily imply sparsity in the encoding of representations from the original data features. HPF is often incorrectly interpreted in the literature as if it possesses encoder sparsity. The distinction between decoder sparsity and encoder sparsity is subtle but important. Due to the lack of encoder sparsity, HPF does not possess the column-clustering property of classical NMF -- the factor loading matrix does not sufficiently define how each factor is formed from the original features. We address this deficiency by self-consistently enforcing encoder sparsity, using a generalized additive model (GAM), thereby allowing one to relate each representation coordinate to a subset of the original data features. In doing so, the method also gains the ability to perform feature selection. We demonstrate our method on simulated data and give an example of how encoder sparsity is of practical use in a concrete application of representing inpatient comorbidities in Medicare patients.
****
Data-efficient Hindsight Off-policy Option Learning
Markus Wulfmeier,Dushyant Rao,Roland Hafner,Thomas Lampe,Abbas Abdolmaleki,Tim Hertweck,Michael Neunert,Dhruva Tirumala,Noah Yamamoto Siegel,Nicolas Heess,Martin Riedmiller
Hierarchical approaches for reinforcement learning aim to improve data efficiency and accelerate learning by incorporating different abstractions. We introduce Hindsight Off-policy Options (HO2), an efficient off-policy option learning algorithm, and isolate the impact of action and temporal abstraction in the option framework by comparing flat policies, mixture policies without temporal abstraction, and finally option policies; all with comparable policy optimization. When aiming for data efficiency, we demonstrate the importance of off-policy optimization, as even flat policies trained off-policy can outperform on-policy option methods. In addition, off-policy training and backpropagation through a dynamic programming inference procedure -- through time and through the policy components for every time-step -- enable us to train all components' parameters independently of the data-generating behavior policy. We continue to illustrate challenges in off-policy option learning and the related importance of trust-region constraints. Experimentally, we demonstrate that HO2 outperforms existing option learning methods and that both action and temporal abstraction provide strong benefits in particular in more demanding simulated robot manipulation tasks from raw pixel inputs. Finally, we develop an intuitive extension to encourage temporal abstraction and investigate differences in its impact between learning from scratch and using pre-trained options.
****
Gated Relational Graph Attention Networks
Denis Lukovnikov,Asja Fischer
Relational Graph Neural Networks (GNN) are a class of GNN that are capable of handling multi-relational graphs. Like all GNNs, they suffer from a drop in performance when training deeper networks, which may be caused by vanishing gradients, over-parameterization, and oversmoothing. Previous works have investigated methods that improve the training of deeper GNNs, which include normalization techniques and various types of skip connection within a node. However, learning long-range patterns in multi-relational graphs using GNNs remains an under-explored topic. In this work, we propose a novel GNN architecture based on the Graph Atten

tion Network (GAT) that uses gated skip connections to improve long-range modeli ng between nodes and uses a more scalable vector-based approach for parameterizi ng relations. We perform an extensive experimental analysis on synthetic and rea l data, focusing explicitly on learning long-range patterns. The results indicat e that the proposed method significantly outperforms several commonly used relat ional GNN variants when used in deeper configurations and stays competitive to e xisting architectures in a shallow setup.
****
Efficient estimates of optimal transport via low-dimensional embeddings
Patric Fulop,Vincent Danos
Optimal transport distances (OT) have been widely used in recent work in Machine Learning as ways to compare probability distributions. These are costly to comp ute when the data lives in high dimension.
Recent work aims specifically at reducing this cost by computing OT using low-ra nk projections of the data (seen as discrete measures)~\citep{paty2019subspace}. We extend this approach and show that one can approximate OT distances by using more general families of maps provided they are 1-Lipschitz. The best estimate is obtained by maximising OT over the given family. As OT calculations are done after mapping data to a lower dimensional space, our method scales well with the original data dimension.
We demonstrate the idea with neural networks.
****
Mapping the Timescale Organization of Neural Language Models
Hsiang-Yun Sherry Chien,Jinhan Zhang,Christopher Honey
In the human brain, sequences of language input are processed within a distribut ed and hierarchical architecture, in which higher stages of processing encode co ntextual information over longer timescales. In contrast, in recurrent neural ne tworks which perform natural language processing, we know little about how the m ultiple timescales of contextual information are functionally organized. Therefo re, we applied tools developed in neuroscience to map the “processing timescales ” of individual units within a word-level LSTM language model. This timescale-ma pping method assigned long timescales to units previously found to track long-ra nge syntactic dependencies. Additionally, the mapping revealed a small subset of the network (less than 15% of units) with long timescales and whose function ha d not previously been explored. We next probed the functional organization of th e network by examining the relationship between the processing timescale of unit s and their network connectivity. We identified two classes of long-timescale un its: “controller” units composed a densely interconnected subnetwork and strongl y projected to the rest of the network, while “integrator” units showed the long est timescales in the network, and expressed projection profiles closer to the m ean projection profile. Ablating integrator and controller units affected model performance at different positions within a sentence, suggesting distinctive fun ctions of these two sets of units. Finally, we tested the generalization of thes e results to a character-level LSTM model and models with different architecture s. In summary, we demonstrated a model-free technique for mapping the timescale organization in recurrent neural networks, and we applied this method to reveal the timescale and functional organization of neural language models
****
Universal Sentence Representations Learning with Conditional Masked Language Mod el
Ziyi Yang,Yinfei Yang,Daniel M Cer,Jax Law,Eric Darve
This paper presents a novel training method, Conditional Masked Language Modelin g (CMLM), to effectively learn sentence representations on large scale unlabeled corpora. CMLM integrates sentence representation learning into MLM training by conditioning on the encoded vectors of adjacent sentences. Our English CMLM mode l achieves state-of-the-art performance on SentEval, even outperforming models l earned using (semi-)supervised signals. As a fully unsupervised learning method, CMLM can be conveniently extended to a broad range of languages and domains. We find that a multilingual CMLM model co-trained with bitext retrieval~(BR) and n atural language inference~(NLI) tasks outperforms the previous state-of-the-art

multilingual models by a large margin. We explore the same language bias of the learned representations, and propose a principle component based approach to remove the language identifying information from the representation while still retaining sentence semantics.
****
GLUECode: A Benchmark for Source Code Machine Learning Models
Anjan Karmakar,Julian Aron Prenner,Miltiadis Allamanis,Romain Robbes
A multitude of machine learning models for source code have been proposed in the recent years capturing various aspects of the inherent rich structure and semantics of code. However, these models are commonly designed to perform well on a single task, failing to capture code's multifaceted nature. To address this, we present GLUECode, Global and Local Understanding Evaluation of Code, a benchmark of diverse tasks to evaluate machine learning models of source code.

Crucially, GLUECode accounts for the distinct characteristics of source code: (1) source code is highly structured and (2) source code is often composed of multiple interacting entities. Existing tasks incentivize researchers to create models and code representations that perform well on a single task - commonly focusing on local reasoning. GLUECode aims to allow researchers to experiment with multiple local and global source code representations, and evaluate these models on their ability to capture the diverse characteristics of source code, thus driving the community towards building robust source code models incorporating global reasoning.

We present results for several baselines. The GLUECode tasks are challenging for the evaluated baselines; no model achieves convincing performance across all tasks. This indicates that there is ample room for progress on GLUECode.
****
Which Model to Transfer? Finding the Needle in the Growing Haystack
Cedric Renggli,André Susano Pinto,Luka Rimanic,Joan Puigcerver,Carlos Riquelme Ruiz,Ce Zhang,Mario Lucic
Transfer learning has been recently popularized as a data-efficient alternative to training models from scratch, in particular in vision and NLP where it provides a remarkably solid baseline. The emergence of rich model repositories, such as TensorFlow Hub, enables the practitioners and researchers to unleash the potential of these models across a wide range of downstream tasks. As these repositories keep growing exponentially, efficiently selecting a good model for the task at hand becomes paramount. We provide a formalization of this problem through a familiar notion of regret and introduce the predominant strategies, namely task-agnostic (e.g. picking the highest scoring ImageNet model) and task-aware search strategies (such as linear or kNN evaluation). We conduct a large-scale empirical study and show that both task-agnostic and task-aware methods can yield high regret. We then propose a simple and computationally efficient hybrid search strategy which outperforms the existing approaches. We highlight the practical benefits of the proposed solution on a set of 19 diverse vision tasks.
****
Disentangling Representations of Text by Masking Transformers
Xiongyi Zhang,Jan-Willem van de Meent,Byron C Wallace
Representations in large language models such as BERT encode a range of features into a single vector, which are predictive in the context of a multitude of downstream tasks. In this paper, we explore whether it is possible to learn disentangled representations by identifying subnetworks in pre-trained models that encode distinct, complementary aspects of the representation. Concretely, we learn binary masks over transformer weights or hidden units to uncover the subset of features that correlate with a specific factor of variation. This sidesteps the need to train a disentangled model from scratch within a particular domain. We evaluate the ability of this method to disentangle representations of syntax and semantics, and sentiment from genre in the context of movie reviews. By combining this method with magnitude pruning we find that we can identify quite sparse subnetworks. Moreover, we find that this disentanglement-via-masking approach perf

orms as well as or better than previously proposed methods based on variational autoencoders and adversarial training.
****
ROMUL: Scale Adaptative Population Based Training
Daniel HAZIZA,Jérémy Rapin,Gabriel Synnaeve
In most pragmatic settings, data augmentation and regularization are essential, and require hyperparameter search.
Population based training (PBT) is an effective tool for efficiently finding the m as well as schedules over hyperparameters.
In this paper, we compare existing PBT algorithms and contribute a new one: ROMU L, for RObust MULtistep search, which adapts its stepsize over the course of tra ining.
We report competitive results with standard models on CIFAR (image classificatio n) as well as Penn Tree Bank (language modeling), which both depend on heavy reg ularization.
We also open-source hoptim, a PBT library agnostic to the training framework, wh ich is simple to use, reentrant, and provides good defaults with ROMUL.
****
Deep Ensembles for Low-Data Transfer Learning
Basil Mustafa,Carlos Riquelme Ruiz,Joan Puigcerver,André Susano Pinto,Daniel Key sers,Neil Houlsby
In the low-data regime, it is difficult to train good supervised models from scr atch.
Instead practitioners turn to pre-trained models, leveraging transfer learning. Ensembling is an empirically and theoretically appealing way to construct powerf ul predictive models, but the predominant approach of training multiple deep net works with different random initialisations collides with the need for transfer via pre-trained weights. In this work, we study different ways of creating ensem bles from pre-trained models. We show that the nature of pre-training itself is a performant source of diversity, and propose a practical algorithm that efficie ntly identifies a subset of pre-trained models for any downstream dataset. The a pproach is simple: Use nearest-neighbour accuracy to rank pre-trained models, fi ne-tune the best ones with a small hyperparameter sweep, and greedily construct an ensemble to minimise validation cross-entropy. When evaluated together with s trong baselines on 19 different downstream tasks (the Visual Task Adaptation Ben chmark), this achieves state-of-the-art performance at a much lower inference bu dget, even when selecting from over 2,000 pre-trained models. We also assess our ensembles on ImageNet variants and show improved robustness to distribution shi ft.
****
Neural networks with late-phase weights
Johannes Von Oswald,Seijin Kobayashi,Joao Sacramento,Alexander Meulemans,Christi an Henning,Benjamin F Grewe
The largely successful method of training neural networks is to learn their weig hts using some variant of stochastic gradient descent (SGD). Here, we show that the solutions found by SGD can be further improved by ensembling a subset of the weights in late stages of learning. At the end of learning, we obtain back a si ngle model by taking a spatial average in weight space. To avoid incurring incre ased computational costs, we investigate a family of low-dimensional late-phase weight models which interact multiplicatively with the remaining parameters. Our results show that augmenting standard models with late-phase weights improves g eneralization in established benchmarks such as CIFAR-10/100, ImageNet and enwik 8. These findings are complemented with a theoretical analysis of a noisy quadra tic problem which provides a simplified picture of the late phases of neural net work learning.
****
Neural Ensemble Search for Uncertainty Estimation and Dataset Shift
Sheheryar Zaidi,Arber Zela,Thomas Elsken,Chris Holmes,Frank Hutter,Yee Whye Teh
Ensembles of neural networks achieve superior performance compared to stand-alon e networks not only in terms of predictive performance, but also uncertainty cal

ibration and robustness to dataset shift. Diversity among networks is believed t
o be key for building strong ensembles, but typical approaches, such as \emph{de
ep ensembles}, only ensemble different weight vectors of a fixed architecture. I
nstead, we propose two methods for constructing ensembles to exploit diversity a
mong networks with \emph{varying} architectures. We find that the resulting ense
mbles are indeed more diverse and also exhibit better uncertainty calibration, p
redictive performance and robustness to dataset shift in comparison with deep en
sembles on a variety of classification tasks.
****

Hindsight Curriculum Generation Based Multi-Goal Experience Replay
Xiaoyun Feng
In multi-goal tasks with sparse rewards, it is challenging to learn from tons of
 experiences with zero rewards. Hindsight experience replay (HER), which replays
 past experiences with additional heuristic goals, has shown it possible for off
-policy reinforcement learning (RL) to make use of failed experiences. However,
the replayed experiences may not lead to well-explored state-action pairs, espec
ially for a pseudo goal, which instead results in a poor estimate of the value f
unction. To tackle the problem, we propose to resample hindsight experiences bas
ed on their likelihood under the current policy and the overall distribution. Ba
sed on the hindsight strategy, we introduce a novel multi-goal experience replay
 method that automatically generates a training curriculum, namely Hindsight Cur
riculum Generation (HCG). As the range of experiences expands, the generated cur
riculum strikes a dynamic balance between exploiting and exploring. We implement
 HCG with the vanilla Deep Deterministic Policy Gradient(DDPG), and experiments
on several tasks with sparse binary rewards demonstrate that HCG improves sample
 efficiency of the state of the art.
****

Uncertainty-aware Active Learning for Optimal Bayesian Classifier
Guang Zhao,Edward Dougherty,Byung-Jun Yoon,Francis Alexander,Xiaoning Qian
For pool-based active learning, in each iteration a candidate training sample is
 chosen for labeling by optimizing an acquisition function. In Bayesian classifi
cation, expected Loss Reduction~(ELR) methods maximize the expected reduction in
 the classification error given a new labeled candidate based on a one-step-look
-ahead strategy. ELR is the optimal strategy with a single query; however, since
 such myopic strategies cannot identify the long-term effect of a query on the c
lassification error, ELR may get stuck before reaching the optimal classifier.
In this paper, inspired by the mean objective cost of uncertainty (MOCU), a metr
ic quantifying the uncertainty directly affecting the classification error, we p
ropose an acquisition function based on a weighted form of MOCU. Similar to ELR,
 the proposed method focuses on the reduction of the uncertainty that pertains t
o the classification error. But unlike any other existing scheme, it provides th
e critical advantage that the resulting Bayesian active learning algorithm guara
ntees convergence to the optimal classifier of the true model. We demonstrate it
s performance with both synthetic and real-world datasets.
****

ResNet After All: Neural ODEs and Their Numerical Solution
Katharina Ott,Prateek Katiyar,Philipp Hennig,Michael Tiemann
A key appeal of the recently proposed Neural Ordinary Differential Equation (ODE
) framework is that it seems to provide a continuous-time extension of discrete
residual neural networks.
As we show herein, though, trained Neural ODE models actually depend on the spec
ific numerical method used during training.
If the trained model is supposed to be a flow generated from an ODE, it should b
e possible to choose another numerical solver with equal or smaller numerical er
ror without loss of performance.
We observe that if training relies on a solver with overly coarse discretization
, then testing with another solver of equal or smaller numerical error results i
n a sharp drop in accuracy.
In such cases, the combination of vector field and numerical method cannot be in
terpreted as a flow generated from an ODE, which arguably poses a fatal breakdow

n of the Neural ODE concept.
We observe, however, that there exists a critical step size beyond which the tra ining yields a valid ODE vector field.
We propose a method that monitors the behavior of the ODE solver during training to adapt its step size, aiming to ensure a valid ODE without unnecessarily incr easing computational cost.
We verify this adaption algorithm on a common bench mark dataset as well as a sy nthetic dataset.

****
Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods
Taiji Suzuki,Shunta Akiyama
Establishing a theoretical analysis that explains why deep learning can outperfo rm shallow learning such as kernel methods is one of the biggest issues in the d eep learning literature. Towards answering this question, we evaluate excess ris k of a deep learning estimator trained by a noisy gradient descent with ridge re gularization on a mildly overparameterized neural network,
and discuss its superiority to a class of linear estimators that includes neural tangent kernel approach, random feature model, other kernel methods, k-NN est imator and so on. We consider a teacher-student regression model, and eventually show that {\it any} linear estimator can be outperformed by deep learning in a sense of the minimax optimal rate especially for a high dimension setting. The o btained excess bounds are so-called fast learning rate which is faster than $O(1 /\sqrt{n})$ that is obtained by usual Rademacher complexity analysis. This discr epancy is induced by the non-convex geometry of the model and the noisy gradient descent used for neural network training provably reaches a near global optimal solution even though the loss landscape is highly non-convex. Although the nois y gradient descent does not employ any explicit or implicit sparsity inducing re gularization, it shows a preferable generalization performance that dominates li near estimators.
****
Deep Ecological Inference
Nic Fishman,Colin McAuliffe
We introduce an efficient approximation to the loss function for the ecological inference problem, where individual labels are predicted from aggregates. This a llows us to construct ecological versions of linear models, deep neural networks , and Bayesian neural networks. Using these models we infer probabilities of vot e choice for candidates in the Maryland 2018 midterm elections for 2,322,277 vot ers in 2055 precincts. We show that increased network depth and joint learning o f multiple races within an election improves the accuracy of ecological inferenc e when compared to benchmark data from polling. Additionally we leverage data on the joint distribution of ballots (available from ballot images which are publi c for election administration purposes) to show that joint learning leads to sig nificantly improved recovery of the covariance structure for multi-task ecologic al inference. Our approach also allows learning latent representations of voters , which we show outperform raw covariates for leave-one-out prediction.
****
Grey-box Extraction of Natural Language Models
Santiago Zanella-Beguelin,Shruti Tople,Andrew Paverd,Boris Köpf
Model extraction attacks attempt to replicate a target machine learning model fr om predictions obtained by querying its inference API. Most existing attacks on Deep Neural Networks achieve this by supervised training of the copy using the v ictim's predictions. An emerging class of attacks exploit algebraic properties o f DNNs to obtain high-fidelity copies using orders of magnitude fewer queries th an the prior state-of-the-art. So far, such powerful attacks have been limited t o networks with few hidden layers and ReLU activations.

In this paper we present algebraic attacks on large-scale natural language model s in a grey-box setting, targeting models with a pre-trained (public) encoder fo

llowed by a single (private) classification layer. Our key observation is that a small set of arbitrary embedding vectors is likely to form a basis of the classification layer's input space, which a grey-box adversary can compute. We show how to use this information to solve an equation system that determines the classification layer from the corresponding probability outputs.

We evaluate the effectiveness of our attacks on different sizes of transformer models and downstream tasks. Our key findings are that (i) with frozen base layers, high-fidelity extraction is possible with a number of queries that is as small as twice the input dimension of the last layer. This is true even for queries that are entirely in-distribution, making extraction attacks indistinguishable from legitimate use; (ii) with fine-tuned base layers, the effectiveness of algebraic attacks decreases with the learning rate, showing that fine-tuning is not only beneficial for accuracy but also indispensable for model confidentiality.
****
Cortico-cerebellar networks as decoupled neural interfaces
Joseph Pemberton,Ellen Boven,Richard Apps,Rui Ponte Costa
The brain solves the credit assignment problem remarkably well. For credit to be correctly assigned across multiple cortical areas a given area should, in principle, wait for others to finish their computation. How the brain deals with this locking problem has remained unclear. Deep learning methods suffer from similar locking constraints both on the forward and backward phase. Recently, decoupled neural interfaces (DNI) were introduced as a solution to the forward and backward locking problems.
Here we propose that a specialised brain region, the cerebellum, helps the cerebral cortex solve the locking problem closely matching the computations and architecture of DNI. In particular, we propose that classical cerebellar forward and inverse models are equivalent to solving the backward and forward locking problems, respectively. To demonstrate the potential of this framework we focus on modelling a given brain area as a recurrent neural network in which the cerebellum approximates temporal feedback signals as provided by BPTT. We tested the cortico-cerebellar-DNI (CC-DNI) model in a range of sensorimotor and cognitive tasks that have been shown to be cerebellar-dependent. First, we show that the CC-DNI unlocking mechanisms can facilitate learning in a simple target reaching task. Next, by building on the sequential MNIST task we demonstrate that these results generalise to more complex sensorimotor tasks. Our cortico-cerebellar model readily applies to a wider range of modalities, to demonstrate this we tested the model in a cognitive task, caption generation. Models without the cerebellar-DNI component exhibit deficits similar to those observed in cerebellar patients in both motor and cognitive tasks. Moreover, we used CC-DNI to generate a set of specific neuroscience predictions. Finally, we introduce a CC-DNI model with highly sparse connectivity as observed in the cerebellum, which substantially reduces the number of parameters while improving learning through decorrelation.
Overall, our work offers a novel perspective on the cerebellum as a brain-wide decoupling machine for efficient credit assignment and opens a new avenue of research between deep learning and neuroscience.
****
Iterative convergent computation is not a useful inductive bias for ResNets
Samuel Lippl,Benjamin Peters,Nikolaus Kriegeskorte
Recent work has suggested that feedforward residual neural networks (ResNets) approximate iterative recurrent computations. Iterative computations are useful in many domains, so they might provide good solutions for neural networks to learn. Here we quantify the degree to which ResNets learn iterative solutions and introduce a regularization approach that encourages learning of iterative solutions. Iterative methods are characterized by two properties: iteration and convergence. To quantify these properties, we define three indices of iterative convergence. Consistent with previous work, we show that, even though ResNets can express iterative solutions, they do not learn them when trained conventionally on computer vision tasks. We then introduce regularizations to encourage iterative convergent computation and test whether this provides a useful inductive bias. To ma

ke the networks more iterative, we manipulate the degree of weight sharing acros s layers using soft gradient coupling. This new method provides a form of recurr ence regularization and can interpolate smoothly between an ordinary ResNet and a ``recurrent" ResNet (i.e., one that uses identical weights across layers and t hus could be physically implemented with a recurrent network computing the succe ssive stages iteratively across time). To make the networks more convergent we i mpose a Lipschitz constraint on the residual functions using spectral normalizat ion. The three indices of iterative convergence reveal that the gradient couplin g and the Lipschitz constraint succeed at making the networks iterative and conv ergent, respectively. However, neither recurrence regularization nor spectral no rmalization improve classification accuracy on standard visual recognition tasks (MNIST, CIFAR-10, CIFAR-100) or on challenging recognition tasks with partial o cclusions (Digitclutter). Iterative convergent computation, in these tasks, does not provide a useful inductive bias for ResNets.

****

Uncertainty for deep image classifiers on out of distribution data.

Tiago Salvador,Alexander Iannantuono,Adam M Oberman

In addition to achieving high accuracy, in many applications, it is important t o estimate the probability that a model prediction is correct. Predictive uncer tainty is particularly important on out of distribution (OOD) data where accurac y degrades. However, models are typically overconfident, and model calibration on OOD data remains a challenge. In this paper we propose a simple post hoc cali bration method that significantly improves on benchmark results [Ovadia et al 2 019] on a wide range of corrupted data. Our method uses outlier exposure to pro perly calibrate the model probabilities.

****

Generalized Variational Continual Learning

Noel Loo,Siddharth Swaroop,Richard E Turner

Continual learning deals with training models on new tasks and datasets in an on line fashion. One strand of research has used probabilistic regularization for c ontinual learning, with two of the main approaches in this vein being Online Ela stic Weight Consolidation (Online EWC) and Variational Continual Learning (VCL). VCL employs variational inference, which in other settings has been improved em pirically by applying likelihood-tempering. We show that applying this modificat ion to VCL recovers Online EWC as a limiting case, allowing for interpolation be tween the two approaches. We term the general algorithm Generalized VCL (GVCL). In order to mitigate the observed overpruning effect of VI, we take inspiration from a common multi-task architecture, neural networks with task-specific FiLM l ayers, and find that this addition leads to significant performance gains, speci fically for variational methods. In the small-data regime, GVCL strongly outperf orms existing baselines. In larger datasets, GVCL with FiLM layers outperforms o r is competitive with existing baselines in terms of accuracy, whilst also provi ding significantly better calibration.

****

Succinct Explanations with Cascading Decision Trees

JIALU ZHANG,Mark Santolucito,Ruzica Piskac

Classic decision tree learning is a binary classification algorithm that constru cts models with first-class transparency - every classification has a directly d erivable explanation. However, learning decision trees on modern datasets genera tes large trees, which in turn generate decision paths of excessive depth, obscu ring the explanation of classifications. To improve the comprehensibility of cla ssifications, we propose a new decision tree model that we call Cascading Decisi on Trees. Cascading Decision Trees shorten the size of explanations of classific ations, without sacrificing model performance overall. Our key insight is to sep arate the notion of a decision path and an explanation path. Utilizing this insi ght, instead of having one monolithic decision tree, we build several smaller de cision subtrees and cascade them in sequence. Our cascading decision subtrees ar e designed to specifically target explanations for positive classifications. Thi s way each subtree identifies the smallest set of features that can classify as many positive samples as possible, without misclassifying any negative samples.

Applying cascading decision trees to new samples results in a significantly shor ter and succinct explanation, if one of the subtrees detects a positive classifi cation. In that case, we immediately stop and report the decision path of only t he current subtree to the user as an explanation for the classification. We eval uate our algorithm on standard datasets, as well as new real-world applications and find that our model shortens the explanation depth by over 40.8\% for positi ve classifications compared to the classic decision tree model.

****
On the Importance of Looking at the Manifold
Nil Adell Mill,Jannis Born,Nathaniel Park,James Hedrick,María Rodríguez Martínez ,Matteo Manica
Data rarely lies on uniquely Euclidean spaces. Even data typically represented i n regular domains, such as images, can have a higher level of relational informa tion, either between data samples or even relations within samples, e.g., how th e objects in an image are linked. With this perspective our data points can be e nriched by explicitly accounting for this connectivity and analyzing them as a g raph. Herein, we analyze various approaches for unsupervised representation lear ning and investigate the importance of considering topological information and i ts impact when learning representations. We explore a spectrum of models, rangin g from uniquely learning representations based on the isolated features of the n odes (focusing on Variational Autoencoders), to uniquely learning representation s based on the topology (using node2vec) passing through models that integrate b oth node features and topological information in a hybrid fashion. For the latte r we use Graph Neural Networks, precisely Deep Graph Infomax (DGI), and an exten sion of the typical formulation of the VAE where the topological structure is ac counted for via an explicit regularization of the loss (Graph-Regularized VAEs, introduced in this work). To extensively investigate these methodologies, we con sider a wide variety of data types: synthetic data point clouds, MNIST, citation networks, and chemical reactions. We show that each of the representations lear ned by these models may have critical importance for further downstream tasks, a nd that accounting for the topological features can greatly improve the modeling capabilities for certain problems. We further provide a framework to analyze th ese, and future models under different scenarios and types of data.
****
Pretrain-to-Finetune Adversarial Training via Sample-wise Randomized Smoothing
Lei Wang,Runtian Zhai,Di He,Liwei Wang,Li Jian
Developing certified models that can provably defense adversarial perturbations is important in machine learning security. Recently, randomized smoothing, combi ned with other techniques (Cohen et al., 2019; Salman et al., 2019), has been s hown to be an effective method to certify models under l_2 perturbations. Exi sting work for certifying l_2 perturbations added the same level of Gaussian n oise to each sample. The noise level determines the trade-off between the test a ccuracy and the average certified robust radius. We propose to further improve t he defense via sample-wise randomized smoothing, which assigns different noise l evels to different samples. Specifically, we propose a pretrain-to-finetune fram ework that first pretrains a model and then adjusts the noise levels for higher performance based on the model’s outputs. For certification, we carefully alloca te specific robust regions for each test sample. We perform extensive experiment s on CIFAR-10 and MNIST datasets and the experimental results demonstrate that o ur method can achieve better accuracy-robustness trade-off in the transductive s etting.
****
Evaluating Gender Bias in Natural Language Inference
Shanya Sharma,Manan Dey,Koustuv Sinha
Gender-bias stereotypes have recently raised significant ethical concerns in nat ural language processing. However, progress in the detection and evaluation of g ender-bias in natural language understanding through inference is limited and re quires further investigation. In this work, we propose an evaluation methodology to measure these biases by constructing a probe task that involves pairing a ge

nder-neutral premise against a gender-specific hypothesis. We use our probe task to investigate state-of-the-art NLI models on the presence of gender stereotypes using occupations. Our findings suggest that three models (BERT, RoBERTa, and BART) trained on MNLI and SNLI data-sets are significantly prone to gender-induced prediction errors. We also find that debiasing techniques such as augmenting the training dataset to ensure that it is a gender-balanced dataset can help reduce such bias in certain cases.
****
Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models
Justin Bayer,Maximilian Soelch,Atanas Mirchev,Baris Kayalibay,Patrick van der Smagt
Amortised inference enables scalable learning of sequential latent-variable models (LVMs) with the evidence lower bound (ELBO). In this setting, variational posteriors are often only partially conditioned. While the true posteriors depend, e.g., on the entire sequence of observations, approximate posteriors are only informed by past observations. This mimics the Bayesian filter---a mixture of smoothing posteriors. Yet, we show that the ELBO objective forces partially-conditioned amortised posteriors to approximate products of smoothing posteriors instead. Consequently, the learned generative model is compromised. We demonstrate these theoretical findings in three scenarios: traffic flow, handwritten digits, and aerial vehicle dynamics. Using fully-conditioned approximate posteriors, performance improves in terms of generative modelling and multi-step prediction.
****
CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning
Ossama Ahmed,Frederik Träuble,Anirudh Goyal,Alexander Neitz,Manuel Wuthrich,Yoshua Bengio,Bernhard Schölkopf,Stefan Bauer
Despite recent successes of reinforcement learning (RL), it remains a challenge for agents to transfer learned skills to related environments. To facilitate research addressing this problem, we proposeCausalWorld, a benchmark for causal structure and transfer learning in a robotic manipulation environment. The environment is a simulation of an open-source robotic platform, hence offering the possibility of sim-to-real transfer. Tasks consist of constructing 3D shapes from a set of blocks - inspired by how children learn to build complex structures. The key strength of CausalWorld is that it provides a combinatorial family of such tasks with common causal structure and underlying factors (including, e.g., robot and object masses, colors, sizes). The user (or the agent) may intervene on all causal variables, which allows for fine-grained control over how similar different tasks (or task distributions) are. One can thus easily define training and evaluation distributions of a desired difficulty level, targeting a specific form of generalization (e.g., only changes in appearance or object mass). Further, this common parametrization facilitates defining curricula by interpolating between an initial and a target task. While users may define their own task distributions, we present eight meaningful distributions as concrete benchmarks, ranging from simple to very challenging, all of which require long-horizon planning as well as precise low-level motor control. Finally, we provide baseline results for a subset of these tasks on distinct training curricula and corresponding evaluation protocols, verifying the feasibility of the tasks in this benchmark.
****
Addressing Extrapolation Error in Deep Offline Reinforcement Learning
Caglar Gulcehre,Sergio Gómez Colmenarejo,ziyu wang,Jakub Sygnowski,Thomas Paine,Konrad Zolna,Yutian Chen,Matthew Hoffman,Razvan Pascanu,Nando de Freitas
Reinforcement learning (RL) encompasses both online and offline regimes. Unlike its online counterpart, offline RL agents are trained using logged-data only, without interaction with the environment. Therefore, offline RL is a promising direction for real-world applications, such as healthcare, where repeated interaction with environments is prohibitive. However, since offline RL losses often involve evaluating state-action pairs not well-covered by training data, they can

suffer due to the errors introduced when the function approximator attempts to extrapolate those pairs' value. These errors can be compounded by bootstrapping when the function approximator overestimates, leading the value function to *grow unbounded*, thereby crippling learning. In this paper, we introduce a three-part solution to combat extrapolation errors: (i) behavior value estimation, (ii) ranking regularization, and (iii) reparametrization of the value function. We provide ample empirical evidence on the effectiveness of our method, showing state of the art performance on the RL Unplugged (RLU) ATARI dataset. Furthermore, we introduce new datasets for bsuite as well as partially observable DeepMind Lab environments, on which our method outperforms state of the art offline RL algorithms.

****

Certified robustness against physically-realizable patch attack via randomized cropping

Wan-Yi Lin,Fatemeh Sheikholeslami,jinghao shi,Leslie Rice,J Zico Kolter

This paper studies a certifiable defense against adversarial patch attacks on image classification. Our approach classifies random crops from the original image independently and the original image is classified as the vote over these crops. This process minimizes changes to the training process, as only the crop classification model needs to be trained, and can be trained in a standard manner without explicit adversarial training. Leveraging the fact that a patch attack can only influence some pixels of the image, we derive certified robustness bounds on the resulting classification. Our method is particularly effective when realistic physical transformations are applied to the adversarial patch, such as affine transformations. Such transformations occur naturally when an adversarial patch is physically introduced to a scene. Our method improves upon the current state of the art in defending against patch attacks on CIFAR10 and ImageNet, both in terms of certified accuracy and inference time.

****

Leveraging affinity cycle consistency to isolate factors of variation in learned representations

Kieran A Murphy,Varun Jampani,Srikumar Ramalingam,Ameesh Makadia

Identifying the dominant factors of variation across a dataset is a central goal of representation learning. Generative approaches lead to descriptions that are rich enough to recreate the data, but often only a partial description is needed to complete downstream tasks or to gain insights about the dataset. In this work, we operate in the setting where limited information is known about the data in the form of groupings, or set membership, and the task is to learn representations which isolate the factors of variation that are common across the groupings. Our key insight is the use of affinity cycle consistency (ACC) between the learned embeddings of images belonging to different sets. In contrast to prior work, we demonstrate that ACC can be applied with significantly fewer constraints on the factors of variation, across a remarkably broad range of settings, and without any supervision for half of the data. By curating datasets from Shapes3D, we quantify the effectiveness of ACC through mutual information between the learned representations and the known generative factors. In addition, we demonstrate the applicability of ACC to the tasks of digit style isolation and synthetic-to-real object pose transfer and compare to generative approaches utilizing the same supervision.

****

Transformer protein language models are unsupervised structure learners

Roshan Rao,Joshua Meier,Tom Sercu,Sergey Ovchinnikov,Alexander Rives

Unsupervised contact prediction is central to uncovering physical, structural, and functional constraints for protein structure determination and design. For decades, the predominant approach has been to infer evolutionary constraints from a set of related sequences. In the past year, protein language models have emerged as a potential alternative, but performance has fallen short of state-of-the-art approaches in bioinformatics. In this paper we demonstrate that Transformer attention maps learn contacts from the unsupervised language modeling objective.

We find the highest capacity models that have been trained to date already outp
erform a state-of-the-art unsupervised contact prediction pipeline, suggesting t
hese pipelines can be replaced with a single forward pass of an end-to-end model
.
****
Neural ODE Processes
Alexander Norcliffe,Cristian Bodnar,Ben Day,Jacob Moss,Pietro Liò
Neural Ordinary Differential Equations (NODEs) use a neural network to model the
 instantaneous rate of change in the state of a system. However, despite their a
pparent suitability for dynamics-governed time-series, NODEs present a few disad
vantages. First, they are unable to adapt to incoming data-points, a fundamental
 requirement for real-time applications imposed by the natural direction of time
. Second, time-series are often composed of a sparse set of measurements that co
uld be explained by many possible underlying dynamics. NODEs do not capture this
 uncertainty. In contrast, Neural Processes (NPs) are a new class of stochastic
processes providing uncertainty estimation and fast data-adaptation, but lack an
 explicit treatment of the flow of time. To address these problems, we introduce
 Neural ODE Processes (NDPs), a new class of stochastic processes determined by
a distribution over Neural ODEs. By maintaining an adaptive data-dependent distr
ibution over the underlying ODE, we show that our model can successfully capture
 the dynamics of low-dimensional systems from just a few data-points. At the sam
e time, we demonstrate that NDPs scale up to challenging high-dimensional time-s
eries with unknown latent dynamics such as rotating MNIST digits.
****
Quickest change detection for multi-task problems under unknown parameters
Firas Jarboui,Vianney Perchet
We consider the quickest change detection problem where both the parameters of p
re- and post- change distributions are unknown, which prevent the use of classic
al simple hypothesis testing. Without additional assumptions, optimal solutions
are not tractable as they rely on some minimax and robust variant of the objecti
ve. As a consequence, change points might be detected too late for practical app
lications (in economics, health care or maintenance for instance).
Other approaches solve a relaxed version of the problem through the use of parti
cular probability distributions or the use of domain knowledge.

We tackle this problem in the more complex Markovian case and we provide a new s
calable approximate algorithm with near optimal performance that runs in $\mathc
al{O}(1)$.
****
Fourier Representations for Black-Box Optimization over Categorical Variables
Hamid Dadkhahi,Jesus Rios,Karthikeyan Shanmugam,Payel Das
Optimization of real-world black-box functions defined over purely categorical v
ariables is an active area of research. In particular, optimization and design o
f biological sequences with specific functional or structural properties have a
profound impact in medicine, materials science, and biotechnology. Standalone ac
quisition methods, such as simulated annealing (SA) and Monte Carlo tree search
(MCTS), are typically used for such optimization problems. In order to improve t
he performance and sample efficiency of such acquisition methods, we propose to
use existing acquisition methods in conjunction with a surrogate model for the b
lack-box evaluations over purely categorical variables. To this end, we present
two different representations, a group-theoretic Fourier expansion and an abridg
ed one-hot encoded Boolean Fourier expansion. To learn such models, characters o
f each representation are considered as experts and their respective coefficient
s are updated via an exponential weight update rule each time the black box is e
valuated. Numerical experiments over synthetic benchmarks as well as real-world
RNA sequence optimization and design problems demonstrate the representational p
ower of the proposed methods, which achieve competitive or superior performance
compared to state-of-the-art counterparts, while improving the computational cos
t and/or sample efficiency substantially.
****

Implicit Acceleration of Gradient Flow in Overparameterized Linear Models
Salma Tarmoun,Guilherme França,Benjamin David Haeffele,Rene Vidal
We study the implicit acceleration of gradient flow in over-parameterized two-la yer linear models. We show that implicit acceleration emerges from a conservatio n law that constrains the dynamics to follow certain trajectories. More precisel y, gradient flow preserves the difference of the Gramian~matrices of the input a nd output weights and we show that the amount of acceleration depends on both th e magnitude of that difference (which is fixed at initialization) and the spectr um of the data. In addition, and generalizing prior work, we prove our results w ithout assuming small, balanced or spectral initialization for the weights, and establish interesting connections between the matrix factorization problem and R iccati type differential equations.
****

Dual-Tree Wavelet Packet CNNs for Image Classification
Hubert Leterme,Kévin Polisano,Valérie Perrier,Karteek Alahari
In this paper, we target an important issue of deep convolutional neural network s (CNNs) — the lack of a mathematical understanding of their properties. We pres ent an explicit formalism that is motivated by the similarities between trained CNN kernels and oriented Gabor filters for addressing this problem. The core ide a is to constrain the behavior of convolutional layers by splitting them into a succession of wavelet packet decompositions, which are modulated by freely-train ed mixture weights. We evaluate our approach with three variants of wavelet deco mpositions with the AlexNet architecture for image classification as an example. The first variant relies on the separable wavelet packet transform while the ot her two implement the 2D dual-tree real and complex wavelet packet transforms, t aking advantage of their feature extraction properties such as directional selec tivity and shift invariance. Our experiments show that we achieve the accuracy r ate of standard AlexNet, but with a significantly lower number of parameters, an d an interpretation of the network that is grounded in mathematical theory.
****

Hyperparameter Transfer Across Developer Adjustments
Danny Stoll,Jörg K.H. Franke,Diane Wagner,Simon Selg,Frank Hutter
After developer adjustments to a machine learning (ML) algorithm, how can the re sults of an old hyperparameter optimization (HPO) automatically be used to speed up a new HPO? This question poses a challenging problem, as developer adjustment s can change which hyperparameter settings perform well, or even the hyperparame ter search space itself. While many approaches exist that leverage knowledge obt ained on previous tasks, so far, knowledge from previous development steps remai ns entirely untapped. In this work, we remedy this situation and propose a new r esearch framework: hyperparameter transfer across adjustments (HT-AA). To lay a solid foundation for this research framework, we provide four simple HT-AA basel ine algorithms and eight benchmarks
changing various aspects of ML algorithms, their hyperparameter search spaces, a nd the neural architectures used. The best baseline, on average and depending on the budgets for the old and new HPO, reaches a given performance 1.2-3.6x faste r than a prominent HPO algorithm without transfer. As HPO is a crucial step in M L development but requires extensive computational resources, this speedup would lead to faster development cycles, lower costs, and reduced environmental impac ts. To make these benefits available to ML developers off-the-shelf and to facil itate future research on HT-AA, we provide python packages for our baselines and benchmarks.
****

The role of Disentanglement in Generalisation
Milton Llera Montero,Casimir JH Ludwig,Rui Ponte Costa,Gaurav Malhotra,Jeffrey B owers
Combinatorial generalisation — the ability to understand and produce novel combi nations of familiar elements — is a core capacity of human intelligence that cur rent AI systems struggle with. Recently, it has been suggested that learning dis entangled representations may help address this problem. It is claimed that such representations should be able to capture the compositional structure of the wo

rld which can then be combined to support combinatorial generalisation. In this study, we systematically tested how the degree of disentanglement affects various forms of generalisation, including two forms of combinatorial generalisation that varied in difficulty. We trained three classes of variational autoencoders (VAEs) on two datasets on an unsupervised task by excluding combinations of generative factors during training. At test time we ask the models to reconstruct the missing combinations in order to measure generalisation performance. Irrespective of the degree of disentanglement, we found that the models supported only weak combinatorial generalisation. We obtained the same outcome when we directly input perfectly disentangled representations as the latents, and when we tested a model on a more complex task that explicitly required independent generative factors to be controlled. While learning disentangled representations does improve interpretability and sample efficiency in some downstream tasks, our results suggest that they are not sufficient for supporting more difficult forms of generalisation.

****

Learning to live with Dale's principle: ANNs with separate excitatory and inhibitory units

Jonathan Cornford,Damjan Kalajdzievski,Marco Leite,Amélie Lamarquette,Dimitri Michael Kullmann,Blake Aaron Richards

The units in artificial neural networks (ANNs) can be thought of as abstractions of biological neurons, and ANNs are increasingly used in neuroscience research. However, there are many important differences between ANN units and real neurons. One of the most notable is the absence of Dale's principle, which ensures that biological neurons are either exclusively excitatory or inhibitory. Dale's principle is typically left out of ANNs because its inclusion impairs learning. This is problematic, because one of the great advantages of ANNs for neuroscience research is their ability to learn complicated, realistic tasks. Here, by taking inspiration from feedforward inhibitory interneurons in the brain we show that we can develop ANNs with separate populations of excitatory and inhibitory units that learn just as well as standard ANNs. We call these networks Dale's ANNs (DANNs). We present two insights that enable DANNs to learn well: (1) DANNs are related to normalization schemes, and can be initialized such that the inhibition centres and standardizes the excitatory activity, (2) updates to inhibitory neuron parameters should be scaled using corrections based on the Fisher Information matrix. These results demonstrate how ANNs that respect Dale's principle can be built without sacrificing learning performance, which is important for future work using ANNs as models of the brain. The results may also have interesting implications for how inhibitory plasticity in the real brain operates.

****

A Simple Approach To Define Curricula For Training Neural Networks

Vinu Sankar Sadasivan,Anirban Dasgupta

In practice, sequence of mini-batches generated by uniform sampling of examples from the entire data is used for training neural networks. Curriculum learning is a training strategy that sorts the training examples by their difficulty and gradually exposes them to the learner. In this work, we propose two novel curriculum learning algorithms and empirically show their improvements in performance with convolutional and fully-connected neural networks on multiple real image datasets. Our dynamic curriculum learning algorithm tries to reduce the distance between the network weight and an optimal weight at any training step by greedily sampling examples with gradients that are directed towards the optimal weight. The curriculum ordering determined by our dynamic algorithm achieves a training speedup of $\sim 45\%$ in our experiments. We also introduce a new task-specific curriculum learning strategy that uses statistical measures such as standard deviation and entropy values to score the difficulty of data points in natural image datasets. We show that this new approach yields a mean training speedup of $\sim 43\%$ in the experiments we perform. Further, we also use our algorithms to learn why curriculum learning works. Based on our study, we argue that curriculum learning removes noisy examples from the initial phases of training, and gradually exposes them to the learner acting like a regularizer that helps in improvin

g the generalization ability of the learner.
****
SALD: Sign Agnostic Learning with Derivatives
Matan Atzmon,Yaron Lipman
Learning 3D geometry directly from raw data, such as point clouds, triangle soup s, or unoriented meshes is still a challenging task that feeds many downstream c omputer vision and graphics applications.

In this paper, we introduce SALD: a method for learning implicit neural represen tations of shapes directly from raw data. We generalize sign agnostic learning (SAL) to include derivatives: given an unsigned distance function to the input ra w data, we advocate a novel sign agnostic regression loss, incorporating both po intwise values and gradients of the unsigned distance function. Optimizing this loss leads to a signed implicit function solution, the zero level set of which i s a high quality and valid manifold approximation to the input 3D data. The moti vation behind SALD is that incorporating derivatives in a regression loss leads to a lower sample complexity, and consequently better fitting. In addition, we p rovide empirical evidence, as well as theoretical motivation in 2D that SAL enjo ys a minimal surface property, favoring minimal area solutions. More importantly , we are able to show that this property still holds for SALD, i.e., with deriv atives included.

We demonstrate the efficacy of SALD for shape space learning on two challenging datasets: ShapeNet that contains inconsistent orientation and non-manifold meshe s, and D-Faust that contains raw 3D scans (triangle soups). On both these datase ts, we present state-of-the-art results.
****
Better sampling in explanation methods can prevent dieselgate-like deception
Domen Vreš,Marko Robnik Šikonja
Machine learning models are used in many sensitive areas where besides predictiv e accuracy their comprehensibility is also important. Interpretability of predic tion models is necessary to determine their biases and causes of errors, and is a necessary prerequisite for users' confidence. For complex state-of-the-art bla ck-box models post-hoc model-independent explanation techniques are an establish ed solution. Popular and effective techniques, such as IME, LIME, and SHAP, use perturbation of instance features to explain individual predictions. Recently, Slack et al. (2020) put their robustness into question by showing that their out comes can be manipulated due to poor perturbation sampling employed. This weakne ss would allow dieselgate type cheating of owners of sensitive models who could deceive inspection and hide potentially unethical or illegal biases existing in their predictive models. This could undermine public trust in machine learning m odels and give rise to legal restrictions on their use.

We show that better sampling in these explanation methods prevents malicious man ipulations. The proposed sampling uses data generators that learn the training s et distribution and generate new perturbation instances much more similar to the training set. We show that the improved sampling increases the robustness of th e LIME and SHAP, while previously untested method IME is already the most robust of all.
****
Ringing ReLUs: Harmonic Distortion Analysis of Nonlinear Feedforward Networks
Christian H.X. Ali Mehmeti-Göpel,David Hartmann,Michael Wand
In this paper, we apply harmonic distortion analysis to understand the effect of nonlinearities in the spectral domain. Each nonlinear layer creates higher-freq uency harmonics, which we call "blueshift", whose magnitude increases with netwo rk depth, thereby increasing the “roughness” of the output landscape. Unlike dif ferential models (such as vanishing gradients, sharpness), this provides a more global view of how network architectures behave across larger areas of their par ameter domain. For example, the model predicts that residual connections are abl e to counter the effect by dampening corresponding higher frequency modes. We em

pirically verify the connection between blueshift and architectural choices, and provide evidence for a connection with trainability.
****
A generalized probability kernel on discrete distributions and its application in two-sample test
Le Niu
We propose a generalized probability kernel(GPK) on discrete distributions with finite support. This probability kernel, defined as kernel between distributions instead of samples, generalizes the existing discrepancy statistics such as maximum mean discrepancy(MMD) as well as probability product kernels, and extends to more general cases. For both existing and newly proposed statistics, we estimate them through empirical frequency and illustrate the strategy to analyze the resulting bias and convergence bounds. We further propose power-MMD, a natural extension of MMD in the framework of GPK, illustrating its usage for the task of two-sample test. Our work connects the fields of discrete distribution-property estimation and kernel-based hypothesis test, which might shed light on more new possibilities.
****
End-to-End on-device Federated Learning: A case study
Hongyi Zhang,Jan Bosch,Helena Holmström Olsson
With the development of computation capability in devices, companies are eager to utilize ML/DL methods to improve their service quality. However, with traditional Machine Learning approaches, companies need to build up a powerful data center to collect data and perform centralized model training, which turns out to be expensive and inefficient. Federated Learning has been introduced to solve this challenge. Because of its characteristics such as model-only exchange and parallel training, the technique can not only preserve user data privacy but also accelerate model training speed. In this paper, we introduce an approach to end-to-end on-device Machine Learning by utilizing Federated Learning. We validate our approach with an important industrial use case, the wheel steering angle prediction in the field of autonomous driving. Our results show that Federated Learning can significantly improve the quality of local edge models and reach the same accuracy level as compared to the traditional centralized Machine Learning approach without its negative effects. Furthermore, Federated Learning can accelerate model training speed and reduce the communication overhead, which proves that this approach has great strength when deploying ML/DL components to real-world embedded systems.
****
On the Explicit Role of Initialization on the Convergence and Generalization Properties of Overparametrized Linear Networks
Hancheng Min,Salma Tarmoun,Rene Vidal,Enrique Mallada
Neural networks trained via gradient descent with random initialization and without any regularization enjoy good generalization performance in practice despite being highly overparametrized. A promising direction to explain this phenomenon is the \emph{Neural Tangent Kernel} (NTK), which characterizes the implicit regularization effect of gradient flow/descent on infinitely wide neural networks with random initialization. However, a non-asymptotic analysis that connects generalization performance, initialization, and optimization for finite width networks remains elusive. In this paper, we present a novel analysis of overparametrized single-hidden layer linear networks, which formally connects initialization, optimization, and overparametrization with generalization performance. We exploit the fact that gradient flow preserves a certain matrix that characterizes the \emph{imbalance} of the network weights, to show that the squared loss converges exponentially at a rate that depends on the level of imbalance of the initialization. Such guarantees on the convergence rate allow us to show that large hidden layer width, together with (properly scaled) random initialization, implicitly constrains the dynamics of the network parameters to be close to a low-dimensional manifold. In turn, minimizing the loss over this manifold leads to solutions with good generalization, which correspond to the min-norm solution in the linear case. Finally, we derive a novel $\mathcal{O}(h^{-1/2})$ upper-bound on the

operator norm distance between the trained network and the min-norm solution, where h is the hidden layer width.

****

CANVASEMB: Learning Layout Representation with Large-scale Pre-training for Graphic Design

Yuxi Xie,Danqing Huang,Jinpeng Wang,Chin-Yew Lin

Layout representation, which models visual elements in a canvas and their interrelations, plays a crucial role in graphic design intelligence.
With a large variety of layout designs and the unique characteristic of layouts that visual elements are defined as a list of categorical (e.g. shape type) and numerical (e.g. position and size) properties, it is challenging to learn a general and compact representation with limited data. Inspired by the recent success of self-supervised pre-training techniques in various natural language processing tasks, in this paper, we propose CanvasEmb (Canvas Embedding), which pre-trains deep representation from unlabeled graphic designs by jointly conditioning on all the context elements in the same canvas, with a multi-dimensional feature encoder and a multi-task learning objective. The pre-trained CanvasEmb model can be fine-tuned with just one additional output layer and with a small size of training data to create models for a wide range of downstream tasks. We verify our approach with presentation slides data. We construct a large-scale dataset with more than one million slides, and propose two novel layout understanding tasks with human labeling sets, namely element role labeling and image captioning. Evaluation results on these two tasks show that our model with fine-tuning achieves state-of-the-art performances. Furthermore, we conduct a deep analysis aiming to understand the modeling mechanism of CanvasEmb, and demonstrate its great potential use on more applications such as layout auto completion and layout retrieval.

****

Federated Learning of a Mixture of Global and Local Models

Filip Hanzely,Peter Richtarik

We propose a new optimization formulation for training federated learning models. The standard formulation has the form of an empirical risk minimization problem constructed to find a single global model trained from the private data stored across all participating devices. In contrast, our formulation seeks an explicit trade-off between this traditional global model and the local models, which can be learned by each device from its own private data without any communication. Further, we develop several efficient variants of SGD (with and without partial participation and with and without variance reduction) for solving the new formulation and prove communication complexity guarantees. Notably, our methods are similar but not identical to federated averaging / local SGD, thus shedding some light on the essence of the elusive method. In particular, our methods do not perform full averaging steps and instead merely take steps towards averaging. We argue for the benefits of this new paradigm for federated learning.

****

Learning to communicate through imagination with model-based deep multi-agent reinforcement learning

Arnu Pretorius,Scott Cameron,Andries Petrus Smit,Elan van Biljon,Lawrence Francis,Femi Azeez,Alexandre Laterre,Karim Beguir

The human imagination is an integral component of our intelligence. Furthermore, the core utility of our imagination is deeply coupled with communication. Language, argued to have been developed through complex interaction within growing collective societies serves as an instruction to the imagination, giving us the ability to share abstract mental representations and perform joint spatiotemporal planning. In this paper, we explore communication through imagination with multi-agent reinforcement learning. Specifically, we develop a model-based approach where agents jointly plan through recurrent communication of their respective predictions of the future. Each agent has access to a learned world model capable of producing model rollouts of future states and predicted rewards, conditioned on the actions sampled from the agent's policy. These rollouts are then encoded i

nto messages and used to learn a communication protocol during training via diff
erentiable message passing. We highlight the benefits of our model-based approac
h, compared to a set of strong baselines, by developing a set of specialised exp
eriments using novel as well as well-known multi-agent environments.
****
Unpacking Information Bottlenecks: Surrogate Objectives for Deep Learning
Andreas Kirsch,Clare Lyle,Yarin Gal
The Information Bottleneck principle offers both a mechanism to explain how deep
 neural networks train and generalize, as well as a regularized objective with w
hich to train models. However, multiple competing objectives are proposed in the
 literature, and the information-theoretic quantities used in these objectives a
re difficult to compute for large deep neural networks, which in turn limits the
ir use as a training objective. In this work, we review these quantities, compar
e and unify previously proposed objectives, which allows us to develop surrogate
 objectives more friendly to optimization without relying on cumbersome tools su
ch as density estimation. We find that these surrogate objectives allow us to ap
ply the information bottleneck to modern neural network architectures. We demons
trate our insights on MNIST, CIFAR-10 and ImageNette with modern DNN architectur
es (ResNets).
****
Putting Theory to Work: From Learning Bounds to Meta-Learning Algorithms
Quentin Bouniot,Ievgen Redko,Romaric Audigier,Angélique Loesch,Amaury Habrard
Most of existing deep learning models rely on excessive amounts of labeled train
ing data in order to achieve state-of-the-art results, even though these data ca
n be hard or costly to get in practice. One attractive alternative is to learn w
ith little supervision, commonly referred to as few-shot learning (FSL), and, in
 particular, meta-learning that learns to learn with few data from related tasks
. Despite the practical success of meta-learning, many of its algorithmic soluti
ons proposed in the literature are based on sound intuitions, but lack a solid t
heoretical analysis of the expected performance on the test task. In this paper,
 we review the recent advances in meta-learning theory and show how they can be
used in practice both to better understand the behavior of popular meta-learning
 algorithms and to improve their generalization capacity. This latter is achieve
d by integrating the theoretical assumptions ensuring efficient meta-learning in
 the form of regularization terms into several popular meta-learning algorithms
for which we provide a large study of their behavior on classic few-shot classif
ication benchmarks. To the best of our knowledge, this is the first contribution
 that puts the most recent learning bounds of meta-learning theory into practice
 for the popular task of few-shot classification.
****
Does injecting linguistic structure into language models lead to better alignmen
t with brain recordings?
Mostafa Abdou,Ana Valeria González,Mariya K Toneva,Daniel Hershcovich,Anders Søg
aard
Neuroscientists evaluate deep neural networks for natural language processing as
 possible candidate models for how language is processed in the brain. These mod
els are often trained without explicit linguistic supervision, but have been sho
wn to learn some linguistic structure in the absence of such supervision (Mannin
g et. al, 2020), potentially questioning the relevance of symbolic linguistic th
eories in modeling such cognitive processes (Warstadt & Bowman, 2020). We evalua
te across two fMRI datasets whether language models align better with brain reco
rdings, if their attention is biased by annotations from syntactic or semantic f
ormalisms. Using structure from dependency or minimal recursion semantic annotat
ions, we find alignments improve significantly for one of the datasets. For anot
her dataset, we see more mixed results. We present an extensive analysis of thes
e results. Our proposed approach enables the evaluation of more targeted hypothe
ses about the composition of meaning in the brain, expanding the range of possib
le scientific inferences a neuroscientist could make, and opens up new opportuni
ties for cross-pollination between computational neuroscience and linguistics.

****
CoCon: A Self-Supervised Approach for Controlled Text Generation
Alvin Chan,Yew-Soon Ong,Bill Pung,Aston Zhang,Jie Fu
Pretrained Transformer-based language models (LMs) display remarkable natural language generation capabilities. With their immense potential, controlling text generation of such LMs is getting attention. While there are studies that seek to control high-level attributes (such as sentiment and topic) of generated text, there is still a lack of more precise control over its content at the word- and phrase-level. Here, we propose Content-Conditioner (CoCon) to control an LM's output text with a content input, at a fine-grained level. In our self-supervised approach, the CoCon block learns to help the LM complete a partially-observed text sequence by conditioning with content inputs that are withheld from the LM. Through experiments, we show that CoCon can naturally incorporate target content into generated texts and control high-level text attributes in a zero-shot manner.
****
αVIL: Learning to Leverage Auxiliary Tasks for Multitask Learning
Rafael Kourdis,Gabriel Gordon-Hall,Philip John Gorinski
Multitask Learning is a Machine Learning paradigm that aims to train a range of (usually related) tasks with the help of a shared model. While the goal is often to improve the joint performance of all training tasks, another approach is to focus on the performance of a specific target task, while treating the remaining ones as auxiliary data from which to possibly leverage positive transfer towards the target during training. In such settings, it becomes important to estimate the positive or negative influence auxiliary tasks will have on the target. While many ways have been proposed to estimate task weights before or during training they typically rely on heuristics or extensive search of the weighting space. We propose a novel method called α-Variable Importance Learning (αVIL) that is able to adjust task weights dynamically during model training, by making direct use of task-specific updates of the underlying model's parameters between training epochs. Experiments indicate that αVIL is able to outperform other Multitask Learning approaches in a variety of settings. To our knowledge, this is the first attempt at making direct use of model updates for task weight estimation.
****
Tracking the progress of Language Models by extracting their underlying Knowledge Graphs
Carlos Aspillaga,Marcelo Mendoza,Alvaro Soto
The state of the art of language models, previously dominated by pre-trained word embeddings, is now being pushed forward by large pre-trained contextual representations. This success has driven growing interest to understand what these models encode inside their inner workings. Despite this, understanding their semantic skills has been elusive, often leading to unsuccessful, non-conclusive, or contradictory results among different works. In this work, we define a probing classifier that we use to extract the underlying knowledge graph of nine of the currently most influential language models, including word embeddings, context encoders, and text generators. This probe is based on concept relatedness, grounded on WordNet. Our results show that this knowledge is present in all the models, but has several inaccuracies. Furthermore, we show that the different pre-training strategies and architectures lead to different model biases. We conduct a systematic evaluation to discover specific factors that explain why some concepts are challenging for the different families of models. We hope our insights will motivate the future development of models that capture concepts more precisely.
****
Bidirectional Variational Inference for Non-Autoregressive Text-to-Speech
Yoonhyung Lee,Joongbo Shin,Kyomin Jung
Although early text-to-speech (TTS) models such as Tacotron 2 have succeeded in generating human-like speech, their autoregressive architectures have several limitations: (1) They require a lot of time to generate a mel-spectrogram consisti

ng of hundreds of steps. (2) The autoregressive speech generation shows a lack o f robustness due to its error propagation property. In this paper, we propose a novel non-autoregressive TTS model called BVAE-TTS, which eliminates the archite ctural limitations and generates a mel-spectrogram in parallel. BVAE-TTS adopts a bidirectional-inference variational autoencoder (BVAE) that learns hierarchica l latent representations using both bottom-up and top-down paths to increase its expressiveness. To apply BVAE to TTS, we design our model to utilize text infor mation via an attention mechanism. By using attention maps that BVAE-TTS generat es, we train a duration predictor so that the model uses the predicted duration of each phoneme at inference. In experiments conducted on LJSpeech dataset, we s how that our model generates a mel-spectrogram 27 times faster than Tacotron 2 w ith similar speech quality. Furthermore, our BVAE-TTS outperforms Glow-TTS, whic h is one of the state-of-the-art non-autoregressive TTS models, in terms of both speech quality and inference speed while having 58% fewer parameters.
****

Uncertainty Calibration Error: A New Metric for Multi-Class Classification
Max-Heinrich Laves,Sontje Ihler,Karl-Philipp Kortmann,Tobias Ortmaier
Various metrics have recently been proposed to measure uncertainty calibration o f deep models for classification. However, these metrics either fail to capture miscalibration correctly or lack interpretability. We propose to use the normali zed entropy as a measure of uncertainty and derive the Uncertainty Calibration E rror (UCE), a comprehensible calibration metric for multi-class classification. In our experiments, we focus on uncertainty from variational Bayesian inference methods and compare UCE to established calibration errors on the task of multi-c lass image classification. UCE avoids several pathologies of other metrics, but does not sacrifice interpretability. It can be used for regularization to improv e calibration during training without penalizing predictions with justified high confidence.
****

Learning continuous-time PDEs from sparse data with graph neural networks
Valerii Iakovlev,Markus Heinonen,Harri Lähdesmäki
The behavior of many dynamical systems follow complex, yet still unknown partial differential equations (PDEs). While several machine learning methods have been proposed to learn PDEs directly from data, previous methods are limited to disc rete-time approximations or make the limiting assumption of the observations arr iving at regular grids. We propose a general continuous-time differential model for dynamical systems whose governing equations are parameterized by message pas sing graph neural networks. The model admits arbitrary space and time discretiza tions, which removes constraints on the locations of observation points and time intervals between the observations. The model is trained with continuous-time a djoint method enabling efficient neural PDE inference. We demonstrate the model' s ability to work with unstructured grids, arbitrary time steps, and noisy obser vations. We compare our method with existing approaches on several well-known ph ysical systems that involve first and higher-order PDEs with state-of-the-art pr edictive performance.
****

NAS-Bench-ASR: Reproducible Neural Architecture Search for Speech Recognition
Abhinav Mehrotra,Alberto Gil C. P. Ramos,Sourav Bhattacharya,Łukasz Dudziak,Ravi chander Vipperla,Thomas Chau,Mohamed S Abdelfattah,Samin Ishtiaq,Nicholas Donald Lane
Powered by innovations in novel architecture design, noise tolerance techniques and increasing model capacity, Automatic Speech Recognition (ASR) has made giant strides in reducing word-error-rate over the past decade. ASR models are often trained with tens of thousand hours of high quality speech data to produce state -of-the-art (SOTA) results. Industry-scale ASR model training thus remains compu tationally heavy and time-consuming, and consequently has attracted little atten tion in adopting automatic techniques. On the other hand, Neural Architecture Se arch (NAS) has gained a lot of interest in the recent years thanks to its succes ses in discovering efficient architectures, often outperforming handcrafted alte rnatives. However, by changing the standard training process into a bi-level opt

imisation problem, NAS approaches often require significantly more time and comp utational power compared to single-model training, and at the same time increase complexity of the overall process. As a result, NAS has been predominately appl ied to problems which do not require as extensive training as ASR, and even then reproducibility of NAS algorithms is often problematic. Lately, a number of ben chmark datasets has been introduced to address reproducibility issues by pro- vi ding NAS researchers with information about performance of different models obta ined through exhaustive evaluation. However, these datasets focus mainly on comp uter vision and NLP tasks and thus suffer from limited coverage of application d omains. In order to increase diversity in the existing NAS benchmarks, and at th e same time provide systematic study of the effects of architectural choices for ASR, we release NAS-Bench-ASR – the first NAS benchmark for ASR models. The dat aset consists of 8, 242 unique models trained on the TIMIT audio dataset for thr ee different target epochs, and each starting from three different initializatio ns. The dataset also includes runtime measurements of all the models on a divers e set of hardware platforms. Lastly, we show that identified good cell structure s in our search space for TIMIT transfer well to a much larger LibriSpeech datas et.

****

MQES: Max-Q Entropy Search for Efficient Exploration in Continuous Reinforcement Learning

Jinyi Liu,Zhi Wang,Jianye HAO,YAN ZHENG

The principle of optimism in the face of (aleatoric and epistemic) uncertainty h as been utilized to design efficient exploration strategies for Reinforcement Le arning (RL). Different from most prior work targeting at discrete action space, we propose a generally information-theoretic exploration principle called Max-Q Entropy Search (MQES) for continuous RL algorithms.

MQES formulates the exploration policy to maximize the information about the glo bally optimal distribution of Q function, which could explore optimistically a nd avoid over-exploration by recognizing the epistemic and aleatoric uncertainty , respectively. To make MQES practically tractable, we firstly incorporate distr ibutional and ensemble Q function approximations to MQES, which could formulat e the epistemic and aleatoric uncertainty accordingly. Then, we introduce a cons traint to stabilize the training and solve the constrained MQES problem to deriv e the exploration policy in closed form. Empirical evaluations show that MQES ou tperforms state-of-the-art algorithms on Mujoco environments.

****

Task-similarity Aware Meta-learning through Nonparametric Kernel Regression

Arun Venkitaraman,Anders Hansson,Bo Wahlberg

This paper investigates the use of nonparametric kernel-regression to obtain a t ask- similarity aware meta-learning algorithm. Our hypothesis is that the use of task- similarity helps meta-learning when the available tasks are limited and m ay contain outlier/ dissimilar tasks. While existing meta-learning approaches im plicitly assume the tasks as being similar, it is generally unclear how this tas k-similarity could be quantified and used in the learning. As a result, most pop ular meta- learning approaches do not actively use the similarity/dissimilarity between the tasks, but rely on availability of huge number of tasks for their wo rking. Our contribution is a novel framework for meta-learning that explicitly u ses task-similarity in the form of kernels and an associated meta-learning algor ithm. We model the task-specific parameters to belong to a reproducing kernel Hi lbert space where the kernel function captures the similarity across tasks. The proposed algorithm iteratively learns a meta-parameter which is used to assign a task-specific descriptor for every task. The task descriptors are then used to quantify the task-similarity through the kernel function. We show how our approa ch conceptually generalizes the popular meta-learning approaches of model-agnost ic meta-learning (MAML) and Meta-stochastic gradient descent (Meta-SGD) approach es. Numerical experiments with regression and classification tasks show that our algorithm outperforms these approaches when the number of tasks is limited, eve n in the presence of out- lier or dissimilar tasks. This supports our hypothesis that task-similarity helps improve the meta-learning performance in task-limite

d and adverse settings.
****
Mutual Calibration between Explicit and Implicit Deep Generative Models
Qitian Wu,Rui Gao,Hongyuan Zha
Deep generative models are generally categorized into explicit models and implic it models. The former defines an explicit density form that allows likelihood in ference; while the latter targets a flexible transformation from random noise to generated samples. To take full advantages of both models, we propose Stein Bri dging, a novel joint training framework that connects an explicit (unnormalized) density estimator and an implicit sample generator via Stein discrepancy. We sh ow that the Stein bridge 1) induces novel mutual regularization via kernel Sobol ev norm penalization and Moreau-Yosida regularization, and 2) stabilizes the tra ining dynamics. Empirically, we demonstrate that Stein Bridging can facilitate t he density estimator to accurately identify data modes and guide the sample gene rator to output more high-quality samples especially when the training samples a re contaminated or limited.
****
Collective Robustness Certificates: Exploiting Interdependence in Graph Neural N etworks
Jan Schuchardt,Aleksandar Bojchevski,Johannes Gasteiger,Stephan Günnemann
In tasks like node classification, image segmentation, and named-entity recognit ion we have a classifier that simultaneously outputs multiple predictions (a vec tor of labels) based on a single input, i.e. a single graph, image, or document respectively. Existing adversarial robustness certificates consider each predict ion independently and are thus overly pessimistic for such tasks. They implicitl y assume that an adversary can use different perturbed inputs to attack differen t predictions, ignoring the fact that we have a single shared input. We propose the first collective robustness certificate which computes the number of predict ions that are simultaneously guaranteed to remain stable under perturbation, i.e . cannot be attacked. We focus on Graph Neural Networks and leverage their local ity property - perturbations only affect the predictions in a close neighborhood - to fuse multiple single-node certificates into a drastically stronger collect ive certificate. For example, on the Citeseer dataset our collective certificate for node classification increases the average number of certifiable feature per turbations from 7 to 351.

****
Zero-shot Fairness with Invisible Demographics
Thomas Kehrenberg,Viktoriia Sharmanska,Myles Scott Bartlett,Novi Quadrianto
In a statistical notion of algorithmic fairness, we partition individuals into g roups based on some key demographic factors such as race and gender, and require that some statistics of a classifier be approximately equalized across those gr oups. Current approaches require complete annotations for demographic factors, o r focus on an abstract worst-off group rather than demographic groups. In this p aper, we consider the setting where the demographic factors are only partially a vailable. For example, we have training examples for white-skinned and dark-skin ned males, and white-skinned females, but we have zero examples for dark-skinned females. We could also have zero examples for females regardless of their skin colors. Without additional knowledge, it is impossible to directly control the d iscrepancy of the classifier's statistics for those invisible groups. We develop a disentanglement algorithm that splits a representation of data into a compone nt that captures the demographic factors and another component that is invariant to them based on a context dataset. The context dataset is much like the deploy ment dataset, it is unlabeled but it contains individuals from all demographics including the invisible. We cluster the context set, equalize the cluster size t o form a "perfect batch", and use it as a supervision signal for the disentangle ment. We propose a new discriminator loss based on a learnable attention mechani sm to distinguish a perfect batch from a non-perfect one. We evaluate our approa ch on standard classification benchmarks and show that it is indeed possible to protect invisible demographics.

****
Rotograd: Dynamic Gradient Homogenization for Multitask Learning
Adrián Javaloy,Isabel Valera
GradNorm (Chen et al., 2018) is a broadly used gradient-based approach for training multitask networks, where different tasks share, and thus compete during learning, for the network parameters. GradNorm eases the fitting of all individual tasks by dynamically equalizing the contribution of each task to the overall gradient magnitude. However, it does not prevent the individual tasks' gradients from conflicting, i.e., pointing towards opposite directions, and thus resulting in a poor multitask performance. In this work we propose Rotograd, an extension to GradNorm that addresses this problem by dynamically homogenizing not only the gradient magnitudes but also their directions across tasks. For this purpose,Rotograd adds a layer of task-specific rotation matrices that aligns all the task gradients. Importantly, we then analyze Rotograd (and its predecessor) through the lens of game theory, providing theoretical guarantees on the algorithm stability and convergence. Finally, our experiments on several real-world datasets and network architectures show that Rotograd outperforms previous approaches for multitask learning.

****
UNSUPERVISED ANOMALY DETECTION FROM SEMANTIC SIMILARITY SCORES
Nima Rafiee,Rahil Gholamipoor,Markus Kollmann
In this paper we present SemSAD, a simple and generic framework for detecting examples that lie out-of-distribution (OOD) for a given training set. The approach is based on learning a semantic similarity measure to find for a given test example the semantically closest example in the training set and then using a discriminator to classify whether the two examples show sufficient semantic dissimilarity such that the test example can be rejected as OOD. We are able to outperform previous approaches for anomaly, novelty, or out-of-distribution detection in the visual domain by a large margin. In particular we obtain AUROC values close to one for the challenging task of detecting examples from CIFAR-10 as out-of-distribution given CIFAR-100 as in-distribution, without making use of label information.
****
Legendre Deep Neural Network (LDNN) and its application for approximation of nonlinear Volterra–Fredholm–Hammerstein integral equations
Kourosh Parand,Zeinab Hajimohammadi,Ali Ghodsi
Various phenomena in biology, physics, and engineering are modeled by differential equations. These differential equations including partial differential equations and ordinary differential equations can be converted and represented as integral equations. In particular, Volterra–Fredholm–Hammerstein integral equations are the main type of these integral equations and researchers are interested in investigating and solving these equations. In this paper, we propose Legendre Deep Neural Network (LDNN) for solving nonlinear Volterra–Fredholm–Hammerstein integral equations (V-F-H-IEs). LDNN utilizes Legendre orthogonal polynomials as activation functions of the Deep structure. We present how LDNN can be used to solve nonlinear V-F-H-IEs. We show using the Gaussian quadrature collocation method in combination with LDNN results in a novel numerical solution for nonlinear V-F-H-IEs. Several examples are given to verify the performance and accuracy of LDNN.
****
Rethinking Parameter Counting: Effective Dimensionality Revisited
Gregory Benton,Wesley Maddox,Andrew Gordon Wilson
Neural networks appear to have mysterious generalization properties when using parameter counting as a proxy for complexity. Indeed, neural networks often have many more parameters than there are data points, yet still provide good generalization performance. Moreover, when we measure generalization as a function of parameters, we see double descent behaviour, where the test error decreases, increases, and then again decreases. We show that many of these properties become und

erstandable when viewed through the lens of effective dimensionality, which meas ures the dimensionality of the parameter space determined by the data. We relate effective dimensionality to posterior contraction in Bayesian deep learning, mo del selection, width-depth tradeoffs, double descent, and functional diversity i n loss surfaces, leading to a richer understanding of the interplay between par ameters and functions in deep models. We also show that effective dimensionality compares favourably to alternative norm- and flatness- based generalization mea sures.
****
Decentralized SGD with Asynchronous, Local and Quantized Updates
Giorgi Nadiradze,Amirmojtaba Sabour,Peter Davies,Ilia Markov,Shigang Li,Dan Alis tarh
The ability to scale distributed optimization to large node counts has been one of the main enablers of recent progress in machine learning. To this end, severa l techniques have been explored, such as asynchronous, quantized and decentrali zed communication--which significantly reduce the impact of communication and sy nchronization, as well as the ability for nodes to perform several local model u pdates before communicating--which reduces the frequency of communication.
In this paper, we show that these techniques, which have so far largely been co nsidered independently, can be jointly leveraged to minimize distribution cost f or training neural network models via stochastic gradient descent (SGD).
We consider a setting with minimal coordination: we have a large number of nodes on a communication graph, each with a local subset of data, performing independ ent SGD updates onto their local models. After some number of local updates, eac h node chooses an interaction partner uniformly at random from its neighbors, an d averages a (possibly quantized) version of its local model with the neighbor's model.
Our first contribution is in proving that, even under such a relaxed setting, SG D can still be guaranteed to converge under standard assumptions. The proof is b ased on a new connection with parallel load-balancing processes, and improves ex isting techniques by handling decentralization, asynchrony, quantization, and lo cal updates, into a single framework, and bounding their impact.
On the practical side, we implement variants of our algorithm and deploy them on to distributed environments, and show that they can successfully converge and sc ale for large-scale neural network training tasks, matching or even slightly imp roving the accuracy of previous methods.
****
Regression from Upper One-side Labeled Data
Takayuki Katsuki
We address a regression problem from weakly labeled data that are correctly labe led only above a regression line, i.e., upper one-side labeled data.
The label values of the data are the results of sensing the magnitude of some ph enomenon.
In this case, the labels often contain missing or incomplete observations whose values are lower than those of correct observations and are also usually lower t han the regression line. It follows that data labeled with lower values than the estimations of a regression function (lower-side data) are mixed with data that should originally be labeled above the regression line (upper-side data).
When such missing label observations are observed in a non-negligible amount, we thus should assume our lower-side data to be unlabeled data that are a mix of o riginal upper- and lower-side data.
We formulate a regression problem from these upper-side labeled and lower-side u nlabeled data. We then derive a learning algorithm in an unbiased and consistent manner to ordinary regression that is learned from data labeled correctly in bo th upper- and lower-side cases. Our key idea is that we can derive a gradient th at requires only upper-side data and unlabeled data as the equivalent expression of that for ordinary regression. We additionally found that a specific class of losses enables us to learn unbiased solutions practically. In numerical experim ents on synthetic and real-world datasets, we demonstrate the advantages of our algorithm.

****
Adversarially Guided Actor-Critic
Yannis Flet-Berliac,Johan Ferret,Olivier Pietquin,Philippe Preux,Matthieu Geist
Despite definite success in deep reinforcement learning problems, actor-critic algorithms are still confronted with sample inefficiency in complex environments, particularly in tasks where efficient exploration is a bottleneck. These methods consider a policy (the actor) and a value function (the critic) whose respective losses are built using different motivations and approaches. This paper introduces a third protagonist: the adversary. While the adversary mimics the actor by minimizing the KL-divergence between their respective action distributions, the actor, in addition to learning to solve the task, tries to differentiate itself from the adversary predictions. This novel objective stimulates the actor to follow strategies that could not have been correctly predicted from previous trajectories, making its behavior innovative in tasks where the reward is extremely rare. Our experimental analysis shows that the resulting Adversarially Guided Actor-Critic (AGAC) algorithm leads to more exhaustive exploration. Notably, AGAC outperforms current state-of-the-art methods on a set of various hard-exploration and procedurally-generated tasks.
****
On the Effectiveness of Deep Ensembles for Small Data Tasks
Lorenzo Brigato,Luca Iocchi
Deep neural networks represent the gold standard for image classification.
However, they usually need large amounts of data to reach superior performance.
In this work, we focus on image classification problems with a few labeled examples per class and improve sample efficiency in the low data regime by using an ensemble of relatively small deep networks.
For the first time, our work broadly studies the existing concept of neural ensembling in small data domains, through an extensive validation using popular datasets and architectures.
We show that deep ensembling is a simple yet effective technique that outperforms current state-of-the-art approaches for learning from small datasets.
We compare different ensemble configurations to their deeper and wider competitors given a total fixed computational budget and provide empirical evidence of their advantage.
Furthermore, we investigate the effectiveness of different losses and show that their choice should be made considering different factors.
****
AT-GAN: An Adversarial Generative Model for Non-constrained Adversarial Examples
Xiaosen Wang,Kun He,Chuanbiao Song,Liwei Wang,John E. Hopcroft
With the rapid development of adversarial machine learning, numerous adversarial attack methods have been proposed. Typical attacks are based on a search in the neighborhood of input image to generate a perturbed adversarial example. Since 2017, generative models are adopted for adversarial attacks, and most of them focus on generating adversarial perturbations from input noise or input image. Thus the output is restricted by input for these works. A recent work targets unrestricted adversarial example using generative model but their method is based on a search in the neighborhood of input noise, so actually their output is still constrained by input. In this work, we propose AT-GAN (Adversarial Transfer on Generative Adversarial Net) to train an adversarial generative model that can directly produce adversarial examples. Different from previous works, we aim to learn the distribution of adversarial examples so as to generate semantically meaningful adversaries. AT-GAN achieves this goal by first learning a generative model for real data, followed by transfer learning to obtain the desired generative model. Once trained and transferred, AT-GAN could generate adversarial examples directly and quickly for any input noise, denoted as non-constrained adversarial examples. Extensive experiments and visualizations show that AT-GAN can efficiently generate diverse adversarial examples that are realistic to human perception, and yields higher attack success rates against adversarially trained models.

****
Training independent subnetworks for robust prediction
Marton Havasi,Rodolphe Jenatton,Stanislav Fort,Jeremiah Zhe Liu,Jasper Snoek,Balaji Lakshminarayanan,Andrew Mingbo Dai,Dustin Tran
Recent approaches to efficiently ensemble neural networks have shown that strong robustness and uncertainty performance can be achieved with a negligible gain in parameters over the original network. However, these methods still require multiple forward passes for prediction, leading to a significant runtime cost. In this work, we show a surprising result:
the benefits of using multiple predictions can be achieved 'for free' under a single model's forward pass. In particular, we show that, using a multi-input multi-output (MIMO) configuration, one can utilize a single model's capacity to train multiple subnetworks that independently learn the task at hand. By ensembling the predictions made by the subnetworks, we improve model robustness without increasing compute. We observe a significant improvement in negative log-likelihood, accuracy, and calibration error on CIFAR10, CIFAR100, ImageNet, and their out-of-distribution variants compared to previous methods.
****
Complex Query Answering with Neural Link Predictors
Erik Arakelyan,Daniel Daza,Pasquale Minervini,Michael Cochez
Neural link predictors are immensely useful for identifying missing edges in large scale Knowledge Graphs. However, it is still not clear how to use these models for answering more complex queries that arise in a number of domains, such as queries using logical conjunctions ($\land$), disjunctions ($\lor$) and existential quantifiers ($\exists$), while accounting for missing edges. In this work, we propose a framework for efficiently answering complex queries on incomplete Knowledge Graphs. We translate each query into an end-to-end differentiable objective, where the truth value of each atom is computed by a pre-trained neural link predictor. We then analyse two solutions to the optimisation problem, including gradient-based and combinatorial search. In our experiments, the proposed approach produces more accurate results than state-of-the-art methods --- black-box neural models trained on millions of generated queries --- without the need of training on a large and diverse set of complex queries. Using orders of magnitude less training data, we obtain relative improvements ranging from 8% up to 40% in Hits@3 across different knowledge graphs containing factual information. Finally, we demonstrate that it is possible to explain the outcome of our model in terms of the intermediate solutions identified for each of the complex query atoms. All our source code and datasets are available online, at https://github.com/uclnlp/cqd.
****
Understanding Mental Representations Of Objects Through Verbs Applied To Them
Ka Chun Lam,Francisco Pereira,Maryam Vaziri-Pashkam,Kristin Woodard,Emalie McMahon
In order to interact with objects in our environment, we rely on an understanding of the actions that can be performed on them, and the extent to which they rely or have an effect on the properties of the object. This knowledge is called the object "affordance". We propose an approach for creating an embedding of objects in an affordance space, in which each dimension corresponds to an aspect of meaning shared by many actions, using text corpora. This embedding makes it possible to predict which verbs will be applicable to a given object, as captured in human judgments of affordance, better than a variety of alternative approaches. Furthermore, we show that the dimensions learned are interpretable, and that they correspond to typical patterns of interaction with objects. Finally, we show that the dimensions can be used to predict a state-of-the-art mental representation of objects, derived purely from human judgements of object similarity.
****
Grounding Language to Autonomously-Acquired Skills via Goal Generation
Ahmed Akakzia,Cédric Colas,Pierre-Yves Oudeyer,Mohamed CHETOUANI,Olivier Sigaud
We are interested in the autonomous acquisition of repertoires of skills. Language-conditioned reinforcement learning (LC-RL) approaches are great tools in this

quest, as they allow to express abstract goals as sets of constraints on the states. However, most LC-RL agents are not autonomous and cannot learn without external instructions and feedback. Besides, their direct language condition cannot account for the goal-directed behavior of pre-verbal infants and strongly limits the expression of behavioral diversity for a given language input. To resolve these issues, we propose a new conceptual approach to language-conditioned RL: the Language-Goal-Behavior architecture (LGB). LGB decouples skill learning and language grounding via an intermediate semantic representation of the world. To showcase the properties of LGB, we present a specific implementation called DECSTR. DECSTR is an intrinsically motivated learning agent endowed with an innate semantic representation describing spatial relations between physical objects. In a first stage G -> B, it freely explores its environment and targets self-generated semantic configurations. In a second stage (L -> G), it trains a language-conditioned goal generator to generate semantic goals that match the constraints expressed in language-based inputs. We showcase the additional properties of LGB w.r.t. both an end-to-end LC-RL approach and a similar approach leveraging non-semantic, continuous intermediate representations. Intermediate semantic representations help satisfy language commands in a diversity of ways, enable strategy switching after a failure and facilitate language grounding.
****

Hopfield Networks is All You Need

Hubert Ramsauer,Bernhard Schäfl,Johannes Lehner,Philipp Seidl,Michael Widrich,Lukas Gruber,Markus Holzleitner,Thomas Adler,David Kreil,Michael K Kopp,Günter Klambauer,Johannes Brandstetter,Sepp Hochreiter

We introduce a modern Hopfield network with continuous states and a corresponding update rule. The new Hopfield network can store exponentially (with the dimension of the associative space) many patterns, retrieves the pattern with one update, and has exponentially small retrieval errors. It has three types of energy minima (fixed points of the update): (1) global fixed point averaging over all patterns, (2) metastable states averaging over a subset of patterns, and (3) fixed points which store a single pattern. The new update rule is equivalent to the attention mechanism used in transformers. This equivalence enables a characterization of the heads of transformer models. These heads perform in the first layers preferably global averaging and in higher layers partial averaging via metastable states. The new modern Hopfield network can be integrated into deep learning architectures as layers to allow the storage of and access to raw input data, intermediate results, or learned prototypes.

These Hopfield layers enable new ways of deep learning, beyond fully-connected, convolutional, or recurrent networks, and provide pooling, memory, association, and attention mechanisms. We demonstrate the broad applicability of the Hopfield layers

across various domains. Hopfield layers improved state-of-the-art on three out of four considered multiple instance learning problems as well as on immune repertoire classification with several hundreds of thousands of instances. On the UCI benchmark collections of small classification tasks, where deep learning methods typically struggle, Hopfield layers yielded a new state-of-the-art when compared to different machine learning methods. Finally, Hopfield layers achieved state-of-the-art on two drug design datasets. The implementation is available at: \url{https://github.com/ml-jku/hopfield-layers}
****

Adapt-and-Adjust: Overcoming the Long-tail Problem of Multilingual Speech Recognition

Genta Indra Winata,Guangsen Wang,Caiming Xiong,Steven Hoi

One crucial challenge of real-world multilingual speech recognition is the long-tailed distribution problem, where some resource-rich languages like English have abundant training data, but a long tail of low-resource languages have varying amounts of limited training data. To overcome the long-tail problem, in this paper, we propose Adapt-and-Adjust (A2), a transformer-based multi-task learning framework for end-to-end multilingual speech recognition. The A2 framework overcomes the long-tail problem via three techniques: (1) exploiting a pretrained mult

ilingual language model (mBERT) to improve the performance of low-resource langu ages; (2) proposing dual adapters consisting of both language-specific and langu age-agnostic adaptation with minimal additional parameters; and (3) overcoming t he class imbalance, either by imposing class priors in the loss during training or adjusting the logits of the softmax output during inference. Extensive experi ments on the CommonVoice corpus show that A2 significantly outperforms conventio nal approaches.
****
Learning to Share in Multi-Agent Reinforcement Learning
Yuxuan Yi,Ge Li,Yaowei Wang,Zongqing Lu
In this paper, we study the problem of networked multi-agent reinforcement learn ing (MARL), where a number of agents are deployed as a partially connected netwo rk. Networked MARL requires all agents make decision in a decentralized manner t o optimize a global objective with restricted communication between neighbors ov er the network. We propose a hierarchically decentralized MARL method, \textit{L ToS}, which enables agents to learn to dynamically share reward with neighbors s o as to encourage agents to cooperate on the global objective. For each agent, t he high-level policy learns how to share reward with neighbors to decompose the global objective, while the low-level policy learns to optimize local objective induced by the high-level policies in the neighborhood. The two policies form a bi-level optimization and learn alternately. We empirically demonstrate that LTo S outperforms existing methods in both social dilemma and two networked MARL sce narios.
****
Neuron Activation Analysis for Multi-Joint Robot Reinforcement Learning
Benedikt Feldotto,Heiko Lengenfelder,Alois Knoll
Recent experiments indicate that pre-training of end-to-end Reinforcement Learni ng neural networks on general tasks can speed up the training process for specif ic robotic applications. However, it remains open if these networks form general feature extractors and a hierarchical organization that are reused as apparent e.g. in Convolutional Neural Networks. In this paper we analyze the intrinsic ne uron activation in networks trained for target reaching of robot manipulators wi th increasing joint number in a vertical plane. We analyze the individual neuron activity distribution in the network, introduce a pruning algorithm to reduce n etwork size keeping the performance, and with these dense network representation s we spot correlations of neuron activity patterns among networks trained for ro bot manipulators with different joint number. We show that the input and output network layers have more distinct neuron activation in contrast to inner layers. Our pruning algorithm reduces the network size significantly, increases the dis tance of neuron activation while keeping a high performance in training and eval uation. Our results demonstrate that neuron activity can be mapped among network s trained for robots with different complexity. Hereby, robots with small joint difference show higher layer-wise projection accuracy whereas more different rob ots mostly show projections to the first layer.
****
AlgebraNets
Jordan Hoffmann,Simon Schmitt,Simon Osindero,Karen Simonyan,Erich Elsen
Neural networks have historically been built layerwise from the set of functions in ${f: \mathbb{R}^n \to \mathbb{R}^m }$, i.e. with activations and weights/par ameters represented by real numbers, $\mathbb{R}$. Our work considers a richer s et of objects for activations and weights, and undertakes a comprehensive study of alternative algebras as number representations by studying their performance on two challenging problems: large-scale image classification using the ImageNet dataset and language modeling using the enwiki8 and WikiText-103 datasets. We d enote this broader class of models as AlgebraNets. Our findings indicate that th e conclusions of prior work, which explored neural networks constructed from $\m athbb{C}$ (complex numbers) and $\mathbb{H}$ (quaternions) on smaller datasets, do not always transfer to these challenging settings. However, our results demo nstrate that there are alternative algebras which deliver better parameter and c omputational efficiency compared with $\mathbb{R}$. We consider $\mathbb{C}$, $\

mathbb{H}$, $M_{2}(\mathbb{R})$ (the set of 2×2 real-valued matrices), $M_{2}(\mathbb{C})$, $M_{3}(\mathbb{R})$, $M_{4}(\mathbb{R})$, dual numbers and the $\mathbb{R}^3$ cross product. Additionally, we note that multiplication in these algebras has higher compute density than real multiplication, a useful property in situations with inherently limited parameter reuse such as auto-regressive inference and sparse neural networks. We therefore investigate how to induce sparsity within AlgebraNets. We hope that our strong results on large-scale, practical benchmarks will spur further exploration of these unconventional architectures which challenge the default choice of using real numbers for neural network weights and activations.

****

Random Network Distillation as a Diversity Metric for Both Image and Text Generation

Liam H Fowl,Micah Goldblum,Arjun Gupta,Amr Sharaf,Tom Goldstein

Generative models are increasingly able to produce remarkably high quality images and text. The community has developed numerous evaluation metrics for comparing generative models. However, these metrics do not effectively quantify data diversity. We develop a new diversity metric that can readily be applied to data, both synthetic and natural, of any type. Our method employs random network distillation, a technique introduced in reinforcement learning. We validate and deploy this metric on both images and text. We further explore diversity in few-shot image generation, a setting which was previously difficult to evaluate.

****

Differentiable Trust Region Layers for Deep Reinforcement Learning

Fabian Otto,Philipp Becker,Vien Anh Ngo,Hanna Carolin Maria Ziesche,Gerhard Neumann

Trust region methods are a popular tool in reinforcement learning as they yield robust policy updates in continuous and discrete action spaces. However, enforcing such trust regions in deep reinforcement learning is difficult. Hence, many approaches, such as Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO), are based on approximations. Due to those approximations, they violate the constraints or fail to find the optimal solution within the trust region. Moreover, they are difficult to implement, often lack sufficient exploration, and have been shown to depend on seemingly unrelated implementation choices. In this work, we propose differentiable neural network layers to enforce trust regions for deep Gaussian policies via closed-form projections. Unlike existing methods, those layers formalize trust regions for each state individually and can complement existing reinforcement learning algorithms. We derive trust region projections based on the Kullback-Leibler divergence, the Wasserstein L2 distance, and the Frobenius norm for Gaussian distributions. We empirically demonstrate that those projection layers achieve similar or better results than existing methods while being almost agnostic to specific implementation choices. The code is available at https://git.io/Jthb0.

****

Self-supervised Visual Reinforcement Learning with Object-centric Representations

Andrii Zadaianchuk,Maximilian Seitzer,Georg Martius

Autonomous agents need large repertoires of skills to act reasonably on new tasks that they have not seen before. However, acquiring these skills using only a stream of high-dimensional, unstructured, and unlabeled observations is a tricky challenge for any autonomous agent. Previous methods have used variational autoencoders to encode a scene into a low-dimensional vector that can be used as a goal for an agent to discover new skills. Nevertheless, in compositional/multi-object environments it is difficult to disentangle all the factors of variation into such a fixed-length representation of the whole scene. We propose to use object-centric representations as a modular and structured observation space, which is learned with a compositional generative world model.

We show that the structure in the representations in combination with goal-conditioned attention policies helps the autonomous agent to discover and learn usefu

l skills. These skills can be further combined to address compositional tasks li
ke the manipulation of several different objects.
****

Balancing training time vs. performance with Bayesian Early Pruning
Mohit Rajpal,Yehong Zhang,Bryan Kian Hsiang Low
Pruning is an approach to alleviate overparameterization of deep neural networks
 (DNN) by zeroing out or pruning DNN elements with little to no efficacy at a gi
ven task. In contrast to related works that do pruning before or after training,
 this paper presents a novel method to perform early pruning of DNN elements (e.
g., neurons or convolutional filters) during the training process while preservi
ng performance upon convergence. To achieve this, we model the future efficacy o
f DNN elements in a Bayesian manner conditioned upon efficacy data collected dur
ing the training and prune DNN elements which are predicted to have low efficacy
 after training completion. Empirical evaluations show that the proposed Bayesia
n early pruning improves the computational efficiency of DNN training with small
 sacrifices in performance. Using our approach we are able to achieve a 48.6%
 faster training time for ResNet-50 on ImageNet to achieve a validation accura
cy of 72.5%.
****

Temporally-Extended ε-Greedy Exploration
Will Dabney,Georg Ostrovski,Andre Barreto
Recent work on exploration in reinforcement learning (RL) has led to a series of
 increasingly complex solutions to the problem. This increase in complexity ofte
n comes at the expense of generality. Recent empirical studies suggest that, whe
n applied to a broader set of domains, some sophisticated exploration methods ar
e outperformed by simpler counterparts, such as ε-greedy. In this paper we propo
se an exploration algorithm that retains the simplicity of ε-greedy while reduci
ng dithering. We build on a simple hypothesis: the main limitation of ε-greedy e
xploration is its lack of temporal persistence, which limits its ability to esca
pe local optima. We propose a temporally extended form of ε-greedy that simply r
epeats the sampled action for a random duration. It turns out that, for many dur
ation distributions, this suffices to improve exploration on a large set of doma
ins. Interestingly, a class of distributions inspired by ecological models of an
imal foraging behaviour yields particularly strong performance.
****

Attention Based Joint Learning for Supervised Electrocardiogram Arrhythmia Diffe
rentiation with Unsupervised Abnormal Beat Segmentation
Xinrong Hu,long wen,shushui wang,Dongpo Liang,Jian Zhuang,Yiyu Shi
Deep learning has shown great promise in arrhythmia classification in electrocar
- diogram (ECG). Existing works, when classifying an ECG segment with multiple b
eats, do not identify the locations of the anomalies, which reduces clinical int
er- pretability. On the other hand, segmenting abnormal beats by deep learning u
su- ally requires annotation for a large number of regular and irregular beats,
which can be laborious, sometimes even challenging, with strong inter-observer v
ariabil- ity between experts. In this work, we propose a method capable of not o
nly dif- ferentiating arrhythmia but also segmenting the associated abnormal bea
ts in the ECG segment. The only annotation used in the training is the type of a
bnormal beats and no segmentation labels are needed. Imitating human’s perceptio
n of an ECG signal, the framework consists of a segmenter and classifier. The se
gmenter outputs an attention map, which aims to highlight the abnormal sections
in the ECG by element-wise modulation. Afterwards, the signals are sent to a cla
ssifier for arrhythmia differentiation. Though the training data is only labeled
 to super- vise the classifier, the segmenter and the classifier are trained in
an end-to-end manner so that optimizing classification performance also adjusts
how the abnor- mal beats are segmented. Validation of our method is conducted on
 two dataset. We observe that involving the unsupervised segmentation in fact bo
osts the clas- sification performance. Meanwhile, a grade study performed by exp
erts suggests that the segmenter also achieves satisfactory quality in identifyi
ng abnormal beats, which significantly enhances the interpretability of the clas
sification results.

****
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 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 the current leading model of the adult ventral stream, 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. 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. When these three strategies are applied in combination, we find that these new models achieve ~80% of a fully trained model's match to the brain, while using two orders of magnitude fewer supervised synaptic updates. 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 Associative Inference Using Fast Weight Memory
Imanol Schlag,Tsendsuren Munkhdalai,Jürgen Schmidhuber
Humans can quickly associate stimuli to solve problems in novel contexts. Our novel neural network model learns state representations of facts that can be composed to perform such associative inference. To this end, we augment the LSTM model with an associative memory, dubbed \textit{Fast Weight Memory} (FWM). Through differentiable operations at every step of a given input sequence, the LSTM \textit{updates and maintains} compositional associations stored in the rapidly changing FWM weights. Our model is trained end-to-end by gradient descent and yields excellent performance on compositional language reasoning problems, meta-reinforcement-learning for POMDPs, and small-scale word-level language modelling.
****
Deep Learning Is Composite Kernel Learning
CHANDRA SHEKAR LAKSHMINARAYANAN,Amit Vikram Singh
Recent works have connected deep learning and kernel methods. In this paper, we show that architectural choices such as convolutional layers with pooling, skip connections, make deep learning a composite kernel learning method, where the kernel is a (architecture dependent) composition of base kernels: even before training, standard deep networks have in-built structural properties that ensure their success. In particular, we build on the recently developed `neural path' framework that characterises the role of gates/masks in fully connected deep networks with ReLU activations.
****
Multiscale Score Matching for Out-of-Distribution Detection
Ahsan Mahmood,Junier Oliva,Martin Andreas Styner
We present a new methodology for detecting out-of-distribution (OOD) images by utilizing norms of the score estimates at multiple noise scales. A score is defined to be the gradient of the log density with respect to the input data. Our methodology is completely unsupervised and follows a straight forward training scheme. First, we train a deep network to estimate scores for L levels of noise. O

nce trained, we calculate the noisy score estimates for N in-distribution samp
les and take the L2-norms across the input dimensions (resulting in an NxL m
atrix). Then we train an auxiliary model (such as a Gaussian Mixture Model) to l
earn the in-distribution spatial regions in this L-dimensional space. This aux
iliary model can now be used to identify points that reside outside the learned
space. Despite its simplicity, our experiments show that this methodology signif
icantly outperforms the state-of-the-art in detecting out-of-distribution images
. For example, our method can effectively separate CIFAR-10 (inlier) and SVHN (O
OD) images, a setting which has been previously shown to be difficult for deep l
ikelihood models.
****

Data Instance Prior for Transfer Learning in GANs
Puneet Mangla,Nupur Kumari,Mayank Singh,Vineeth N. Balasubramanian,Balaji Krishn
amurthy
Recent advances in generative adversarial networks (GANs) have shown remarkable
progress in generating high-quality images. However, this gain in performance de
pends on the availability of a large amount of training data. In limited data re
gimes, training typically diverges, and therefore the generated samples are of l
ow quality and lack diversity. Previous works have addressed training in low dat
a setting by leveraging transfer learning and data augmentation techniques. We p
ropose a novel transfer learning method for GANs in the limited data domain by l
everaging informative data prior derived from self-supervised/supervised pre-tra
ined networks trained on a diverse source domain. We perform experiments on seve
ral standard vision datasets using various GAN architectures (BigGAN, SNGAN, Sty
leGAN2) to demonstrate that the proposed method effectively transfers knowledge
to domains with few target images, outperforming existing state-of-the-art techn
iques in terms of image quality and diversity. We also show the utility of data
instance prior in large-scale unconditional image generation and image editing t
asks.
****

WordsWorth Scores for Attacking CNNs and LSTMs for Text Classification
Nimrah Shakeel
Black box attacks on traditional deep learning models trained for text classific
a- tion target important words in a piece of text, in order to change model pred
iction. Current approaches towards highlighting important features are time cons
uming and require large number of model queries. We present a simple yet novel m
ethod to calculate word importance scores, based on model predictions on single
words. These scores, which we call WordsWorth scores, need to be calculated only
 once for the training vocabulary. They can be used to speed up any attack metho
d that requires word importance, with negligible loss of attack performance. We
run ex- periments on a number of datasets trained on word-level CNNs and LSTMs,
for sentiment analysis and topic classification and compare to state-of-the-art
base- lines. Our results show the effectiveness of our method in attacking these
 models with success rates that are close to the original baselines. We argue th
at global importance scores act as a very good proxy for word importance in a lo
cal context because words are a highly informative form of data. This aligns wit
h the manner in which humans interpret language, with individual words having we
ll- defined meaning and powerful connotations. We further show that these scores
 can be used as a debugging tool to interpret a trained model by highlighting re
le- vant words for each class. Additionally, we demonstrate the effect of overtr
aining on word importance, compare the robustness of CNNs and LSTMs, and explain
 the transferability of adversarial examples across a CNN and an LSTM using thes
e scores. We highlight the fact that neural networks make highly informative pre
- dictions on single words.
****

Inductive Collaborative Filtering via Relation Graph Learning
Qitian Wu,Hengrui Zhang,Xiaofeng Gao,Hongyuan Zha
Collaborative filtering has shown great power in predicting potential user-item
ratings by factorizing an observed user-item rating matrix into products of two
sets of latent factors. However, the user-specific latent factors can only be le

arned in transductive setting and a model trained on existing users cannot adapt to new users without retraining the model. In this paper, we propose an inductive collaborative filtering framework that learns a hidden relational graph among users from the rating matrix. We first consider a base matrix factorization model trained on one group of users' ratings and devise a relation inference model that estimates their underlying relations (as dense weighted graphs) to other users with respect to historical rating patterns. The relational graphs enable attentive message passing from users to users in the latent space and are updated in end-to-end manner. The key advantage of our model is the capability for inductively computing user-specific representations using no feature, with good scalability and superior expressiveness compared to other feature-driven inductive models. Extensive experiments demonstrate that our model achieves state-of-the-art performance for inductive learning on several matrix completion benchmarks, provides very close performance to transductive models when given many training ratings and exceeds them significantly on cold-start users.
****

Learning to Noise: Application-Agnostic Data Sharing with Local Differential Privacy

Alex Mansbridge,Gregory Barbour,Davide Piras,Christopher Frye,Ilya Feige,David Barber

In recent years, the collection and sharing of individuals’ private data has become commonplace in many industries. Local differential privacy (LDP) is a rigorous approach which uses a randomized algorithm to preserve privacy even from the database administrator, unlike the more standard central differential privacy. For LDP, when applying noise directly to high-dimensional data, the level of noise required all but entirely destroys data utility. In this paper we introduce a novel, application-agnostic privatization mechanism that leverages representation learning to overcome the prohibitive noise requirements of direct methods, while maintaining the strict guarantees of LDP. We further demonstrate that data privatized with this mechanism can be used to train machine learning algorithms. Applications of this model include private data collection, private novel-class classification, and the augmentation of clean datasets with additional privatized features. We achieve significant gains in performance on downstream classification tasks relative to benchmarks that noise the data directly, which are state-of-the-art in the context of application-agnostic LDP mechanisms for high-dimensional data sharing tasks.
****

Learning to Sample with Local and Global Contexts in Experience Replay Buffer

Youngmin Oh,Kimin Lee,Jinwoo Shin,Eunho Yang,Sung Ju Hwang

Experience replay, which enables the agents to remember and reuse experience from the past, has played a significant role in the success of off-policy reinforcement learning (RL). To utilize the experience replay efficiently, the existing sampling methods allow selecting out more meaningful experiences by imposing priorities on them based on certain metrics (e.g. TD-error). However, they may result in sampling highly biased, redundant transitions since they compute the sampling rate for each transition independently, without consideration of its importance in relation to other transitions. In this paper, we aim to address the issue by proposing a new learning-based sampling method that can compute the relative importance of transition. To this end, we design a novel permutation-equivariant neural architecture that takes contexts from not only features of each transition (local) but also those of others (global) as inputs. We validate our framework, which we refer to as Neural Experience Replay Sampler (NERS), on multiple benchmark tasks for both continuous and discrete control tasks and show that it can significantly improve the performance of various off-policy RL methods. Further analysis confirms that the improvements of the sample efficiency indeed are due to sampling diverse and meaningful transitions by NERS that considers both local and global contexts.
****

Active Deep Probabilistic Subsampling

Hans van Gorp,Iris A.M. Huijben,Bastiaan S. Veeling,Nicola Pezzotti,Ruud Van Slo

un
Subsampling a signal of interest can reduce costly data transfer, battery drain, radiation exposure and acquisition time in a wide range of problems. The recently proposed Deep Probabilistic Subsampling (DPS) method effectively integrates subsampling in an end-to-end deep learning model, but learns a static pattern for all datapoints. We generalize DPS to a sequential method that actively picks the next sample based on the information acquired so far; dubbed Active-DPS (A-DPS). We validate that A-DPS improves over DPS for MNIST classification at high subsampling rates. We observe that A-DPS learns to actively adapt based on the previously sampled elements, yielding different sampling sequences across the dataset. Moreover, we demonstrate strong performance in active acquisition Magnetic Resonance Image (MRI) reconstruction, outperforming DPS and other deep learning methods.
****
Improving Sequence Generative Adversarial Networks with Feature Statistics Alignment
Yekun Chai,Qiyue Yin,Junge Zhang
Generative Adversarial Networks (GAN) are facing great challenges in synthesizing sequences of discrete elements, such as mode dropping and unstable training. The binary classifier in the discriminator may limit the capacity of learning signals and thus hinder the advance of adversarial training. To address such issues, apart from the binary classification feedback, we harness a Feature Statistics Alignment (FSA) paradigm to deliver fine-grained signals in the latent high-dimensional representation space. Specifically, FSA forces the mean statistics of the fake data distribution to approach that of real data as close as possible in a finite-dimensional feature space. Experiments on synthetic and real benchmark datasets show the superior performance in quantitative evaluation and demonstrate the effectiveness of our approach to discrete sequence generation. To the best of our knowledge, the proposed architecture is the first that employs feature alignment regularization in the Gumbel-Softmax based GAN framework for sequence generation.
****
Using Deep Reinforcement Learning to Train and Evaluate Instructional Sequencing Policies for an Intelligent Tutoring System
Jithendaraa Subramanian,David Mostow
We present STEP, a novel Deep Reinforcement Learning solution to the problem of learning instructional sequencing. STEP has three components: 1. Simulate the student by fitting a knowledge tracing model to data logged by an intelligent tutoring system. 2. Train instructional sequencing policies by using Proximal Policy Optimization. 3. Evaluate the learned instructional policies by estimating their local and global impact on learning gains. STEP leverages the student model by representing the student’s knowledge state as a probability vector of knowing each skill and using the student’s estimated learning gains as its reward function to evaluate candidate policies. A learned policy represents a mapping from each state to an action that maximizes the reward, i.e. the upward distance to the next state in the multi-dimensional space. We use STEP to discover and evaluate potential improvements to a literacy and numeracy tutor used by hundreds of children in Tanzania.
****
Sparsifying Networks via Subdifferential Inclusion
Sagar Verma,Jean-Christophe Pesquet
Sparsifying deep neural networks is of paramount interest in many areas, especially when those networks have to be implemented on low-memory devices. In this article, we propose a new formulation of the problem of generating sparse weights for a neural network. By leveraging the properties of standard nonlinear activation functions, we show that the problem is equivalent to an approximate subdifferential inclusion problem. The accuracy of the approximation controls the sparsity. We show that the proposed approach is valid for a broad class of activation functions (ReLU, sigmoid, softmax). We propose an iterative optimization algorithm to induce sparsity whose convergence is guaranteed. Because of the algorithm

flexibility, the sparsity can be ensured from partial training data in a minibat
ch manner. To demonstrate the effectiveness of our method, we perform experiment
s on various networks in different applicative contexts: image classification, s
peech recognition, natural language processing, and time-series forecasting.
****

Two steps at a time --- taking GAN training in stride with Tseng's method
Axel Böhm,Michael Sedlmayer,Ernö Robert Csetnek,Radu Ioan Bot
Motivated by the training of Generative Adversarial Networks (GANs), we study me
thods for solving minimax problems with additional nonsmooth regularizers.
We do so by employing \emph{monotone operator} theory, in particular the \emph{F
orward-Backward-Forward (FBF)} method, which avoids the known issue of limit cyc
ling by correcting each update by a second gradient evaluation.
Furthermore, we propose a seemingly new scheme which recycles old gradients to m
itigate the additional computational cost.
In doing so we rediscover a known method, related to \emph{Optimistic Gradient D
escent Ascent (OGDA)}.
For both schemes we prove novel convergence rates for convex-concave minimax pro
blems via a unifying approach. The derived error bounds are in terms of the gap
function for the ergodic iterates.
For the deterministic and the stochastic problem we show a convergence rate of $
\mathcal{O}(\nicefrac{1}{k})$ and $\mathcal{O}(\nicefrac{1}{\sqrt{k}})$, respect
ively.
We complement our theoretical results with empirical improvements in the trainin
g of Wasserstein GANs on the CIFAR10 dataset.
****

Invariant Causal Representation Learning
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 causal relationship with the outcome.
 This is done by leveraging a diverse set of training environments to reduce the
 effect of spurious features, on top of which an invariant classifier is then bu
ilt. However, these methods have generalization guarantees only when both data r
epresentation and classifiers come from a linear model class. As an alternative,
 we propose Invariant Causal Representation Learning (ICRL), a learning paradigm
 that enables out-of-distribution generalization in the nonlinear setting (i.e.,
 nonlinear representations and nonlinear classifiers). It builds upon a practica
l and general assumption: data representations factorize when conditioning on th
e outcome and the environment. Based on this, we show identifiability up to a pe
rmutation and pointwise transformation. We also prove that all direct causes of
the outcome can be fully discovered, which further enables us to obtain generali
zation guarantees in the nonlinear setting. Extensive experiments on both synthe
tic and real-world datasets show that our approach significantly outperforms a v
ariety of baseline methods.
****

A priori guarantees of finite-time convergence for Deep Neural Networks
Anushree Rankawat,Mansi Rankawat,Harshal B. Oza
In this paper, we perform Lyapunov based analysis of the loss function to derive
 an a priori upper bound on the settling time of deep neural networks. While pre
vious studies have attempted to understand deep learning using control theory fr
amework, there is limited work on a priori finite time convergence analysis. Dra
wing from the advances in analysis of finite-time control of non-linear systems,
 we provide a priori guarantees of finite-time convergence in a deterministic co
ntrol theoretic setting. We formulate the supervised learning framework as a con
trol problem where weights of the network are control inputs and learning transl
ates into a tracking problem. An analytical formula for finite-time upper bound
on settling time is provided a priori under the assumptions of boundedness of in
put. Finally, we prove that our loss function is robust against input perturbati
ons.

****
Distributed Associative Memory Network with Association Reinforcing Loss
Taewon Park,Inchul Choi,Minho Lee
Despite recent progress in memory augmented neural network research, associative memory networks with a single external memory still show limited performance on complex relational reasoning tasks. The main reason for this problem comes from the lossy representation of a content-based addressing memory and its insufficient associating performance for long temporal sequence data. To address these problems, here we introduce a novel Distributed Associative Memory architecture (DAM) with Association Reinforcing Loss (ARL) function which enhances the relation reasoning performance of memory augmented neural network. In this framework, instead of relying on a single large external memory, we form a set of multiple smaller associative memory blocks and update these sub-memory blocks simultaneously and independently with the content-based addressing mechanism. Based on DAM architecture, we can effectively retrieve complex relational information by integrating diverse representations distributed across multiple sub-memory blocks with an attention mechanism. Moreover, to further enhance the relation modeling performance of memory network, we propose ARL which assists a task's target objective while learning relational information exist in data. ARL enables the memory augmented neural network to reinforce an association between input data and task objective by reproducing stochastically sampled input data from stored memory contents. With this content reproducing task, it enriches the representations with relational information. In experiments, we apply our two main approaches to Differential Neural Computer (DNC), which is one of the representative content-based addressing memory model and achieves state-of-the-art performance on both memorization and relational reasoning tasks.
****
Unsupervised Discovery of Interpretable Latent Manipulations in Language VAEs
Max Ryabinin,Artem Babenko,Elena Voita
Language generation models are attracting more and more attention due to their constantly increasing quality and remarkable generation results. State-of-the-art NLG models like BART/T5/GPT-3 do not have latent spaces, therefore there is no natural way to perform controlled generation. In contrast, less popular models with explicit latent spaces have the innate ability to manipulate text attributes by moving along latent directions. For images, properties of latent spaces are well-studied: there exist interpretable directions (e.g. zooming, aging, background removal) and they can even be found without supervision. This success is expected: latent space image models, especially GANs, achieve state-of-the-art generation results and hence have been the focus of the research community. For language, this is not the case: text GANs are hard to train because of non-differentiable discrete data generation, and language VAEs suffer from posterior collapse and fill the latent space poorly. This makes finding interpetable text controls challenging. In this work, we make the first step towards unsupervised discovery of interpretable directions in language latent spaces. For this, we turn to methods shown to work in the image domain. Surprisingly, we find that running PCA on VAE representations of training data consistently outperforms shifts along the coordinate and random directions. This approach is simple, data-adaptive, does not require training and discovers meaningful directions, e.g. sentence length, subject age, and verb tense. Our work lays foundations for two important areas: first, it allows to compare models in terms of latent space interpretability, and second, it provides a baseline for unsupervised latent controls discovery.
****
Variational inference for diffusion modulated Cox processes
Prateek Jaiswal,Harsha Honnappa,Vinayak Rao
This paper proposes a stochastic variational inference (SVI) method for computing an approximate posterior path measure of a Cox process. These processes are widely used in natural and physical sciences, engineering and operations research, and represent a non-trivial model of a wide array of phenomena. In our work, we model the stochastic intensity as the solution of a diffusion stochastic differential equation (SDE), and our objective is to infer the posterior, or smoothing

, measure over the paths given Poisson process realizations. We first derive a s
ystem of stochastic partial differential equations (SPDE) for the pathwise smoot
hing posterior density function, a non-trivial result, since the standard soluti
on of SPDEs typically involves an It\^o stochastic integral, which is not define
d pathwise. Next, we propose an SVI approach to approximating the solution of th
e system. We parametrize the class of approximate smoothing posteriors using a n
eural network, derive a lower bound on the evidence of the observed point proces
s sample-path, and optimize the lower bound using stochastic gradient descent (S
GD). We demonstrate the efficacy of our method on both synthetic and real-world
problems, and demonstrate the advantage of the neural network solution over stan
dard numerical solvers.
****

Discovering a set of policies for the worst case reward
Tom Zahavy,Andre Barreto,Daniel J Mankowitz,Shaobo Hou,Brendan O'Donoghue,Iurii
Kemaev,Satinder Singh
We study the problem of how to construct a set of policies that can be composed
together to solve a collection of reinforcement learning tasks. Each task is a d
ifferent reward function defined as a linear combination of known features. We
consider a specific class of policy compositions which we call set improving pol
icies (SIPs): given a set of policies and a set of tasks, a SIP is any compositi
on of the former whose performance is at least as good as that of its constituen
ts across all the tasks. We focus on the most conservative instantiation of SIPs
, set-max policies (SMPs), so our analysis extends to any SIP. This includes kno
wn policy-composition operators like generalized policy improvement. Our main co
ntribution is an algorithm that builds a set of policies in order to maximize th
e worst-case performance of the resulting SMP on the set of tasks. The algorithm
 works by successively adding new policies to the set. We show that the worst-ca
se performance of the resulting SMP strictly improves at each iteration, and the
 algorithm only stops when there does not exist a policy that leads to improved
performance. We empirically evaluate our algorithm on a grid world and also on a
 set of domains from the DeepMind control suite. We confirm our theoretical resu
lts regarding the monotonically improving performance of our algorithm. Interest
ingly, we also show empirically that the sets of policies computed by the algori
thm are diverse, leading to different trajectories in the grid world and very di
stinct locomotion skills in the control suite.
****

Language Controls More Than Top-Down Attention: Modulating Bottom-Up Visual Proc
essing with Referring Expressions
Ozan Arkan Can,Ilker Kesen,Deniz Yuret
How to best integrate linguistic and perceptual processing in multimodal tasks i
s an important open problem. In this work we argue that the common technique of
using language to direct visual attention over high-level visual features may no
t be optimal. Using language throughout the bottom-up visual pathway, going from
 pixels to high-level features, may be necessary. Our experiments on several Eng
lish referring expression datasets show significant improvements when language i
s used to control the filters for bottom-up visual processing in addition to top
-down attention.
****

Automatic Music Production Using Generative Adversarial Networks
Giorgio Barnabò,Giovanni Trappolini,Lorenzo Lastilla,Cesare Campagnano,Angela Fa
n,Fabio Petroni,Fabrizio Silvestri
When talking about computer-based music generation, two are the main threads of
research: the construction of $\textit{autonomous music-making systems}$, and th
e design of $\textit{computer-based environments to assist musicians}$. However,
 even though creating accompaniments for melodies is an essential part of every
producer's and songwriter's work, little effort has been done in the field of au
tomatic music arrangement in the audio domain. In this contribution, we propose
a novel framework for $\textit{automatic music accompaniment}$ $\textit{in the M
el-frequency domain}$. Using several songs converted into Mel-spectrograms, a tw
o-dimensional time-frequency representation of audio signals, we were able to au

tomatically generate original arrangements for both bass and voice lines. Treating music pieces as images (Mel-spectrograms) allowed us to reformulate our problem as an $\textit{unpaired image-to-image translation}$ problem, and to tackle it with CycleGAN, a well-established framework. Moreover, the choice to deploy raw audio and Mel-spectrograms enabled us to more effectively model long-range dependencies, to better represent how humans perceive music, and to potentially draw sounds for new arrangements from the vast collection of music recordings accumulated in the last century. Our approach was tested on two different downstream tasks: given a bass line creating credible and on-time drums, and given an acapella song arranging it to a full song. In absence of an objective way of evaluating the output of music generative systems, we also defined a possible metric for the proposed task, partially based on human (and expert) judgment.
****
Parameter-Based Value Functions
Francesco Faccio,Louis Kirsch,Jürgen Schmidhuber
Traditional off-policy actor-critic Reinforcement Learning (RL) algorithms learn value functions of a single target policy. However, when value functions are updated to track the learned policy, they forget potentially useful information about old policies. We introduce a class of value functions called Parameter-Based Value Functions (PBVFs) whose inputs include the policy parameters. They can generalize across different policies. PBVFs can evaluate the performance of any policy given a state, a state-action pair, or a distribution over the RL agent's initial states. First we show how PBVFs yield novel off-policy policy gradient theorems. Then we derive off-policy actor-critic algorithms based on PBVFs trained by Monte Carlo or Temporal Difference methods. We show how learned PBVFs can zero-shot learn new policies that outperform any policy seen during training. Finally our algorithms are evaluated on a selection of discrete and continuous control tasks using shallow policies and deep neural networks. Their performance is comparable to state-of-the-art methods.
****
Global Node Attentions via Adaptive Spectral Filters
Shouheng Li,Dongwoo Kim,Qing Wang
Graph neural networks (GNNs) have been extensively studied for prediction tasks on graphs. Most GNNs assume local homophily, i.e., strong similarities in local neighborhoods. This assumption limits the generalizability of GNNs, which has been demonstrated by recent work on disassortative graphs with weak local homophily. In this paper, we argue that GNN's feature aggregation scheme can be made flexible and adaptive to data without the assumption of local homophily. To demonstrate, we propose a GNN model with a global self-attention mechanism defined using learnable spectral filters, which can attend to any nodes, regardless of distance. We evaluated the proposed model on node classification tasks over six benchmark datasets. The proposed model has been shown to generalize well to both assortative and disassortative graphs. Further, it outperforms all state-of-the-art baselines on disassortative graphs and performs comparably with them on assortative graphs.
****
Addressing the Topological Defects of Disentanglement
Diane Bouchacourt,Mark Ibrahim,Stephane Deny
A core challenge in Machine Learning is to disentangle natural factors of variation in data (e.g. object shape vs pose). A popular approach to disentanglement consists in learning to map each of these factors to distinct subspaces of a model's latent representation. However, this approach has shown limited empirical success to date. Here, we show that this approach to disentanglement introduces topological defects (i.e. discontinuities in the encoder) for a broad family of transformations acting on images ---encompassing simple affine transformations such as rotations and translations. Moreover, motivated by classical results from group representation theory, we propose an alternative, more flexible approach to disentanglement which relies on distributed equivariant operators, potentially acting on the entire latent space. We theoretically and empirically demonstrate the effectiveness of our approach to disentangle affine transformations. Our w

ork lays a theoretical foundation for the recent success of a new generation of models using distributed operators for disentanglement (see Discussion).
****
Machine Reading Comprehension with Enhanced Linguistic Verifiers
Xianchao Wu
We propose two linguistic verifiers for span-extraction style machine reading comprehension to respectively tackle two challenges: how to evaluate the syntactic completeness of predicted answers and how to utilize the rich context of long documents. Our first verifier rewrites a question through replacing its interrogatives by the predicted answer phrases and then builds a cross-attention scorer between the rewritten question and the segment, so that the answer candidates are scored in a \emph{position-sensitive} context. Our second verifier builds a hierarchical attention network to represent segments in a passage where neighbour segments in long passages are \emph{recurrently connected} and can contribute to current segment-question pair's inference for answerablility classification and boundary determination. We then combine these two verifiers together into a pipeline and apply it to SQuAD2.0, NewsQA and TriviaQA benchmark sets. Our pipeline achieves significantly better improvements of both exact matching and F1 scores than state-of-the-art baselines.
****
New Bounds For Distributed Mean Estimation and Variance Reduction
Peter Davies,Vijaykrishna Gurunanthan,Niusha Moshrefi,Saleh Ashkboos,Dan Alistarh
We consider the problem of distributed mean estimation (DME), in which n machines are each given a local d-dimensional vector $\mathbf x_v \in \mathbb R^d$, and must cooperate to estimate the mean of their inputs $\mathbf \mu = \frac 1 n\sum_{v = 1}^n \mathbf x_v$, while minimizing total communication cost. DME is a fundamental construct in distributed machine learning, and there has been considerable work on variants of this problem, especially in the context of distributed variance reduction for stochastic gradients in parallel SGD. Previous work typically assumes an upper bound on the norm of the input vectors, and achieves an error bound in terms of this norm. However, in many real applications, the input vectors are concentrated around the correct output $\mathbf \mu$, but $\mathbf \mu$ itself has large norm. In such cases, previous output error bounds perform poorly.

In this paper, we show that output error bounds need not depend on input norm. We provide a method of quantization which allows distributed mean estimation to be performed with solution quality dependent only on the distance between inputs, not on input norm, and show an analogous result for distributed variance reduction. The technique is based on a new connection with lattice theory. We also provide lower bounds showing that the communication to error trade-off of our algorithms is asymptotically optimal. As the lattices achieving optimal bounds under ℓ_2-norm can be computationally impractical, we also present an extension which leverages easy-to-use cubic lattices, and is loose only up to a logarithmic factor in d. We show experimentally that our method yields practical improvements for common applications, relative to prior approaches.
****
Class Balancing GAN with a Classifier in the Loop
Harsh Rangwani,Konda Reddy Mopuri,Venkatesh Babu Radhakrishnan
Generative Adversarial Networks (GANs) have swiftly evolved to imitate increasingly complex image distributions. However, majority of the developments focus on performance of GANs on balanced datasets. We find that the existing GANs and their training regimes which work well on balanced datasets fail to be effective in case of imbalanced (i.e. long-tailed) datasets. In this work we introduce a novel and theoretically motivated Class Balancing regularizer for training GANs. Our regularizer makes use of the knowledge from a pre-trained classifier to ensure balanced learning of all the classes in the dataset. This is achieved via modelling the effective class frequency based on the exponential forgetting observed in neural networks and encouraging the GAN to focus on underrepresented classes. We demonstrate the utility of our contribution in two diverse scenarios: (i) Le

arning representations for long-tailed distributions, where we achieve better pe rformance than existing approaches, and (ii) Generation of Universal Adversarial Perturbations (UAPs) in the data-free scenario for the large scale datasets, wh ere we bridge the gap between data-driven and data-free approaches for crafting UAPs.
****
Learning Contextual Perturbation Budgets for Training Robust Neural Networks
Jing Xu,Zhouxing Shi,Huan Zhang,Jinfeng Yi,Cho-Jui Hsieh,Liwei Wang
Existing methods for training robust neural networks generally aim to make model s uniformly robust on all input dimensions. However, different input dimensions are not uniformly important to the prediction. In this paper, we propose a novel framework to train certifiably robust models and learn non-uniform perturbation budgets on different input dimensions, in contrast to using the popular $\ell_\ infty$ threat model. We incorporate a perturbation budget generator into the exi sting certified defense framework, and perform certified training with generated perturbation budgets. In comparison to the radius of ℓ_∞ ball in prev ious works, the robustness intensity is measured by robustness volume which is t he multiplication of perturbation budgets on all input dimensions. We evaluate o ur method on MNIST and CIFAR-10 datasets and show that we can achieve lower clea n and certified errors on relatively larger robustness volumes, compared to meth ods using uniform perturbation budgets. Further with two synthetic datasets cons tructed from MNIST and CIFAR-10, we also demonstrate that the perturbation budge t generator can produce semantically-meaningful budgets, which implies that the generator can capture contextual information and the sensitivity of different fe atures in input images.
****
Enabling Binary Neural Network Training on the Edge
Erwei Wang,James J. Davis,Daniele Moro,Piotr Zielinski,Claudionor Coelho,Satraji t Chatterjee,Peter Y. K. Cheung,George Anthony Constantinides
The ever-growing computational demands of increasingly complex machine learning models frequently necessitate the use of powerful cloud-based infrastructure for their training. Binary neural networks are known to be promising candidates for on-device inference due to their extreme compute and memory savings over higher -precision alternatives. In this paper, we demonstrate that they are also strong ly robust to gradient quantization, thereby making the training of modern models on the edge a practical reality. We introduce a low-cost binary neural network training strategy exhibiting sizable memory footprint reductions and energy savi ngs vs Courbariaux & Bengio's standard approach. Against the latter, we see coin cident memory requirement and energy consumption drops of 2--6$\times$, while re aching similar test accuracy, across a range of small-scale models trained to cl assify popular datasets. We also showcase ImageNet training of ResNetE-18, achie ving a 3.12$\times$ memory reduction over the aforementioned standard. Such savi ngs will allow for unnecessary cloud offloading to be avoided, reducing latency and increasing energy efficiency while also safeguarding user privacy.
****
Deep Coherent Exploration For Continuous Control
Yijie Zhang,Herke van Hoof
In policy search methods for reinforcement learning (RL), exploration is often p erformed by injecting noise either in action space at each step independently or in parameter space over each full trajectory. In prior work, it has been shown that with linear policies, a more balanced trade-off between these two explorati on strategies is beneficial. However, that method did not scale to policies usin g deep neural networks. In this paper, we introduce Deep Coherent Exploration, a general and scalable exploration framework for deep RL algorithms on continuous control, that generalizes step-based and trajectory-based exploration. This fra mework models the last layer parameters of the policy network as latent variable s and uses a recursive inference step within the policy update to handle these l atent variables in a scalable manner. We find that Deep Coherent Exploration imp roves the speed and stability of learning of A2C, PPO, and SAC on several contin uous control tasks.

****
Median DC for Sign Recovery: Privacy can be Achieved by Deterministic Algorithms
Jiyuan Tu,Weidong Liu,Xiaojun Mao
Privacy-preserving data analysis becomes prevailing in recent years. It is a com mon sense in privacy literature that strict differential privacy can only be obt ained by imposing additional randomness in the algorithm. In this paper, we stud y the problem of private sign recovery for sparse mean estimation and sparse lin ear regression in a distributed setup. By taking a coordinate-wise median among the reported local sign vectors, which can be referred to as a median divide-and -conquer (Med-DC) approach, we can recover the signs of the true parameter with a provable consistency guarantee. Moreover, without adding any extra randomness to the algorithm, our Med-DC method can protect data privacy with high probabili ty. Simulation studies are conducted to demonstrate the effectiveness of our pro posed method.
****
Isometric Autoencoders
Amos Gropp,Matan Atzmon,Yaron Lipman
High dimensional data is often assumed to be concentrated on or near a low-dimen sional manifold. Autoencoders (AE) is a popular technique to learn representatio ns of such data by pushing it through a neural network with a low dimension bott leneck while minimizing a reconstruction error. Using high capacity AE often lea ds to a large collection of minimizers, many of which represent a low dimensiona l manifold that fits the data well but generalizes poorly.

Two sources of bad generalization are: extrinsic, where the learned manifold pos sesses extraneous parts that are far from the data; and intrinsic, where the enc oder and decoder introduce arbitrary distortion in the low dimensional parameter ization. An approach taken to alleviate these issues is to add a regularizer tha t favors a particular solution; common regularizers promote sparsity, small deri vatives, or robustness to noise.

In this paper, we advocate an isometry (i.e., local distance preserving) regular izer. Specifically, our regularizer encourages: (i) the decoder to be an isomet ry; and (ii) the encoder to be the decoder’s pseudo-inverse, that is, the encode r extends the inverse of the decoder to the ambient space by orthogonal projecti on. In a nutshell, (i) and (ii) fix both intrinsic and extrinsic degrees of fre edom and provide a non-linear generalization to principal component analysis (PC A). Experimenting with the isometry regularizer on dimensionality reduction task s produces useful low-dimensional data representations.
****
Pseudo Label-Guided Multi Task Learning for Scene Understanding
Sunkyung Kim,Hyesong Choi,Dongbo Min
Multi-task learning (MTL) for scene understanding has been actively studied by e xploiting correlation of multiple tasks. This work focuses on improving the perf ormance of the MTL network that infers depth and semantic segmentation maps from a single image. Specifically, we propose a novel MTL architecture, called Pseud o-MTL, that introduces pseudo labels for joint learning of monocular depth estim ation and semantic segmentation tasks. The pseudo ground truth depth maps, gener ated from pretrained stereo matching methods, are leveraged to supervise the mon ocular depth estimation. More importantly, the pseudo depth labels serve to impo se a cross-view consistency on the estimated monocular depth and segmentation ma ps of two views. This enables for mitigating the mismatch problem incurred by in consistent prediction results across two views. A thorough ablation study valida tes that the cross-view consistency leads to a substantial performance gain by e nsuring inference-view invariance for the two tasks.
****
Stochastic Proximal Point Algorithm for Large-scale Nonconvex Optimization: Conv ergence, Implementation, and Application to Neural Networks
Aysegul Bumin,Kejun Huang
We revisit the stochastic proximal point algorithm (SPPA) for large-scale noncon

vex optimization problems. SPPA has been shown to converge faster and more stabl
e than the celebrated stochastic gradient descent (SGD) algorithm, and its many
variations, for convex problems. However, the per-iteration update of SPPA is de
fined abstractly and has long been considered expensive. In this paper, we show
that efficient implementation of SPPA can be achieved. If the problem is a nonli
near least squares, each iteration of SPPA can be efficiently implemented by Gau
ss-Newton; with some linear algebra trick the resulting complexity is in the sam
e order of SGD. For more generic problems, SPPA can still be implemented with L-
BFGS or accelerated gradient with high efficiency. Another contribution of this
work is the convergence of SPPA to a stationary point in expectation for nonconv
ex problems. The result is encouraging that it admits more flexible choices of t
he step sizes under similar assumptions. The proposed algorithm is elaborated fo
r both regression and classification problems using different neural network str
uctures. Real data experiments showcase its effectiveness in terms of convergenc
e and accuracy compared to SGD and its variants.
****

A Chaos Theory Approach to Understand Neural Network Optimization
Michele Sasdelli,Thalaiyasingam Ajanthan,Tat-Jun Chin,Gustavo Carneiro
Despite the complicated structure of modern deep neural network architectures, t
hey are still optimized with algorithms based on Stochastic Gradient Descent (SG
D). However, the reason behind the effectiveness of SGD is not well understood,
making its study an active research area. In this paper, we formulate deep neura
l network optimization as a dynamical system and show that the rigorous theory d
eveloped to study chaotic systems can be useful to understand SGD and its varian
ts. In particular, we first observe that the inverse of the instability timescal
e of SGD optimization, represented by the largest Lyapunov exponent, corresponds
to the most negative eigenvalue of the Hessian of the loss. This observation en
ables the introduction of an efficient method to estimate the largest eigenvalue
of the Hessian. Then, we empirically show that for a large range of learning ra
tes, SGD traverses the loss landscape across regions with largest eigenvalue of
the Hessian similar to the inverse of the learning rate. This explains why effec
tive learning rates can be found to be within a large range of values and shows
that SGD implicitly uses the largest eigenvalue of the Hessian while traversing
the loss landscape. This sheds some light on the effectiveness of SGD over more
sophisticated second-order methods. We also propose a quasi-Newton method that d
ynamically estimates an optimal learning rate for the optimization of deep learn
ing models. We demonstrate that our observations and methods are robust across d
ifferent architectures and loss functions on CIFAR-10 dataset.
****

Learning to Set Waypoints for Audio-Visual Navigation
Changan Chen,Sagnik Majumder,Ziad Al-Halah,Ruohan Gao,Santhosh Kumar Ramakrishna
n,Kristen Grauman
In audio-visual navigation, an agent intelligently travels through a complex, un
mapped 3D environment using both sights and sounds to find a sound source (e.g.,
a phone ringing in another room). Existing models learn to act at a fixed granu
larity of agent motion and rely on simple recurrent aggregations of the audio ob
servations. We introduce a reinforcement learning approach to audio-visual navig
ation with two key novel elements: 1) waypoints that are dynamically set and lea
rned end-to-end within the navigation policy, and 2) an acoustic memory that pro
vides a structured, spatially grounded record of what the agent has heard as it
moves. Both new ideas capitalize on the synergy of audio and visual data for rev
ealing the geometry of an unmapped space. We demonstrate our approach on two cha
llenging datasets of real-world 3D scenes, Replica and Matterport3D. Our model i
mproves the state of the art by a substantial margin, and our experiments reveal
that learning the links between sights, sounds, and space is essential for audi
o-visual navigation.
****

Log representation as an interface for log processing applications
Mohammad Amin Sadeghi,Shameem Parambath,Ji Lucas,Youssef Meguebli,Maguette Toure
,Fawaz Al Qahtani,Ting Yu,Sanjay Chawla

Log files are files that record events, messages, or transactions. Logs are rich containers of data because they can store a sequence of structured textual and numerical data. Many sequential forms of data including natural languages and temporal signals can be represented as logs.

We propose to represent logs at a few levels of abstraction including field level, log level, and log sequence level. The representation for each level can be computed from the previous level. These representations are in vector format and serve as interfaces to downstream applications. We use a version of transformer networks to encode numerical information as well as textual information that is suitable for log embedding. We show how a number of log processing applications can be readily solved with our representation.
****
Image Modeling with Deep Convolutional Gaussian Mixture Models
Alexander Gepperth,Benedikt Pfülb
In this conceptual work, we present DCGMM, a deep hierarchical Gaussian Mixture Model (GMM) that is particularly suited for describing and generating images.
Vanilla (i.e., "flat") GMMs require a very large number of components to well describe images, leading to long training times and memory issues.
DCGMMs avoid this by a stacked architecture of multiple GMM layers, linked by convolution and pooling operations.
This allows to exploit the compositionality of images in a similar way as deep CNNs do.
This sets them apart from vanilla GMMs which are trained by EM, requiring a prior k-means initialization which is infeasible in a layered structure.
For generating sharp images with DCGMM, we introduce a new gradient-based technique for sampling through non-invertible operations like convolution and pooling.
Based on the MNIST and FashionMNIST datasets, we validate the DCGMM model by demonstrating its superiority over "flat" GMMs for clustering, sampling and outlier detection.
We additionally demonstrate the applicability of DCGMM to variant generation, in-painting and class-conditional sampling.
****
Towards Finding Longer Proofs
Zsolt Zombori,Adrián Csiszárik,Henryk Michalewski,Cezary Kaliszyk,Josef Urban
We present a reinforcement learning (RL) based guidance system for automated theorem proving geared towards Finding Longer Proofs (FLoP). FLoP is a step towards learning to reason by analogy, reducing the dependence on large scale search in automated theorem provers. We use several simple, structured datasets with very long proofs to show that FLoP can successfully generalise a single training proof to a large class of related problems, implementing a simple form of analogical reasoning. On these benchmarks, FLoP is competitive with strong theorem provers despite using very limited search.
****
GL-Disen: Global-Local disentanglement for unsupervised learning of graph-level representations
Thilini Cooray,Ngai-man Cheung,Wei Lu
Graph-level representation learning plays a crucial role in a variety of tasks such as molecular property prediction and community analysis. Currently, several models based on mutual information maximization have shown strong performance on the task of unsupervised graph representation learning. In this paper, instead, we consider a disentanglement approach to learn graph-level representations in the unsupervised setting. Our work is the first to study disentanglement learning for graph-level representations. Our key observation is that the formation of many real-world graphs is a complex process with global and local generative factors. We hypothesize that disentangled representations which capture these global and local generative factors into independent latent units can be highly beneficial. Specifically, for graph-level representation learning, our disentanglement approach can alleviate distraction due to local variations of individual nodes or individual local neighbourhoods. We propose a VAE based learning algorithm

to disentangle the global graph-level information, which is common across the en
tire graph, and local patch-level information, which varies across individual pa
tches (the local subgraphs centered around the nodes). Through extensive experim
ents and analysis, we show that our method achieves the state-of-the-art perform
ance on the task of unsupervised graph representation learning.

****
Disambiguating Symbolic Expressions in Informal Documents
Dennis Müller,Cezary Kaliszyk
We propose the task of \emph{disambiguating} symbolic expressions in informal ST
EM documents in the form of \LaTeX files -- that is, determining their precise s
emantics and abstract syntax tree -- as a neural machine translation task. We di
scuss the distinct challenges involved and present a dataset with roughly 33,000
 entries. We evaluated several baseline models on this dataset, which failed to
yield even syntactically valid \LaTeX before overfitting. Consequently, we descr
ibe a methodology using a \emph{transformer} language model pre-trained on sourc
es obtained from \url{arxiv.org}, which yields promising results despite the sma
ll size of the dataset. We evaluate our model using a plurality of dedicated tec
hniques, taking syntax and semantics of symbolic expressions into account.
****
Hard Masking for Explaining Graph Neural Networks
Thorben Funke,Megha Khosla,Avishek Anand
Graph Neural Networks (GNNs) are a flexible and powerful family of models that b
uild nodes' representations on irregular graph-structured data. This paper focus
es on explaining or interpreting the rationale underlying a given prediction of
already trained graph neural networks for the node classification task. Existing
 approaches for interpreting GNNs try to find subsets of important features and
nodes by learning a continuous mask. Our objective is to find discrete masks tha
t are arguably more interpretable while minimizing the expected deviation from t
he underlying model's prediction. We empirically show that our explanations are
both more predictive and sparse. Additionally, we find that multiple diverse exp
lanations are possible, which sufficiently explain a prediction. Finally, we ana
lyze the explanations to find the effect of network homophily on the decision-ma
king process of GNNs.
****
Single Layers of Attention Suffice to Predict Protein Contacts
Nick Bhattacharya,Neil Thomas,Roshan Rao,Justas Daupras,Peter K Koo,David Baker,
Yun S. Song,Sergey Ovchinnikov
The established approach to unsupervised protein contact prediction estimates co
evolving positions using undirected graphical models. This approach trains a Pot
ts model on a Multiple Sequence Alignment, then predicts that the edges with hig
hest weight correspond to contacts in the 3D structure. On the other hand, incre
asingly large Transformers are being pretrained on protein sequence databases bu
t have demonstrated mixed results for downstream tasks, including contact predic
tion. This has sparked discussion about the role of scale and attention-based mo
dels in unsupervised protein representation learning. We argue that attention is
 a principled model of protein interactions, grounded in real properties of prot
ein family data. We introduce a simplified attention layer, factored attention,
and show that it achieves comparable performance to Potts models, while sharing
parameters both within and across families. Further, we extract contacts from th
e attention maps of a pretrained Transformer and show they perform competitively
 with the other two approaches. This provides evidence that large-scale pretrain
ing can learn meaningful protein features when presented with unlabeled and unal
igned data. We contrast factored attention with the Transformer to indicate that
 the Transformer leverages hierarchical signal in protein family databases not c
aptured by our single-layer models. This raises the exciting possibility for the
 development of powerful structured models of protein family databases.
****
Trust, but verify: model-based exploration in sparse reward environments
Konrad Czechowski,Tomasz Odrzygó■d■,Micha■ Izworski,Marek Zbysi■ski,■ukasz Kuci■

ski,Piotr Mi■o■
We propose $\textit{trust-but-verify}$ (TBV) mechanism, a new method which uses model uncertainty estimates to guide exploration. The mechanism augments graph search planning algorithms by the capacity to deal with learned model's imperfections. We identify certain type of frequent model errors, which we dub $\textit{false loops}$, and which are particularly dangerous for graph search algorithms in discrete environments. These errors impose falsely pessimistic expectations and thus hinder exploration. We confirm this experimentally and show that TBV can effectively alleviate them. TBV combined with MCTS or Best First Search forms an effective model-based reinforcement learning solution, which is able to robustly solve sparse reward problems.
****
Visual Question Answering From Another Perspective: CLEVR Mental Rotation Tests
Christopher Beckham,Martin Weiss,Florian Golemo,Sina Honari,Derek Nowrouzezahrai,Christopher Pal
Different types of \emph{mental rotation tests} have been used extensively in psychology to understand human visual reasoning and perception. Understanding what an object or visual scene would look like from another viewpoint is a challenging problem that is made even harder if it must be performed from a single image. 3D computer vision has a long history of examining related problems. However, often what one is most interested in is the answer to a relatively simple question posed in another visual frame of reference -- as opposed to creating a full 3D reconstruction.
Mental rotations tests can also manifest as consequential questions in the real world such as: does the pedestrian that I see, see the car that I am driving?
We explore a controlled setting whereby questions are posed about the properties of a scene if the scene were observed from another viewpoint. To do this we have created a new version of the CLEVR VQA problem setup and dataset that we call CLEVR Mental Rotation Tests or CLEVR-MRT, where the goal is to answer questions about the original CLEVR viewpoint given a single image obtained from a different viewpoint of the same scene. Using CLEVR Mental Rotation Tests we examine standard state of the art methods, show how they fall short, then explore novel neural architectures that involve inferring representations encoded as feature volumes describing a scene. Our new methods use rigid transformations of feature volumes conditioned on the viewpoint camera. We examine the efficacy of different model variants through performing a rigorous ablation study. Furthermore, we examine the use of contrastive learning to infer a volumetric encoder in a self-supervised manner and find that this approach yields the best results of our study using CLEVR-MRT.
****
Deep Gated Canonical Correlation Analysis
Ofir Lindenbaum,Moshe Salhov,Amir Averbuch,Yuval Kluger
Canonical Correlation Analysis (CCA) models can extract informative correlated representations from multimodal unlabelled data. Despite their success, CCA models may break if the number of variables exceeds the number of samples. We propose Deep Gated-CCA, a method for learning correlated representations based on a sparse subset of variables from two observed modalities. The proposed procedure learns two non-linear transformations and simultaneously gates the input variables to identify a subset of most correlated variables. The non-linear transformations are learned by training two neural networks to maximize a shared correlation loss defined based on their outputs. Gating is obtained by adding an approximate ℓ_0 regularization term applied to the input variables. This approximation relies on a recently proposed continuous Gaussian based relaxation for Bernoulli variables which act as gates. We demonstrate the efficacy of the method using several synthetic and real examples. Most notably, the method outperforms other linear and non-linear CCA models.
****
Colorization Transformer
Manoj Kumar,Dirk Weissenborn,Nal Kalchbrenner
We present the Colorization Transformer, a novel approach for diverse high fidel

ity image colorization based on self-attention. Given a grayscale image, the colorization proceeds in three steps. We first use a conditional autoregressive transformer to produce a low resolution coarse coloring of the grayscale image. Our architecture adopts conditional transformer layers to effectively condition grayscale input. Two subsequent fully parallel networks upsample the coarse colored low resolution image into a finely colored high resolution image. Sampling from the Colorization Transformer produces diverse colorings whose fidelity outperforms the previous state-of-the-art on colorising ImageNet based on FID results and based on a human evaluation in a Mechanical Turk test. Remarkably, in more than 60\% of cases human evaluators prefer the highest rated among three generated colorings over the ground truth. The code and pre-trained checkpoints for Colorization Transformer are publicly available at https://github.com/google-research/google-research/tree/master/coltran
****
More Side Information, Better Pruning: Shared-Label Classification as a Case Study
Omer Leibovitch,Nir Ailon
Pruning of neural networks, also known as compression or sparsification, is the task of converting a given network, which may be too expensive to use (in prediction) on low resource platforms, with another 'lean' network which performs almost as well as the original one, while using considerably fewer resources. By turning the compression ratio knob, the practitioner can trade off the information gain versus the necessary computational resources, where information gain is a measure of reduction of uncertainty in the prediction.

In certain cases, however, the practitioner may readily possess some information on the prediction from other sources. The main question we study here is, whether it is possible to take advantage of the additional side information, in order to further reduce the computational resources, in tandem with the pruning process?

Motivated by a real-world application, we distill the following elegantly stated problem. We are given a multi-class prediction problem, combined with a (possibly pre-trained) network architecture for solving it on a given instance distribution, and also a method for pruning the network to allow trading off prediction speed with accuracy. We assume the network and the pruning methods are state-of-the-art, and it is not our goal here to improve them. However, instead of being asked to predict a single drawn instance x, we are being asked to predict the label of an n-tuple of instances $(x_1,\dots x_n)$, with the additional side information of all tuple instances share the same label. The shared label distribution is identical to the distribution on which the network was trained.

One trivial way to do this is by obtaining individual raw predictions for each of the n instances (separately), using our given network, pruned for a desired accuracy, then taking the average to obtain a single more accurate prediction. This is simple to implement but intuitively sub-optimal, because the n independent instantiations of the network do not share any information, and would probably waste resources on overlapping computation.

We propose various methods for performing this task, and compare them using extensive experiments on public benchmark data sets for image classification. Our comparison is based on measures of relative information (RI) and n-accuracy, which we define. Interestingly, we empirically find that I) sharing information between the n independently computed hidden representations of $x_1,..,x_n$, using an LSTM based gadget, performs best, among all methods we experiment with, ii) for all methods studied, we exhibit a sweet spot phenomenon, which sheds light on the compression-information trade-off and may assist a practitioner to choose the desired compression ratio.
****
Gradient-based training of Gaussian Mixture Models for High-Dimensional Streamin

g Data
Alexander Gepperth,Benedikt Pfülb
We present an approach for efficiently training Gaussian Mixture Models by SGD o
n non-stationary, high-dimensional streaming data.
Our training scheme does not require data-driven parameter initialization (e.g.,
 k-means) and has the ability to process high-dimensional samples without numeri
cal problems.
Furthermore, the approach allows mini-batch sizes as low as 1, typical for strea
ming-data settings, and it is possible to react and adapt to changes in data sta
tistics (concept drift/shift) without catastrophic forgetting.
Major problems in such streaming-data settings are undesirable local optima duri
ng early training phases and numerical instabilities due to high data dimensiona
lities.%, and catastrophic forgetting when encountering concept drift.
We introduce an adaptive annealing procedure to address the first problem,%, whi
ch additionally plays a decisive role in controlling the \acp{GMM}' reaction to
concept drift.
whereas numerical instabilities are eliminated by using an exponential-free appr
oximation to the standard \ac{GMM} log-likelihood.
Experiments on a variety of visual and non-visual benchmarks show that our SGD a
pproach can be trained completely without, for instance, k-means based centroid
initialization, and compares favorably to sEM, an online variant of EM.
****
Succinct Network Channel and Spatial Pruning via Discrete Variable QCQP
Yeonwoo Jeong,Deokjae Lee,Gaon An,Changyong Son,Hyun Oh Song
    Reducing the heavy computational cost of large convolutional neural networks
 is crucial when deploying the networks to resource-constrained environments. In
 this context, recent works propose channel pruning via greedy channel selection
 to achieve practical acceleration and memory footprint reduction. We first show
 this channel-wise approach ignores the inherent quadratic coupling between chan
nels in the neighboring layers and cannot safely remove inactive weights during
the pruning procedure. Furthermore, we show that these pruning methods cannot gu
arantee the given resource constraints are satisfied and cause discrepancy with
the true objective. To this end, we formulate a principled optimization framewor
k with discrete variable QCQP, which provably prevents any inactive weights and
 enables the exact guarantee of meeting the resource constraints in terms of FLO
Ps and memory. Also, we extend the pruning granularity beyond channels and joint
ly prune individual 2D convolution filters spatially for greater efficiency. Our
 experiments show competitive pruning results under the target resource constrai
nts on CIFAR-10 and ImageNet datasets on various network architectures.

****
Theoretical bounds on estimation error for meta-learning
James Lucas,Mengye Ren,Irene Raissa KAMENI KAMENI,Toniann Pitassi,Richard Zemel
Machine learning models have traditionally been developed under the assumption t
hat the training and test distributions match exactly. However, recent success i
n few-shot learning and related problems are encouraging signs that these models
 can be adapted to more realistic settings where train and test distributions di
ffer. Unfortunately, there is severely limited theoretical support for these alg
orithms and little is known about the difficulty of these problems. In this work
, we provide novel information-theoretic lower-bounds on minimax rates of conver
gence for algorithms that are trained on data from multiple sources and tested o
n novel data. Our bounds depend intuitively on the information shared between so
urces of data, and characterize the difficulty of learning in this setting for a
rbitrary algorithms. We demonstrate these bounds on a hierarchical Bayesian mode
l of meta-learning, computing both upper and lower bounds on parameter estimatio
n via maximum-a-posteriori inference.
****
InstantEmbedding: Efficient Local Node Representations
Stefan Postavaru,Anton Tsitsulin,Filipe Miguel Goncalves de Almeida,Yingtao Tian
,Silvio Lattanzi,Bryan Perozzi

In this paper, we introduce InstantEmbedding, an efficient method for generating single-node representations using local PageRank computations. We prove that our approach produces globally consistent representations in sublinear time. We demonstrate this empirically by conducting extensive experiments on real-world datasets with over a billion edges. Our experiments confirm that InstantEmbedding requires drastically less computation time (over 9,000 times faster) and less memory (by over 8,000 times) to produce a single node's embedding than traditional methods including DeepWalk, node2vec, VERSE, and FastRP. We also show that our method produces high quality representations, demonstrating results that meet or exceed the state of the art for unsupervised representation learning on tasks like node classification and link prediction.
****

Learning Robust Models using the Principle of Independent Causal Mechanisms
Jens Müller,Robert Schmier,Lynton Ardizzone,Carsten Rother,Ullrich Koethe
Standard supervised learning breaks down under data distribution shift. However, the principle of independent causal mechanisms (ICM, Peters et al. (2017)) can turn this weakness into an opportunity: one can take advantage of distribution shift between different environments during training in order to obtain more robust models. We propose a new gradient-based learning framework whose objective function is derived from the ICM principle. We show theoretically and experimentally that neural networks trained in this framework focus on relations remaining invariant across environments and ignore unstable ones. Moreover, we prove that the recovered stable relations correspond to the true causal mechanisms under certain conditions. In both regression and classification, the resulting models generalize well to unseen scenarios where traditionally trained models fail.
****

Implicit Normalizing Flows
Cheng Lu,Jianfei Chen,Chongxuan Li,Qiuhao Wang,Jun Zhu
Normalizing flows define a probability distribution by an explicit invertible transformation $\boldsymbol{\mathbf{z}}=f(\boldsymbol{\mathbf{x}})$. In this work, we present implicit normalizing flows (ImpFlows), which generalize normalizing flows by allowing the mapping to be implicitly defined by the roots of an equation $F(\boldsymbol{\mathbf{z}}, \boldsymbol{\mathbf{x}})= \boldsymbol{\mathbf{0}}$. ImpFlows build on residual flows (ResFlows) with a proper balance between expressiveness and tractability. Through theoretical analysis, we show that the function space of ImpFlow is strictly richer than that of ResFlows. Furthermore, for any ResFlow with a fixed number of blocks, there exists some function that ResFlow has a non-negligible approximation error. However, the function is exactly representable by a single-block ImpFlow. We propose a scalable algorithm to train and draw samples from ImpFlows. Empirically, we evaluate ImpFlow on several classification and density modeling tasks, and ImpFlow outperforms ResFlow with a comparable amount of parameters on all the benchmarks.
****

Ablation Path Saliency
Olivier Verdier,Justus Sagemüller
We consider the saliency problem for black-box classification. In image classification, this means highlighting the part of the image that is most relevant for the current decision.
We cast the saliency problem as finding an optimal ablation path between two images. An ablation path consists of a sequence of ever smaller masks, joining the current image to a reference image in another decision region. The optimal path will stay as long as possible in the current decision region. This approach extends the ablation tests in [Sturmfels et al. (2020)]. The gradient of the corresponding objective function is closely related to the integrated gradient method [Sundararajan et al. (2017)]. In the saturated case (when the classifier outputs a binary value) our method would reduce to the meaningful perturbation approach [Fong & Vedaldi (2017)], since crossing the decision boundary as late as possible would then be equivalent to finding the smallest possible mask lying on the decision boundary.
Our interpretation provides geometric understanding of existing saliency methods

, and suggests a novel approach based on ablation path optimisation.
****

NETWORK ROBUSTNESS TO PCA PERTURBATIONS

Anan Kabaha,Dana Drachsler Cohen

A key challenge in analyzing neural networks' robustness is identifying input features for which networks are robust to perturbations. Existing work focuses on direct perturbations to the inputs, thereby studies network robustness to the lowest-level features. In this work, we take a new approach and study the robustness of networks to the inputs' semantic features. We show a black-box approach to determine features for which a network is robust or weak. We leverage these features to obtain provably robust neighborhoods defined using robust features and adversarial examples defined by perturbing weak features. We evaluate our approach with PCA features. We show (1) provably robust neighborhoods are larger: on average by 1.8x and up to 4.5x, compared to the standard neighborhoods, and (2) our adversarial examples are generated using at least 8.7x fewer queries and have at least 2.8x lower L2 distortion compared to state-of-the-art. We further show that our attack is effective even against ensemble adversarial training.
****

ARELU: ATTENTION-BASED RECTIFIED LINEAR UNIT

Chen Dengsheng,Jun Li,Kai Xu

Element-wise activation functions play a critical role in deep neural networks via affecting the expressivity power and the learning dynamics. Learning-based activation functions have recently gained increasing attention and success. We propose a new perspective of learnable activation function through formulating them with element-wise attention mechanism. In each network layer, we devise an attention module which learns an element-wise, sign-based attention map for the pre-activation feature map. The attention map scales an element based on its sign. Adding the attention module with a rectified linear unit (ReLU) results in an amplification of positive elements and a suppression of negative ones, both with learned, data-adaptive parameters. We coin the resulting activation function Attention-based Rectified Linear Unit (AReLU). The attention module essentially learns an element-wise residue of the activated part of the input, as ReLU can be viewed as an identity transformation. This makes the network training more resis- tant to gradient vanishing. The learned attentive activation leads to well-focused activation of relevant regions of a feature map. Through extensive evaluations, we show that AReLU significantly boosts the performance of most mainstream network architectures with only two extra learnable parameters per layer introduced. Notably, AReLU facilitates fast network training under small learning rates, which makes it especially suited in the case of transfer learning and meta learning.
****

Fast 3D Acoustic Scattering via Discrete Laplacian Based Implicit Function Encoders

Hsien-Yu Meng,Zhenyu Tang,Dinesh Manocha

Acoustic properties of objects corresponding to scattering characteristics are frequently used for 3D audio content creation, environmental acoustic effects, localization and acoustic scene analysis, etc. The numeric solvers used to compute these acoustic properties are too slow for interactive applications. We present a novel geometric deep learning algorithm based on discrete-laplacian and implicit encoders to compute these characteristics for rigid or deformable objects at interactive rates. We use a point cloud approximation of each object, and each point is encoded in a high-dimensional latent space. Our multi-layer network can accurately estimate these acoustic properties for arbitrary topologies and takes less than 1ms per object on a NVIDIA GeForce RTX 2080 Ti GPU. We also prove that our learning method is permutation and rotation invariant and demonstrate high accuracy on objects that are quite different from the training data. We highlight its application to generating environmental acoustic effects in dynamic environments.
****

XLVIN: eXecuted Latent Value Iteration Nets

Andreea Deac,Petar Veli■kovi■,Ognjen Milinkovic,Pierre-Luc Bacon,Jian Tang,Mladen Nikolic
Value Iteration Networks (VINs) have emerged as a popular method to perform implicit planning within deep reinforcement learning, enabling performance improvements on tasks requiring long-range reasoning and understanding of environment dynamics. This came with several limitations, however: the model is not explicitly incentivised to perform meaningful planning computations, the underlying state space is assumed to be discrete, and the Markov decision process (MDP) is assumed fixed and known. We propose eXecuted Latent Value Iteration Networks (XLVINs), which combine recent developments across contrastive self-supervised learning, graph representation learning and neural algorithmic reasoning to alleviate all of the above limitations, successfully deploying VIN-style models on generic environments. XLVINs match the performance of VIN-like models when the underlying MDP is discrete, fixed and known, and provide significant improvements to model-free baselines across three general MDP setups.
****
On the Importance of Sampling in Training GCNs: Convergence Analysis and Variance Reduction
Weilin Cong,Morteza Ramezani,Mehrdad Mahdavi
Graph Convolutional Networks (GCNs) have achieved impressive empirical advancement across a wide variety of graph-related applications. Despite their great success, training GCNs on large graphs suffers from computational and memory issues. A potential path to circumvent these obstacles is sampling-based methods, where at each layer a subset of nodes is sampled. Although recent studies have empirically demonstrated the effectiveness of sampling-based methods, these works lack theoretical convergence guarantees under realistic settings and cannot fully leverage the information of evolving parameters during optimization. In this paper, we describe and analyze a general \textbf{\textit{doubly variance reduction}} schema that can accelerate any sampling method under the memory budget. The motivating impetus for the proposed schema is a careful analysis for the variance of sampling methods where it is shown that the induced variance can be decomposed into node embedding approximation variance (\emph{zeroth-order variance}) during forward propagation and layerwise-gradient variance (\emph{first-order variance}) during backward propagation. We theoretically analyze the convergence of the proposed schema and show that it enjoys an $\mathcal{O}(1/T)$ convergence rate. We complement our theoretical results by integrating the proposed schema in different sampling methods and applying them to different large real-world graphs.
****
Quantifying and Learning Disentangled Representations with Limited Supervision
Loek Tonnaer,Luis Armando Pérez Rey,Vlado Menkovski,Mike Holenderski,Jacobus W. Portegies
Learning low-dimensional representations that disentangle the underlying factors of variation in data has been posited as an important step towards interpretable machine learning with good generalization. To address the fact that there is no consensus on what disentanglement entails, Higgins et al. (2018) propose a formal definition for Linear Symmetry-Based Disentanglement, or LSBD, arguing that underlying real-world transformations give exploitable structure to data.

Although several works focus on learning LSBD representations, such methods require supervision on the underlying transformations for the entire dataset, and cannot deal with unlabeled data. Moreover, none of these works provide a metric to quantify LSBD.

We propose a metric to quantify LSBD representations that is easy to compute under certain well-defined assumptions. Furthermore, we present a method that can leverage unlabeled data, such that LSBD representations can be learned with limited supervision on transformations. Using our LSBD metric, our results show that limited supervision is indeed sufficient to learn LSBD representations.
****
Variational Information Bottleneck for Effective Low-Resource Fine-Tuning

Rabeeh Karimi mahabadi,Yonatan Belinkov,James Henderson
While large-scale pretrained language models have obtained impressive results wh
en fine-tuned on a wide variety of tasks, they still often suffer from overfitti
ng in low-resource scenarios. Since such models are general-purpose feature extr
actors, many of these features are inevitably irrelevant for a given target task
. We propose to use Variational Information Bottleneck (VIB) to suppress irrele
vant features when fine-tuning on low-resource target tasks, and show that our m
ethod successfully reduces overfitting. Moreover, we show that our VIB model fi
nds sentence representations that are more robust to biases in natural language
inference datasets, and thereby obtains better generalization to out-of-domain d
atasets. Evaluation on seven low-resource datasets in different tasks shows that
our method significantly improves transfer learning in low-resource scenarios,
surpassing prior work. Moreover, it improves generalization on 13 out of 15 out-
of-domain natural language inference benchmarks. Our code is publicly available
in https://github.com/rabeehk/vibert.
****

TropEx: An Algorithm for Extracting Linear Terms in Deep Neural Networks
Martin Trimmel,Henning Petzka,Cristian Sminchisescu
Deep neural networks with rectified linear (ReLU) activations are piecewise line
ar functions, where hyperplanes partition the input space into an astronomically
high number of linear regions. Previous work focused on counting linear regions
to measure the network's expressive power and on analyzing geometric properties
of the hyperplane configurations. In contrast, we aim to understand the impact
of the linear terms on network performance, by examining the information encoded
in their coefficients. To this end, we derive TropEx, a nontrivial tropical alg
ebra-inspired algorithm to systematically extract linear terms based on data. Ap
plied to convolutional and fully-connected networks, our algorithm uncovers sign
ificant differences in how the different networks utilize linear regions for gen
eralization. This underlines the importance of systematic linear term exploratio
n, to better understand generalization in neural networks trained with complex d
ata sets.
****

Seq2Tens: An Efficient Representation of Sequences by Low-Rank Tensor Projection
s
Csaba Toth,Patric Bonnier,Harald Oberhauser
Sequential data such as time series, video, or text can be challenging to analys
e as the ordered structure gives rise to complex dependencies. At the heart of t
his is non-commutativity, in the sense that reordering the elements of a sequenc
e can completely change its meaning. We use a classical mathematical object -- t
he free algebra -- to capture this non-commutativity. To address the innate comp
utational complexity of this algebra, we use compositions of low-rank tensor pro
jections. This yields modular and scalable building blocks that give state-of-th
e-art performance on standard benchmarks such as multivariate time series classi
fication, mortality prediction and generative models for video.
****

Self-Organizing Intelligent Matter: A blueprint for an AI generating algorithm
Karol Gregor,Frederic Besse
We propose an artificial life framework aimed at facilitating the emergence of i
ntelligent organisms. In this framework there is no explicit notion of an agent:
instead there is an environment made of atomic elements. These elements contain
neural operations and interact through exchanges of information and through phy
sics-like rules contained in the environment. We discuss how an evolutionary pro
cess can lead to the emergence of different organisms made of many such atomic e
lements which can coexist and thrive in the environment. We discuss how this for
ms the basis of a general AI generating algorithm. We provide a simplified imple
mentation of such system and discuss what advances need to be made to scale it u
p further.
****

Representation learning for improved interpretability and classification accurac
y of clinical factors from EEG

Garrett Honke,Irina Higgins,Nina Thigpen,Vladimir Miskovic,Katie Link,Sunny Duan,Pramod Gupta,Julia Klawohn,Greg Hajcak
Despite extensive standardization, diagnostic interviews for mental health disorders encompass substantial subjective judgment. Previous studies have demonstrated that EEG-based neural measures can function as reliable objective correlates of depression, or even predictors of depression and its course. However, their clinical utility has not been fully realized because of 1) the lack of automated ways to deal with the inherent noise associated with EEG data at scale, and 2) the lack of knowledge of which aspects of the EEG signal may be markers of a clinical disorder. Here we adapt an unsupervised pipeline from the recent deep representation learning literature to address these problems by 1) learning a disentangled representation using β-VAE to denoise the signal, and 2) extracting interpretable features associated with a sparse set of clinical labels using a Symbol-Concept Association Network (SCAN). We demonstrate that our method is able to outperform the canonical hand-engineered baseline classification method on a number of factors, including participant age and depression diagnosis. Furthermore, our method recovers a representation that can be used to automatically extract denoised Event Related Potentials (ERPs) from novel, single EEG trajectories, and supports fast supervised re-mapping to various clinical labels, allowing clinicians to re-use a single EEG representation regardless of updates to the standardized diagnostic system. Finally, single factors of the learned disentangled representations often correspond to meaningful markers of clinical factors, as automatically detected by SCAN, allowing for human interpretability and post-hoc expert analysis of the recommendations made by the model.
****

Shuffle to Learn: Self-supervised learning from permutations via differentiable ranking
Andrew N Carr,Quentin Berthet,Mathieu Blondel,Olivier Teboul,Neil Zeghidour
Self-supervised pre-training using so-called "pretext" tasks has recently shown impressive performance across a wide range of tasks. In this work we advance self-supervised learning from permutations, that consists in shuffling parts of input and training a model to reorder them, improving downstream performance in classification. To do so, we overcome the main challenges of integrating permutation inversions (a discontinuous operation) into an end-to-end training scheme, heretofore sidestepped by casting the reordering task as classification, fundamentally reducing the space of permutations that can be exploited. These advances rely on two main, independent contributions. First, we use recent advances in differentiable ranking to integrate the permutation inversion flawlessly into a neural network, enabling us to use the full set of permutations, at no additional computing cost. Our experiments validate that learning from all possible permutations (up to 10^{18}) improves the quality of the pre-trained representations over using a limited, fixed set. Second, we successfully demonstrate that inverting permutations is a meaningful pretext task in a diverse range of modalities, beyond images, which does not require modality-specific design. In particular, we also improve music understanding by reordering spectrogram patches in the frequency space, as well as video classification by reordering frames along the time axis. We furthermore analyze the influence of the patches that we use (vertical, horizontal, 2-dimensional), as well as the benefit of our approach in different data regimes.
****

Minimum Description Length Recurrent Neural Networks
Nur Lan,Emmanuel Chemla,Roni Katzir
Recurrent neural networks (RNNs) face two well-known challenges: (a) the difficulty of such networks to generalize appropriately as opposed to memorizing, especially from very short input sequences (generalization); and (b) the difficulty for us to understand the knowledge that the network has attained (transparency). We explore the implications to these challenges of employing a general search through neural architectures using a genetic algorithm with Minimum Description Length (MDL) as an objective function. We find that MDL leads the networks to reach adequate levels of generalization from very small corpora, improving over back

propagation-based alternatives. We demonstrate this approach by evolving networks which perform tasks of increasing complexity with absolute correctness. The resulting networks are small, easily interpretable, and unlike classical RNNs, are provably appropriate for sequences of arbitrary length even when trained on very limited corpora. One case study is addition, for which our system grows a network with just four cells, reaching 100% accuracy (and at least .999 certainty) for arbitrary large numbers.
****
AUL is a better optimization metric in PU learning
Shangchuan Huang,Songtao Wang,Dan Li,Liwei Jiang
Traditional binary classification models are trained and evaluated with fully labeled data which is not common in real life. In non-ideal dataset, only a small fraction of positive data are labeled. Training a model from such partially labeled data is named as positive-unlabeled (PU) learning. A naive solution of PU learning is treating unlabeled samples as negative. However, using biased data, the trained model may converge to non-optimal point and its real performance cannot be well estimated. Recent works try to recover the unbiased result by estimating the proportion of positive samples with mixture proportion estimation (MPE) algorithms, but the model performance is still limited and heavy computational cost is introduced (particularly for big datasets). In this work, we theoretically prove that Area Under Lift curve (AUL) is an unbiased metric in PU learning scenario, and the experimental evaluation on 9 datasets shows that the average absolute error of AUL estimation is only 1/6 of AUC estimation. By experiments we also find that, compared with state-of-the-art AUC-optimization algorithm, AULoptimization algorithm can not only significantly save the computational cost, but also improve the model performance by up to 10%.
****
Variational Deterministic Uncertainty Quantification
Joost van Amersfoort,Lewis Smith,Andrew Jesson,Oscar Key,Yarin Gal
Building on recent advances in uncertainty quantification using a single deep deterministic model (DUQ), we introduce variational Deterministic Uncertainty Quantification (vDUQ). We overcome several shortcomings of DUQ by recasting it as a Gaussian process (GP) approximation. Our principled approximation is based on an inducing point GP in combination with Deep Kernel Learning. This enables vDUQ to use rigorous probabilistic foundations, and work not only on classification but also on regression problems. We avoid uncertainty collapse away from the training data by regularizing the spectral norm of the deep feature extractor. Our method matches SotA accuracy, 96.2\% on CIFAR-10, while maintaining the speed of softmax models, and provides uncertainty estimates competitive with Deep Ensembles. We demonstrate our method in regression problems and by estimating uncertainty in causal inference for personalized medicine
****
ChemistryQA: A Complex Question Answering Dataset from Chemistry
Zhuoyu Wei,Wei Ji,Xiubo Geng,Yining Chen,Baihua Chen,Tao Qin,Daxin Jiang
Many Question Answering (QA) tasks have been studied in NLP and employed to evaluate the progress of machine intelligence. One kind of QA tasks, such as Machine Reading Comprehension QA, is well solved by end-to-end neural networks; another kind of QA tasks, such as Knowledge Base QA, needs to be translated to a formatted representations and then solved by a well-designed solver. We notice that some real-world QA tasks are more complex, which cannot be solved by end-to-end neural networks or translated to any kind of formal representations. To further stimulate the research of QA and development of QA techniques, in this work, we create a new and complex QA dataset, ChemistryQA, based on real-world chemical calculation questions. To answer chemical questions, machines need to understand questions, apply chemistry and Math knowledge, and do calculation and reasoning. To help researchers ramp up, we build two baselines: the first one is BERT-based sequence to sequence model, and the second one is an extraction system plus a graph search based solver. These two methods achieved 0.164 and 0.169 accuracy on the development set, respectively, which clearly demonstrate that new techniques are needed for complex QA tasks. ChemistryQA dataset will be available for p

ublic download once the paper is published.
****
Neural Random Projection: From the Initial Task To the Input Similarity Problem
Alan Savushkin,Nikita Benkovich,Dmitry Golubev
The data representation plays an important role in evaluating similarity between objects. In this paper, we propose a novel approach for implicit data representation to evaluate similarity of input data using a trained neural network. In contrast to the previous approach, which uses gradients for representation, we utilize only the outputs from the last hidden layer of a neural network and do not use a backward step. The proposed technique explicitly takes into account the initial task and significantly reduces the size of the vector representation, as well as the computation time. Generally, a neural network obtains representations related only to the problem being solved, which makes the last hidden layer representation useless for input similarity task.
In this paper, we consider two reasons for the decline in the quality of representations: correlation between neurons and insufficient size of the last hidden layer. To reduce the correlation between neurons we use orthogonal weight initialization for each layer and modify the loss function to ensure orthogonality of the weights during training. Moreover, we show that activation functions can potentially increase correlation. To solve this problem, we apply modified Batch-Normalization with Dropout. Using orthogonal weight matrices allow us to consider such neural networks as an application of the Random Projection method and get a lower bound estimate for the size of the last hidden layer. We perform experiments on MNIST and physical examination datasets. In both experiments, initially, we split a set of labels into two disjoint subsets to train a neural network for binary classification problem, and then use this model to measure similarity between input data and define hidden classes. We also cluster the inputs to evaluate how well objects from the same hidden class are grouped together. Our experimental results show that the proposed approach achieves competitive results on the input similarity task while reducing both computation time and the size of the input representation.
****
Generalizing and Tensorizing Subgraph Search in the Supernet
Hansi Yang,quanming yao
Recently, a special kind of graph, i.e., supernet, which allows two nodes connected by multi-choice edges, has exhibited its power in neural architecture search (NAS) by searching better architectures for computer vision (CV) and natural language processing (NLP) tasks. In this paper, we discover that the design of such discrete architectures also appears in many other important learning tasks, e.g., logical chain inference in knowledge graphs (KGs) and meta-path discovery in heterogeneous information networks (HINs). Thus, we are motivated to generalize the supernet search problem on a broader horizon. However, none of the existing works are effective since the supernet's topology is highly task-dependent and diverse. To address this issue, we propose to tensorize the supernet, i.e. unify the subgraph search problems by a tensor formulation and encode the topology inside the supernet by a tensor network. We further propose an efficient algorithm that admits both stochastic and deterministic objectives to solve the search problem. Finally, we perform extensive experiments on diverse learning tasks, i.e., architecture design for CV, logic inference for KG, and meta-path discovery for HIN. Empirical results demonstrate that our method leads to better performance and architectures.

****
Identifying Treatment Effects under Unobserved Confounding by Causal Representation Learning
Pengzhou Abel Wu,Kenji Fukumizu
As an important problem of causal inference, we discuss the estimation of treatment effects under the existence of unobserved confounding. By representing the confounder as a latent variable, we propose Counterfactual VAE, a new variant of variational autoencoder, based on recent advances in identifiability of represen

tation learning. Combining the identifiability and classical identification resu lts of causal inference, under mild assumptions on the generative model and with small noise on the outcome, we theoretically show that the confounder is identi fiable up to an affine transformation and then the treatment effects can be iden tified. Experiments on synthetic and semi-synthetic datasets demonstrate that o ur method matches the state-of-the-art, even under settings violating our formal assumptions.
****
Learning disentangled representations with the Wasserstein Autoencoder
Benoit Gaujac,Ilya Feige,David Barber
Disentangled representation learning has undoubtedly benefited from objective fu nction surgery. However, a delicate balancing act of tuning is still required in order to trade off reconstruction fidelity versus disentanglement. Building on previous successes of penalizing the total correlation in the latent variables, we propose TCWAE (Total Correlation Wasserstein Autoencoder). Working in the WAE paradigm naturally enables the separation of the total-correlation term, thus p roviding disentanglement control over the learned representation, while offering more flexibility in the choice of reconstruction cost. We propose two variants using different KL estimators and perform extensive quantitative comparisons on data sets with known generative factors, showing competitive results relative to state-of-the-art techniques. We further study the trade off between disentangle ment and reconstruction on more-difficult data sets with unknown generative fact ors, where we expect improved reconstructions due to the flexibility of the WAE paradigm.
****
Language-Agnostic Representation Learning of Source Code from Structure and Cont ext
Daniel Zügner,Tobias Kirschstein,Michele Catasta,Jure Leskovec,Stephan Günnemann
Source code (Context) and its parsed abstract syntax tree (AST; Structure) are t wo complementary representations of the same computer program. Traditionally, de signers of machine learning models have relied predominantly either on Structure or Context. We propose a new model, which jointly learns on Context and Structu re of source code. In contrast to previous approaches, our model uses only langu age-agnostic features, i.e., source code and features that can be computed direc tly from the AST. Besides obtaining state-of-the-art on monolingual code summari zation on all five programming languages considered in this work, we propose the first multilingual code summarization model. We show that jointly training on n on-parallel data from multiple programming languages improves results on all ind ividual languages, where the strongest gains are on low-resource languages. Rema rkably, multilingual training only from Context does not lead to the same improv ements, highlighting the benefits of combining Structure and Context for represe ntation learning on code.
****
Generalized Multimodal ELBO
Thomas M. Sutter,Imant Daunhawer,Julia E Vogt
Multiple data types naturally co-occur when describing real-world phenomena and learning from them is a long-standing goal in machine learning research. However , existing self-supervised generative models approximating an ELBO are not able to fulfill all desired requirements of multimodal models: their posterior approx imation functions lead to a trade-off between the semantic coherence and the abi lity to learn the joint data distribution. We propose a new, generalized ELBO fo rmulation for multimodal data that overcomes these limitations. The new objectiv e encompasses two previous methods as special cases and combines their benefits without compromises. In extensive experiments, we demonstrate the advantage of t he proposed method compared to state-of-the-art models in self-supervised, gener ative learning tasks.
****
Saliency Grafting: Innocuous Attribution-Guided Mixup with Calibrated Label Mixi ng
Joonhyung Park,June Yong Yang,Jinwoo Shin,Sung Ju Hwang,Eunho Yang

The Mixup scheme of mixing a pair of samples to create an augmented training sample has gained much attention recently for better training of neural networks. A straightforward and widely used extension is to combine Mixup and regional dropout methods: removing random patches from a sample and replacing it with the features from another sample. Albeit their simplicity and effectiveness, these methods are prone to create harmful samples due to their randomness. In recent studies, attempts to prevent such a phenomenon by selecting only the most informative features are gradually emerging. However, this maximum saliency strategy acts against their fundamental duty of sample diversification as they always deterministically select regions with maximum saliency, injecting bias into the augmented data. To address this problem, we present Saliency Grafting, a novel Mixup-like data augmentation method that captures the best of both ways. By stochastically sampling the features and ‘grafting’ them onto another sample, our method effectively generates diverse yet meaningful samples. The second ingredient of Saliency Grafting is to produce the label of the grafted sample by mixing the labels in a saliency-calibrated fashion, which rectifies supervision misguidance introduced by the random sampling procedure. Our experiments under CIFAR and ImageNet datasets show that our scheme outperforms the current state-of-the-art augmentation strategies not only in terms of classification accuracy, but is also superior in coping under stress conditions such as data corruption and data scarcity. The code will be released.
****

Rethinking Compressed Convolution Neural Network from a Statistical Perspective
Feiqing Huang,Yuefeng Si,Guodong Li
Many designs have recently been proposed to improve the model efficiency of convolutional neural networks (CNNs) at a fixed resource budget, while there is a lack of theoretical analysis to justify them. This paper first formulates CNNs with high-order inputs into statistical models, which have a special "Tucker-like" formulation. This makes it possible to further conduct the sample complexity analysis to CNNs as well as compressed CNNs via tensor decomposition. Tucker and CP decompositions are commonly adopted to compress CNNs in the literature. The low rank assumption is usually imposed on the output channels, which according to our study, may not be beneficial to obtain a computationally efficient model while a similar accuracy can be maintained. Our finding is further supported by ablation studies on CIFAR10, SVNH and UCF101 datasets.
****

DQSGD: DYNAMIC QUANTIZED STOCHASTIC GRADIENT DESCENT FOR COMMUNICATION-EFFICIENT DISTRIBUTED LEARNING
Guangfeng Yan,Shao-Lun Huang,Tian Lan,Linqi Song
Gradient quantization is widely adopted to mitigate communication costs in distributed learning systems. Existing gradient quantization algorithms often rely on design heuristics and/or empirical evidence to tune the quantization strategy for different learning problems. To the best of our knowledge, there is no theoretical framework characterizing the trade-off between communication cost and model accuracy under dynamic gradient quantization strategies. This paper addresses this issue by proposing a novel dynamic quantized SGD (DQSGD) framework, which enables us to optimize the quantization strategy for each gradient descent step by exploring the trade-off between communication cost and modeling error. In particular, we derive an upper bound, tight in some cases, of the modeling error for arbitrary dynamic quantization strategy. By minimizing this upper bound, we obtain an enhanced quantization algorithm with significantly improved modeling error under given communication overhead constraints. Besides, we show that our quantization scheme achieves a strengthened communication cost and model accuracy trade-off in a wide range of optimization models. Finally, through extensive experiments on large-scale computer vision and natural language processing tasks on CIFAR-10, CIFAR-100, and AG-News datasets, respectively. we demonstrate that our quantization scheme significantly outperforms the state-of-the-art gradient quantization methods in terms of communication costs.
****

Secure Federated Learning of User Verification Models

Hossein Hosseini,Hyunsin Park,Sungrack Yun,Christos Louizos,Joseph Soriaga,Max W elling
We consider the problem of training User Verification (UV) models in federated s etup, where the conventional loss functions are not applicable due to the constr aints that each user has access to the data of only one class and user embedding s cannot be shared with the server or other users. To address this problem, we propose Federated User Verification (FedUV), a framework for private and secure training of UV models. In FedUV, users jointly learn a set of vectors and maximi ze the correlation of their instance embeddings with a secret user-defi ned linear combination of those vectors. We show that choosing the linear combi nations from the codewords of an error-correcting code allows users to collabora tively train the model without revealing their embedding vectors. We present th e experimental results for user verification with voice, face, and handwriting d ata and show that FedUV is on par with existing approaches, while not sharing th e embeddings with other users or the server.
****

Multi-scale Network Architecture Search for Object Detection
Yuxin Yue,Quanquan Li,Yujie Wang
Many commonly-used detection frameworks aim to handle the multi-scale object det ection problem. The input image is always encoded to multi-scale features and ob jects grouped by scale range are assigned to the corresponding features. However , the design of multi-scale feature production is quite hand-crafted or partiall y automatic. In this paper, we show that more possible architectures of encoder network and different strategies of feature utilization can lead to superior per formance. Specifically, we propose an efficient and effective multi-scale networ k architecture search method (MSNAS) to improve multi-scale object detection by jointly optimizing network stride search of the encoder and appropriate feature selection for detection heads. We demonstrate the effectiveness of the method on COCO dataset and obtain a remarkable performance gain with respect to the origi nal Feature Pyramid Networks.
****

Nonvacuous Loss Bounds with Fast Rates for Neural Networks via Conditional Infor mation Measures
Fredrik Hellström,Giuseppe Durisi
We present a framework to derive bounds on the test loss of randomized learning algorithms for the case of bounded loss functions. This framework leads to bound s that depend on the conditional information density between the the output hypo thesis and the choice of the training set, given a larger set of data samples fr om which the training set is formed. Furthermore, the bounds pertain to the aver age test loss as well as to its tail probability, both for the PAC-Bayesian and the single-draw settings. If the conditional information density is bounded unif ormly in the size n of the training set, our bounds decay as $1/n$, which is r eferred to as a fast rate. This is in contrast with the tail bounds involving co nditional information measures available in the literature, which have a less be nign $1/\sqrt{n}$ dependence. We demonstrate the usefulness of our tail bounds by showing that they lead to estimates of the test loss achievable with several neural network architectures trained on MNIST and Fashion-MNIST that match the s tate-of-the-art bounds available in the literature.
****

Non-Markovian Predictive Coding For Planning In Latent Space
Tung Nguyen,Rui Shu,Tuan Pham,Hung Bui,Stefano Ermon
High-dimensional observations are a major challenge in the application of model- based reinforcement learning (MBRL) to real-world environments. In order to hand le high-dimensional sensory inputs, existing MBRL approaches use representation learning to map high-dimensional observations into a lower-dimensional latent sp ace that is more amenable to dynamics estimation and planning. Crucially, the ta sk-relevance and predictability of the learned representations play critical rol es in the success of planning in latent space. In this work, we present Non-Mark ovian Predictive Coding (NMPC), an information-theoretic approach for planning f rom high-dimensional observations with two key properties: 1) it formulates a mu

tual information objective that prioritizes the encoding of task-relevant compon ents of the environment; and 2) it employs a recurrent neural network capable of modeling non-Markovian latent dynamics. To demonstrate NMPC’s ability to priori tize task-relevant information, we evaluate our new model on a challenging modif ication of standard DMControl tasks where the DMControl background is replaced w ith natural videos, containing complex but irrelevant information to the plannin g task. Our experiments show that NMPC is superior to existing methods in the ch allenging complex-background setting while remaining competitive with current st ate-of-the-art MBRL models in the standard setting.
****

Sparse matrix products for neural network compression
Luc Giffon,hachem kadri,Stephane Ayache,Ronan Sicre,thierry artieres
Over-parameterization of neural networks is a well known issue that comes along with their great performance. Among the many approaches proposed to tackle this problem, low-rank tensor decompositions are largely investigated to compress dee p neural networks. Such techniques rely on a low-rank assumption of the layer we ight tensors that does not always hold in practice. Following this observation, this paper studies sparsity inducing techniques to build new sparse matrix produ ct layer for high-rate neural networks compression. Specifically, we explore rec ent advances in sparse optimization to replace each layer's weight matrix, eithe r convolutional or fully connected, by a product of sparse matrices. Our experim ents validate that our approach provides a better compression-accuracy trade-off than most popular low-rank-based compression techniques.

****

SiamCAN:Simple yet Effective Method to enhance Siamese Short-Term Tracking
Yue Zhao,Zhibin Yu
Most traditional Siamese trackers are used to regard the location of the max res ponse map as the center of target. However, it is difficult for these traditiona l methods to calculate response value accurately when face the similar object, deformation, background clutters and other challenges. So how to get the reliabl e response map is the key to improve tracking performance. Accordingly, a simple yet effective short-term tracking framework (called SiamCAN),by which bridging the information flow between search branch and template branch, is proposed to s olve the above problem in this paper. Moreover, in order to get more accurate ta rget estimation, an anchor-free mechanism and specialized training strategy are applied to narrow the gap between the predicted bounding box and groundtruth. Th e proposed method achieves state-of-the-art performance on four visual tracking benchmarks including UAV123, OTB100, VOT2018 and VOT2019, outperforming the stro ng baseline, SiamBAN, by 0.327 $\displaystyle \rightarrow$ 0.331 on VOT2019 and 0.631 $\displaystyle \rightarrow$ 0.638 success score, 0.833 $\displaystyle \rig htarrow$ 0.850 precision score on UAV123.
****

ON NEURAL NETWORK GENERALIZATION VIA PROMOTING WITHIN-LAYER ACTIVATION DIVERSITY
Firas Laakom,Jenni Raitoharju,Alexandros Iosifidis,Moncef Gabbouj
During the last decade, neural networks have been intensively used to tackle var ious problems and they have often led to state-of-the-art results. These network s are composed of multiple jointly optimized layers arranged in a hierarchical s tructure. At each layer, the aim is to learn to extract hidden patterns needed t o solve the problem at hand and forward it to the next layers. In the standard f orm, a neural network is trained with gradient-based optimization, where the err ors are back-propagated from the last layer back to the first one. Thus at each optimization step, neurons at a given layer receive feedback from neurons belong ing to higher layers of the hierarchy. In this paper, we propose to complement t his traditional 'between-layer' feedback with additional 'within-layer' feedback to encourage diversity of the activations within the same layer. To this end, w e 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 neuron to learn a distinctive representation and, thus, to enrich the data representation learned within the layer 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 in a supervised context and we prove that increasing the diversity of hidden activations reduces the estimation error. In addition to the theoretical guarantees, we present an empirical study confirming that the proposed approach enhances the performance of neural networks.
****

Factored Action Spaces in Deep Reinforcement Learning
Thomas PIERROT,Valentin Macé,Jean-Baptiste Sevestre,Louis Monier,Alexandre Laterre,Nicolas Perrin,Karim Beguir,Olivier Sigaud
Very large action spaces constitute a critical challenge for deep Reinforcement Learning (RL) algorithms. An existing approach consists in splitting the action space into smaller components and choosing either independently or sequentially actions in each dimension. This approach led to astonishing results for the StarCraft and Dota 2 games, however it remains underexploited and understudied. In this paper, we name this approach Factored Actions Reinforcement Learning (FARL) and study both its theoretical impact and practical use. Notably, we provide a theoretical analysis of FARL on the Proximal Policy Optimization (PPO) and Soft Actor Critic (SAC) algorithms and evaluate these agents in different classes of problems. We show that FARL is a very versatile and efficient approach to combinatorial and continuous control problems.
****

A Near-Optimal Recipe for Debiasing Trained Machine Learning Models
Ibrahim Alabdulmohsin,Mario Lucic
We present an efficient and scalable algorithm for debiasing trained models, including deep neural networks (DNNs), which we prove to be near-optimal by bounding its excess Bayes risk. Unlike previous black-box reduction methods to cost-sensitive classification rules, the proposed algorithm operates on models that have been trained without having to retrain the model. Furthermore, as the algorithm is based on projected stochastic gradient descent (SGD), it is particularly attractive for deep learning applications. We empirically validate the proposed algorithm on standard benchmark datasets across both classical algorithms and modern DNN architectures and demonstrate that it outperforms previous post-processing approaches for unbiased classification.
****

Semi-supervised counterfactual explanations
SURYA SHRAVAN KUMAR SAJJA,Sumanta Mukherjee,Satyam Dwivedi,Vikas C. Raykar
Counterfactual explanations for machine learning models are used to find minimal interventions to the feature values such that the model changes the prediction to a different output or a target output. A valid counterfactual explanation should have likely feature values. Here, we address the challenge of generating counterfactual explanations that lie in the same data distribution as that of the training data and more importantly, they belong to the target class distribution. This requirement has been addressed through the incorporation of auto-encoder reconstruction loss in the counterfactual search process. Connecting the output behavior of the classifier to the latent space of the auto-encoder has further improved the speed of the counterfactual search process and the interpretability of the resulting counterfactual explanations. Continuing this line of research, we show further improvement in the interpretability of counterfactual explanations when the auto-encoder is trained in a semi-supervised fashion with class tagged input data. We empirically evaluate our approach on several datasets and show considerable improvement in-terms of several metrics.
****

Attacking Few-Shot Classifiers with Adversarial Support Sets
Elre Talea Oldewage,John F Bronskill,Richard E Turner
Few-shot learning systems, especially those based on meta-learning, have recently made significant advances, and are now being considered for real world problems in healthcare, personalization, and science. In this paper, we examine the robustness of such deployed few-shot learning systems when they are fed an imperceptibly perturbed few-shot dataset, showing that the resulting predictions on test

inputs can become worse than chance. This is achieved by developing a novel Adversarial Support Set Attack which crafts a poisoned set of examples. When even a small subset of malicious data points is inserted into the support set of a meta-learner, accuracy is significantly reduced. For example, the average classification accuracy of CNAPs on the Aircraft dataset in the META-DATASET benchmark drops from 69.2% to 9.1% when only 20% of the support set is poisoned by imperceptible perturbations. We evaluate the new attack on a variety of few-shot classification algorithms including MAML, prototypical networks, and CNAPs, on both small scale (miniImageNet) and large scale (META-DATASET) few-shot classification problems. Interestingly, adversarial support sets produced by attacking a meta-learning based few-shot classifier can also reduce the accuracy of a fine-tuning based few-shot classifier when both models use similar feature extractors.
****
Encoded Prior Sliced Wasserstein AutoEncoder for learning latent manifold representations
Sanjukta Krishnagopal,Jacob Bedrossian
While variational autoencoders have been successful in a variety of tasks, the use of conventional Gaussian or Gaussian mixture priors are limited in their ability to encode underlying structure of data in the latent representation.
In this work, we introduce an Encoded Prior Sliced Wasserstein AutoEncoder (EPSWAE) wherein an additional prior-encoder network facilitates learns an embedding of the data manifold which preserves topological and geometric properties of the data, thus improving the structure of latent space.
The autoencoder and prior-encoder networks are iteratively trained using the Sliced Wasserstein (SW) distance, which efficiently measures the distance between two \textit{arbitrary} sampleable distributions without being constrained to a specific form as in the KL divergence, and without requiring expensive adversarial training.
To improve the representation, we use (1) a structural consistency term in the loss that encourages isometry between feature space and latent space and (2) a nonlinear variant of the SW distance which averages over random nonlinear shearing.
The effectiveness of the learned manifold encoding is best explored by traversing the latent space through interpolations along \textit{geodesics} which generate samples that lie on the manifold and hence are advantageous compared to standard Euclidean interpolation.
To this end, we introduce a graph-based algorithm for interpolating along network-geodesics in latent space by maximizing the density of samples along the path while minimizing total energy. We use the 3D-spiral data to show that the prior does indeed encode the geometry underlying the data and to demonstrate the advantages of the network-algorithm for interpolation.
Additionally, we apply our framework to MNIST, and CelebA datasets, and show that outlier generations, latent representations, and geodesic interpolations are comparable to the state of the art.
****
Model-based micro-data reinforcement learning: what are the crucial model properties and which model to choose?
Balázs Kégl,Gabriel Hurtado,Albert Thomas
We contribute to micro-data model-based reinforcement learning (MBRL) by rigorously comparing popular generative models using a fixed (random shooting) control agent. We find that on an environment that requires multimodal posterior predictives, mixture density nets outperform all other models by a large margin. When multimodality is not required, our surprising finding is that we do not need probabilistic posterior predictives: deterministic models are on par, in fact they consistently (although non-significantly) outperform their probabilistic counterparts. We also found that heteroscedasticity at training time, perhaps acting as a regularizer, improves predictions at longer horizons. At the methodological side, we design metrics and an experimental protocol which can be used to evaluate the various models, predicting their asymptotic performance when using them on the control problem. Using this framework, we improve the state-of-the-art sampl

e complexity of MBRL on Acrobot by two to four folds, using an aggressive traini ng schedule which is outside of the hyperparameter interval usually considered.
****

Set Prediction without Imposing Structure as Conditional Density Estimation
David W Zhang,Gertjan J. Burghouts,Cees G. M. Snoek
Set prediction is about learning to predict a collection of unordered variables with unknown interrelations. Training such models with set losses imposes the st ructure of a metric space over sets. We focus on stochastic and underdefined cas es, where an incorrectly chosen loss function leads to implausible predictions. Example tasks include conditional point-cloud reconstruction and predicting futu re states of molecules. In this paper we propose an alternative to training via set losses, by viewing learning as conditional density estimation. Our learning framework fits deep energy-based models and approximates the intractable likelih ood with gradient-guided sampling. Furthermore, we propose a stochastically augm ented prediction algorithm that enables multiple predictions, reflecting the pos sible variations in the target set. We empirically demonstrate on a variety of d atasets the capability to learn multi-modal densities and produce different plau sible predictions. Our approach is competitive with previous set prediction mode ls on standard benchmarks. More importantly, it extends the family of addressabl e tasks beyond those that have unambiguous predictions.
****

Learning Value Functions in Deep Policy Gradients using Residual Variance
Yannis Flet-Berliac,reda ouhamma,odalric-ambrym maillard,Philippe Preux
Policy gradient algorithms have proven to be successful in diverse decision maki ng and control tasks. However, these methods suffer from high sample complexity and instability issues. In this paper, we address these challenges by providing a different approach for training the critic in the actor-critic framework. Our work builds on recent studies indicating that traditional actor-critic algorithm s do not succeed in fitting the true value function, calling for the need to ide ntify a better objective for the critic. In our method, the critic uses a new st ate-value (resp. state-action-value) function approximation that learns the valu e of the states (resp. state-action pairs) relative to their mean value rather t han the absolute value as in conventional actor-critic. We prove the theoretical consistency of the new gradient estimator and observe dramatic empirical improv ement across a variety of continuous control tasks and algorithms. Furthermore, we validate our method in tasks with sparse rewards, where we provide experiment al evidence and theoretical insights.
****

IDF++: Analyzing and Improving Integer Discrete Flows for Lossless Compression
Rianne van den Berg,Alexey A. Gritsenko,Mostafa Dehghani,Casper Kaae Sønderby,Ti m Salimans
In this paper we analyse and improve integer discrete flows for lossless compres sion. Integer discrete flows are a recently proposed class of models that learn invertible transformations for integer-valued random variables. Their discrete n ature makes them particularly suitable for lossless compression with entropy cod ing schemes. We start by investigating a recent theoretical claim that states th at invertible flows for discrete random variables are less flexible than their c ontinuous counterparts. We demonstrate with a proof that this claim does not hol d for integer discrete flows due to the embedding of data with finite support in to the countably infinite integer lattice. Furthermore, we zoom in on the effect of gradient bias due to the straight-through estimator in integer discrete flow s, and demonstrate that its influence is highly dependent on architecture choice s and less prominent than previously thought. Finally, we show how different arc hitecture modifications improve the performance of this model class for lossless compression, and that they also enable more efficient compression: a model with half the number of flow layers performs on par with or better than the original integer discrete flow model.
****

Combining Imitation and Reinforcement Learning with Free Energy Principle
Ryoya Ogishima,Izumi Karino,Yasuo Kuniyoshi

Imitation Learning (IL) and Reinforcement Learning (RL) from high dimensional se
nsory inputs are often introduced as separate problems, but a more realistic pro
blem setting is how to merge the techniques so that the agent can reduce explora
tion costs by partially imitating experts at the same time it maximizes its retu
rn. Even when the experts are suboptimal (e.g. Experts learned halfway with othe
r RL methods or human-crafted experts), it is expected that the agent outperform
s the suboptimal experts' performance. In this paper, we propose to address the
issue by using and theoretically extending Free Energy Principle, a unified brai
n theory that explains perception, action and model learning in a Bayesian proba
bilistic way. We find that both IL and RL can be achieved based on the same free
 energy objective function. Our results show that our approach is promising in v
isual control tasks especially with sparse-reward environments.
****

Addressing Some Limitations of Transformers with Feedback Memory
Angela Fan,Thibaut Lavril,Edouard Grave,Armand Joulin,Sainbayar Sukhbaatar
Transformers have been successfully applied to sequential tasks despite being fe
edforward networks. Unlike recurrent neural networks, Transformers use attention
 to capture temporal relations while processing input tokens in parallel. While
this parallelization makes them computationally efficient, it restricts the mod
el from fully exploiting the sequential nature of the input. The representation
at a given layer can only access representations from lower layers, rather than
the higher level representations already available. In this work, we propose the
 Feedback Transformer architecture that exposes all previous representations to
all future representations, meaning the lowest representation of the current tim
estep is formed from the highest-level abstract representation of the past. We d
emonstrate on a variety of benchmarks in language modeling, machine translation,
 and reinforcement learning that the increased representation capacity can creat
e small, shallow models with much stronger performance than comparable Transform
ers.
****

Black-Box Optimization Revisited: Improving Algorithm Selection Wizards through
Massive Benchmarking
Laurent Meunier,Herilalaina Rakotoarison,Jeremy Rapin,Paco Wong,Baptiste Roziere
,Olivier Teytaud,Antoine Moreau,Carola Doerr
Existing studies in black-box optimization for machine learning suffer from low
generalizability, caused by a typically selective choice of problem instances us
ed
for training and testing different optimization algorithms. Among other issues,
this practice promotes overfitting and poor-performing user guidelines. To addre
ss
this shortcoming, we propose in this work a benchmark suite, OptimSuite,
which covers a broad range of black-box optimization problems, ranging from
academic benchmarks to real-world applications, from discrete over numerical
to mixed-integer problems, from small to very large-scale problems, from noisy
over dynamic to static problems, etc. We demonstrate the advantages of such a
broad collection by deriving from it Automated Black Box Optimizer (ABBO), a
general-purpose algorithm selection wizard. Using three different types of algor
ithm
selection techniques, ABBO achieves competitive performance on all
benchmark suites. It significantly outperforms previous state of the art on some
 of
them, including YABBOB and LSGO. ABBO relies on many high-quality base
components. Its excellent performance is obtained without any task-specific
parametrization. The benchmark collection, the ABBO wizard, its base solvers,
as well as all experimental data are reproducible and open source in OptimSuite.
****

On the Importance of Distraction-Robust Representations for Robot Learning
Andy Wang,Antoine Cully
Representation Learning methods can allow the application of Reinforcement Learn
ing algorithms when a high dimensionality in a robot's perceptions would otherwi

se prove prohibitive. Consequently, unsupervised Representation Learning compone nts often feature in robot control algorithms that assume high-dimensional camer a images as the principal source of information.
In their design and performance, these algorithms often benefit from the control led nature of the simulation or laboratory conditions they are evaluated in. How ever, these settings fail to acknowledge the stochasticity of most real-world en vironments.
In this work, we introduce the concept of Distraction-Robust Representation Lear ning. We argue that environment noise and other distractions require learned rep resentations to encode the robot's expected perceptions rather than the observed ones. Our experimental evaluations demonstrate that representations learned wit h a traditional dimensionality reduction algorithm are strongly susceptible to d istractions in a robot's environment.
We propose an Encoder-Decoder architecture that produces representations that al low the learning outcomes of robot control tasks to remain unaffected by these d istractions.
****

Fully Unsupervised Diversity Denoising with Convolutional Variational Autoencode rs
Mangal Prakash,Alexander Krull,Florian Jug
Deep Learning based methods have emerged as the indisputable leaders for virtual ly all image restoration tasks. Especially in the domain of microscopy images, v arious content-aware image restoration (CARE) approaches are now used to improve the interpretability of acquired data. Naturally, there are limitations to what can be restored in corrupted images, and like for all inverse problems, many po tential solutions exist, and one of them must be chosen. Here, we propose DivNoi sing, a denoising approach based on fully convolutional variational autoencoders (VAEs), overcoming the problem of having to choose a single solution by predict ing a whole distribution of denoised images. First we introduce a principled way of formulating the unsupervised denoising problem within the VAE framework by e xplicitly incorporating imaging noise models into the decoder. Our approach is f ully unsupervised, only requiring noisy images and a suitable description of the imaging noise distribution. We show that such a noise model can either be measu red, bootstrapped from noisy data, or co-learned during training. If desired, co nsensus predictions can be inferred from a set of DivNoising predictions, leadin g to competitive results with other unsupervised methods and, on occasion, even with the supervised state-of-the-art. DivNoising samples from the posterior enab le a plethora of useful applications. We are (i) showing denoising results for 1 3 datasets, (ii) discussing how optical character recognition (OCR) applications can benefit from diverse predictions, and are (iii) demonstrating how instance cell segmentation improves when using diverse DivNoising predictions.
****

Zero-shot Transfer Learning for Gray-box Hyper-parameter Optimization
Hadi Samer Jomaa,Lars Schmidt-Thieme,Josif Grabocka
Zero-shot hyper-parameter optimization refers to the process of selecting hyper- parameter configurations that are expected to perform well for a given dataset upfront, without access to any observations of the losses of the target response . Existing zero-shot approaches are posed as initialization strategies for Bayes ian Optimization and they often rely on engineered meta-features to measure data set similarity, operating under the assumption that the responses of similar dat asets behaves similarly with respect to the same hyper-parameters. Solutions for zero- shot HPO are embarrassingly parallelizable and thus can reduce vastly the required wallclock time of learning a single model. We propose a very simple HP O model called Gray-box Zero(0)-Shot Initialization (GROSI) as a conditional par ametric surrogate that learns a universal response model by exploiting the relat ionship between the hyper-parameters and the dataset meta-features directly. In contrast to existing HPO solutions, we achieve transfer of knowledge without eng ineered meta- features, but rather through a shared model that is trained simult aneously across all datasets. We design and optimize a novel loss function that allows us to regress from the dataset/hyper-parameter pair unto the response. Ex

periments on 120 datasets demonstrate the strong performance of GROSI, compared to conventional initialization strategies. We also show that by fine-tuning GROSI to the target dataset, we can outperform state-of-the-art sequential HPO algorithms.

****

Not All Memories are Created Equal: Learning to Expire

Sainbayar Sukhbaatar,Da JU,Spencer Poff,Stephen Roller,Arthur Szlam,Jason E Weston,Angela Fan

Attention mechanisms have shown promising results in sequence modeling tasks that require long-term memory. Recent work has investigated mechanisms to reduce the computational cost of preserving and storing the memories. However, not all content in the past is equally important to remember. We propose Expire-Span, a method that learns to retain the most important information and
expire the irrelevant information. This enables Transformers to scale to attend to tens of thousands of previous timesteps efficiently, as not all hidden states from previous timesteps are preserved. We demonstrate that Expire-Span can help models identify and retain critical information and show it can achieve state of the art results on long-context language modeling, reinforcement learning, and algorithmic tasks. Finally, we show that Expire-Span can scale to memories that are tens of thousands in size, which is helpful on incredibly long context tasks such as character-level PG-19 and a frame-by-frame moving objects task.

****

Ordering-Based Causal Discovery with Reinforcement Learning

Xiaoqiang Wang,Yali Du,Shengyu Zhu,Liangjun Ke,Zhitang Chen,Jianye HAO,Jun Wang

It is a long-standing question to discover causal relations among a set of variables in many empirical sciences. Recently, Reinforcement Learning (RL) has achieved promising results in causal discovery. However, searching the space of directed graphs directly and enforcing acyclicity by implicit penalties tend to be inefficient and restrict the method to the small problems. In this work, we alternatively consider searching an ordering by RL from the variable ordering space that is much smaller than that of directed graphs, which also helps avoid dealing with acyclicity. Specifically, we formulate the ordering search problem as a Markov decision process, and then use different reward designs to optimize the ordering generating model. A generated ordering is then processed using variable selection methods to obtain the final directed acyclic graph. In contrast to other causal discovery methods, our method can also utilize a pretrained model to accelerate training. We conduct experiments on both synthetic and real-world datasets, and show that the proposed method outperforms other baselines on important metrics even on large graph tasks.

****

Is Attention Better Than Matrix Decomposition?

Zhengyang Geng,Meng-Hao Guo,Hongxu Chen,Xia Li,Ke Wei,Zhouchen Lin

As an essential ingredient of modern deep learning, attention mechanism, especially self-attention, plays a vital role in the global correlation discovery. However, is hand-crafted attention irreplaceable when modeling the global context? Our intriguing finding is that self-attention is not better than the matrix decomposition~(MD) model developed 20 years ago regarding the performance and computational cost for encoding the long-distance dependencies. We model the global context issue as a low-rank completion problem and show that its optimization algorithms can help design global information blocks. This paper then proposes a series of Hamburgers, in which we employ the optimization algorithms for solving MDs to factorize the input representations into sub-matrices and reconstruct a low-rank embedding. Hamburgers with different MDs can perform favorably against the popular global context module self-attention when carefully coping with gradients back-propagated through MDs. Comprehensive experiments are conducted in the vision tasks where it is crucial to learn the global context, including semantic segmentation and image generation, demonstrating significant improvements over self-attention and its variants. Code is available at https://github.com/Gsunshine/Enjoy-Hamburger.

****
Deep Quotient Manifold Modeling
Jiseob Kim,Seungjae Jung,Hyundo Lee,Byoung-Tak Zhang
One of the difficulties in modeling real-world data is their complex multi-manif old structure due to discrete features. In this paper, we propose quotient manif old modeling (QMM), a new data-modeling scheme that considers generic manifold s tructure independent of discrete features, thereby deriving efficiency in modeli ng and allowing generalization over untrained manifolds. QMM considers a deep en coder inducing an equivalence between manifolds; but we show it is sufficient to consider it only implicitly via a bias-regularizer we derive. This makes QMM ea sily applicable to existing models such as GANs and VAEs, and experiments show t hat these models not only present superior FID scores but also make good general izations across different datasets. In particular, we demonstrate an MNIST model that synthesizes EMNIST alphabets.
****
Learning Aggregation Functions
Giovanni Pellegrini,Alessandro Tibo,Paolo Frasconi,Andrea Passerini,Manfred Jaeg er
Learning on sets is increasingly gaining attention in the machine learning comm unity, due to its widespread applicability. Typically, representations over sets are computed by using fixed aggregation functions such as sum or maximum. Howev er, recent results showed that universal function representation by sum- (or max -) decomposition requires either highly discontinuous (and thus poorly learnable) mappings, or a latent dimension equal to the maximum number of elements in the set. To mitigate this problem, we introduce LAF (Learning Aggregation Functions), a learnable aggregator for sets of arbitrary cardinality. LAF can approximate several extensively used aggregators (such as average, sum, maximum) as well a s more complex functions (e.g. variance and skewness). We report experiments on semi-synthetic and real data showing that LAF outperforms state-of-the-art sum- (max-) decomposition architectures such as DeepSets and library-based architect ures like Principal Neighborhood Aggregation.
****
Improving Transformation Invariance in Contrastive Representation Learning
Adam Foster,Rattana Pukdee,Tom Rainforth
We propose methods to strengthen the invariance properties of representations ob tained by contrastive learning. While existing approaches implicitly induce a de gree of invariance as representations are learned, we look to more directly enfo rce invariance in the encoding process. To this end, we first introduce a traini ng objective for contrastive learning that uses a novel regularizer to control h ow the representation changes under transformation. We show that representations trained with this objective perform better on downstream tasks and are more rob ust to the introduction of nuisance transformations at test time. Second, we pro pose a change to how test time representations are generated by introducing a fe ature averaging approach that combines encodings from multiple transformations o f the original input, finding that this leads to across the board performance ga ins. Finally, we introduce the novel Spirograph dataset to explore our ideas in the context of a differentiable generative process with multiple downstream task s, showing that our techniques for learning invariance are highly beneficial.
****
Don't be picky, all students in the right family can learn from good teachers
Roy Henha Eyono,Fabio Maria Carlucci,Pedro M Esperança,Binxin Ru,Philip Torr
State-of-the-art results in deep learning have been improving steadily, in good part due to the use of larger models. However, widespread use is constrained by device hardware limitations, resulting in a substantial performance gap between state-of-the-art models and those that can be effectively deployed on small devi ces.

While Knowledge Distillation (KD) theoretically enables small student models to emulate larger teacher models, in practice selecting a good student architecture requires considerable human expertise. Neural Architecture Search (NAS) appears

as a natural solution to this problem but most approaches can be inefficient, as most of the computation is spent comparing architectures sampled from the same distribution, with negligible differences in performance.

In this paper, we propose to instead search for a family of student architectures sharing the property of being good at learning from a given teacher.
Our approach AutoKD, powered by Bayesian Optimization, explores a flexible graph-based search space, enabling us to automatically learn the optimal student architecture distribution and KD parameters, while being 20x more sample efficient compared to existing state-of-the-art. We evaluate our method on 3 datasets; on large images specifically, we reach the teacher performance while using 3x less memory and 10x less parameters. Finally, while AutoKD uses the traditional KD loss, it outperforms more advanced KD variants using hand-designed students.
****
Sparse Uncertainty Representation in Deep Learning with Inducing Weights
Hippolyt Ritter,Martin Kukla,Cheng Zhang,Yingzhen Li
Bayesian neural networks and deep ensembles represent two modern paradigms of uncertainty quantification in deep learning. Yet these approaches struggle to scale mainly due to memory inefficiency issues, since they require parameter storage several times higher than their deterministic counterparts. To address this, we augment the weight matrix of each layer with a small number of inducing weights, thereby projecting the uncertainty quantification into such low dimensional spaces. We further extend Matheron's conditional Gaussian sampling rule to enable fast weight sampling, whichenable our inference method to maintain reasonable run-time as compared with ensembles. Importantly, our approach achieves competitive performance to the state-of-the-art in prediction and uncertainty estimation tasks with fully connected neural networks and ResNets, while reducing the parameter size to $\leq 47.9\%$ of that of a single neural network.
****
Explicit Pareto Front Optimization for Constrained Reinforcement Learning
Sandy Huang,Abbas Abdolmaleki,Philemon Brakel,Steven Bohez,Nicolas Heess,Martin Riedmiller,raia hadsell
Many real-world problems require that reinforcement learning (RL) agents learn policies that not only maximize a scalar reward, but do so while meeting constraints, such as remaining below an energy consumption threshold. Typical approaches for solving constrained RL problems rely on Lagrangian relaxation, but these suffer from several limitations. We draw a connection between multi-objective RL and constrained RL, based on the key insight that the constraint-satisfying optimal policy must be Pareto optimal. This leads to a novel, multi-objective perspective for constrained RL. We propose a framework that uses a multi-objective RL algorithm to find a Pareto front of policies that trades off between the reward and constraint(s), and simultaneously searches along this front for constraint-satisfying policies. We show that in practice, an instantiation of our framework outperforms existing approaches on several challenging continuous control domains, both in terms of solution quality and sample efficiency, and enables flexibility in recovering a portion of the Pareto front rather than a single constraint-satisfying policy.
****
On the Origin of Implicit Regularization in Stochastic Gradient Descent
Samuel L Smith,Benoit Dherin,David Barrett,Soham De
For infinitesimal learning rates, stochastic gradient descent (SGD) follows the path of gradient flow on the full batch loss function. However moderately large learning rates can achieve higher test accuracies, and this generalization benefit is not explained by convergence bounds, since the learning rate which maximizes test accuracy is often larger than the learning rate which minimizes training loss. To interpret this phenomenon we prove that for SGD with random shuffling, the mean SGD iterate also stays close to the path of gradient flow if the learning rate is small and finite, but on a modified loss. This modified loss is composed of the original loss function and an implicit regularizer, which penalizes the norms of the minibatch gradients. Under mild assumptions, when the batch siz

e is small the scale of the implicit regularization term is proportional to the ratio of the learning rate to the batch size. We verify empirically that explicitly including the implicit regularizer in the loss can enhance the test accuracy when the learning rate is small.
****
AC-VAE: Learning Semantic Representation with VAE for Adaptive Clustering
Xingyu Xie,Minjuan Zhu,Yan Wang,Lei Zhang
Unsupervised representation learning is essential in the field of machine learning, and accurate neighbor clusters of representation show great potential to support unsupervised image classification. This paper proposes a VAE (Variational Autoencoder) based network and a clustering method to achieve adaptive neighbor clustering to support the self-supervised classification. The proposed network encodes the image into the representation with boundary information, and the proposed cluster method takes advantage of the boundary information to deliver adaptive neighbor cluster results. Experimental evaluations show that the proposed method outperforms state-of-the-art representation learning methods in terms of neighbor clustering accuracy. Particularly, AC-VAE achieves 95\% and 82\% accuracy on CIFAR10 dataset when the average neighbor cluster sizes are 10 and 100. Furthermore, the neighbor cluster results are found converge within the clustering range ($\alpha\leq2$), and the converged neighbor clusters are used to support the self-supervised classification. The proposed method delivers classification results that are competitive with the state-of-the-art and reduces the super parameter k in KNN (K-nearest neighbor), which is often used in self-supervised classification.
****
On the use of linguistic similarities to improve Neural Machine Translation for African Languages
Tikeng Notsawo Pascal,NANDA ASSOBJIO Brice Yvan,James Assiene
In recent years, there has been a resurgence in research on empirical methods for machine translation. Most of this research has been focused on high-resource, European languages. Despite the fact that around 30% of all languages spoken worldwide are African, the latter have been heavily under investigated and this, partly due to the lack of public parallel corpora online. Furthermore, despite their large number (more than 2,000) and the similarities between them, there is currently no publicly available study on how to use this multilingualism (and associated similarities) to improve machine translation systems performance on African languages. So as to address these issues:
We propose a new dataset for African languages that provides parallel data for vernaculars not present in commonly used dataset like JW300 [1]. To exploit multilingualism, we first use a historical approach based on historical origins of these languages, their morphologies, their geographical and cultural distributions as well as migrations of population to identify similar vernaculars.
We also propose a new metric to automatically evaluate similarities between languages. This new metric does not require word level parallelism like traditional methods but only paragraph level parallelism.
We then show that performing Masked Language Modelling and Translation Language Modeling in addition to multi-task learning on a cluster of similar languages leads to a strong boost of performance in translating individual pairs inside this cluster.
In particular, we record an improvement of 29 BLEU on the pair Bafia-Ewondo using our approaches compared to previous work methods that did not exploit multilingualism in any way.

[1] http://opus.nlpl.eu/JW300.php
****
BAFFLE: TOWARDS RESOLVING FEDERATED LEARNING’S DILEMMA - THWARTING BACKDOOR AND INFERENCE ATTACKS
Thien Duc Nguyen,Phillip Rieger,Hossein Yalame,Helen Möllering,Hossein Fereidooni,Samuel Marchal,Markus Miettinen,Azalia Mirhoseini,Ahmad-Reza Sadeghi,Thomas Schneider,Shaza Zeitouni

Recently, federated learning (FL) has been subject to both security and privacy attacks posing a dilemmatic challenge on the underlying algorithmic designs: On the one hand, FL is shown to be vulnerable to backdoor attacks that stealthily manipulate the global model output using malicious model updates, and on the other hand, FL is shown vulnerable to inference attacks by a malicious aggregator inferring information about clients' data from their model updates. Unfortunately, existing defenses against these attacks are insufficient and mitigating both attacks at the same time is highly challenging, because while defeating backdoor attacks requires the analysis of model updates, protection against inference attacks prohibits access to the model updates to avoid information leakage. In this work, we introduce BAFFLE, a novel in-depth defense for FL that tackles this challenge. To mitigate backdoor attacks, it applies a multilayered defense by using a Model Filtering layer to detect and reject malicious model updates and a Poison Elimination layer to eliminate any effect of a remaining undetected weak manipulation. To impede inference attacks, we build private BAFFLE that securely evaluates the BAFFLE algorithm under encryption using sophisticated secure computation techniques. We extensively evaluate BAFFLE against state-of-the-art backdoor attacks on several datasets and applications, including image classification, word prediction, and IoT intrusion. We show that BAFFLE can entirely remove backdoors with a negligible effect on accuracy and that private BAFFLE is practical.
****

Share or Not? Learning to Schedule Language-Specific Capacity for Multilingual Translation

Biao Zhang,Ankur Bapna,Rico Sennrich,Orhan Firat

Using a mix of shared and language-specific (LS) parameters has shown promise in multilingual neural machine translation (MNMT), but the question of when and where LS capacity matters most is still under-studied. We offer such a study by proposing conditional language-specific routing (CLSR). CLSR employs hard binary gates conditioned on token representations to dynamically select LS or shared paths. By manipulating these gates, it can schedule LS capacity across sub-layers in MNMT subject to the guidance of translation signals and budget constraints. Moreover, CLSR can easily scale up to massively multilingual settings. Experiments with Transformer on OPUS-100 and WMT datasets show that: 1) MNMT is sensitive to both the amount and the position of LS modeling: distributing 10%-30% LS computation to the top and/or bottom encoder/decoder layers delivers the best performance; and 2) one-to-many translation benefits more from CLSR compared to many-to-one translation, particularly with unbalanced training data. Our study further verifies the trade-off between the shared capacity and LS capacity for multilingual translation. We corroborate our analysis by confirming the soundness of our findings as foundation of our improved multilingual Transformers. Source code and models are available at https://github.com/bzhangGo/zero/tree/iclr2021_clsr.
****

Robust Constrained Reinforcement Learning for Continuous Control with Model Misspecification

Daniel J Mankowitz,Dan Andrei Calian,Rae Jeong,Cosmin Paduraru,Nicolas Heess,Sumanth Dathathri,Martin Riedmiller,Timothy Mann

Many real-world physical control systems are required to satisfy constraints upon deployment. Furthermore, real-world systems are often subject to effects such as non-stationarity, wear-and-tear, uncalibrated sensors and so on. Such effects effectively perturb the system dynamics and can cause a policy trained successfully in one domain to perform poorly when deployed to a perturbed version of the same domain. This can affect a policy's ability to maximize future rewards as well as the extent to which it satisfies constraints. We refer to this as constrained model misspecification. We present an algorithm with theoretical guarantees that mitigates this form of misspecification, and showcase its performance in multiple Mujoco tasks from the Real World Reinforcement Learning (RWRL) suite.
****

Transient Non-stationarity and Generalisation in Deep Reinforcement Learning

Maximilian Igl,Gregory Farquhar,Jelena Luketina,Wendelin Boehmer,Shimon Whiteson

Non-stationarity can arise in Reinforcement Learning (RL) even in stationary env

ironments. For example, most RL algorithms collect new data throughout training, using a non-stationary behaviour policy. Due to the transience of this non-stationarity, it is often not explicitly addressed in deep RL and a single neural network is continually updated. However, we find evidence that neural networks exhibit a memory effect, where these transient non-stationarities can permanently impact the latent representation and adversely affect generalisation performance. Consequently, to improve generalisation of deep RL agents, we propose Iterated Relearning (ITER). ITER augments standard RL training by repeated knowledge transfer of the current policy into a freshly initialised network, which thereby experiences less non-stationarity during training. Experimentally, we show that ITER improves performance on the challenging generalisation benchmarks ProcGen and Multiroom.

****

Gradient descent temporal difference-difference learning

Rong Zhu,James Murray

Off-policy algorithms, in which a behavior policy differs from the target policy and is used to gain experience for learning, have proven to be of great practical value in reinforcement learning. However, even for simple convex problems such as linear value function approximation, these algorithms are not guaranteed to be stable. To address this, alternative algorithms that are provably convergent in such cases have been introduced, the most well known being gradient descent temporal difference (GTD) learning. This algorithm and others like it, however, tend to converge much more slowly than conventional temporal difference learning.

In this paper we propose gradient descent temporal difference-difference (Gradient-DD) learning in order to accelerate GTD learning by introducing second-order differences in successive parameter updates.

We investigate this algorithm in the framework of linear value function approximation and analytically showing its improvement over GTD learning. Studying the model empirically on the random walk and Boyan-chain prediction tasks, we find substantial improvement over GTD learning and, in several cases, better performance even than conventional TD learning.

****

Lossless Compression of Structured Convolutional Models via Lifting

Gustav Sourek,Filip Zelezny,Ondrej Kuzelka

Lifting is an efficient technique to scale up graphical models generalized to relational domains by exploiting the underlying symmetries. Concurrently, neural models are continuously expanding from grid-like tensor data into structured representations, such as various attributed graphs and relational databases. To address the irregular structure of the data, the models typically extrapolate on the idea of convolution, effectively introducing parameter sharing in their, dynamically unfolded, computation graphs. The computation graphs themselves then reflect the symmetries of the underlying data, similarly to the lifted graphical models. Inspired by lifting, we introduce a simple and efficient technique to detect the symmetries and compress the neural models without loss of any information. We demonstrate through experiments that such compression can lead to significant speedups of structured convolutional models, such as various Graph Neural Networks, across various tasks, such as molecule classification and knowledge-base completion.

****

Distribution-Based Invariant Deep Networks for Learning Meta-Features

Gwendoline de Bie,Herilalaina Rakotoarison,Gabriel Peyré,Michèle Sebag

Recent advances in deep learning from probability distributions successfully achieve classification or regression from distribution samples, thus invariant under permutation of the samples. The first contribution of the paper is to extend these neural architectures to achieve invariance under permutation of the features, too. The proposed architecture, called Dida, inherits the NN properties of universal approximation, and its robustness with respect to Lipschitz-bounded transformations of the input distribution is established. The second contribution i

s to empirically and comparatively demonstrate the merits of the approach on two tasks defined at the dataset level. On both tasks, Dida learns meta-features supporting the characterization of a (labelled) dataset. The first task consists of predicting whether two dataset patches are extracted from the same initial dataset. The second task consists of predicting whether the learning performance achieved by a hyper-parameter configuration under a fixed algorithm (ranging in k-NN, SVM, logistic regression and linear SGD) dominates that of another configuration, for a dataset extracted from the OpenML benchmarking suite. On both tasks, Dida outperforms the state of the art: DSS and Dataset2Vec architectures, as well as the models based on the hand-crafted meta-features of the literature.
****

HyperSAGE: Generalizing Inductive Representation Learning on Hypergraphs
Devanshu Arya,Deepak Gupta,Stevan Rudinac,Marcel Worring
Graphs are the most ubiquitous form of structured data representation used in machine learning. They model, however, only pairwise relations between nodes and are not designed for encoding the higher-order relations found in many real-world datasets. To model such complex relations, hypergraphs have proven to be a natural representation. Learning the node representations in a hypergraph is more complex than in a graph as it involves information propagation at two levels: within every hyperedge and across the hyperedges. Most current approaches first transform a hypergraph structure to a graph for use in existing geometric deep learning algorithms. This transformation leads to information loss, and sub-optimal exploitation of the hypergraph's expressive power. We present HyperSAGE, a novel hypergraph learning framework that uses a two-level neural message passing strategy to accurately and efficiently propagate information through hypergraphs. The flexible design of HyperSAGE facilitates different ways of aggregating neighborhood information. Unlike the majority of related work which is transductive, our approach, inspired by the popular GraphSAGE method, is inductive. Thus, it can also be used on previously unseen nodes, facilitating deployment in problems such as evolving or partially observed hypergraphs. Through extensive experimentation, we show that HyperSAGE outperforms state-of-the-art hypergraph learning methods on representative benchmark datasets. We also demonstrate that the higher expressive power of HyperSAGE makes it more stable in learning node representations as compared to the alternatives.
****

A Unifying Perspective on Neighbor Embeddings along the Attraction-Repulsion Spectrum
Niklas Böhm,Philipp Berens,Dmitry Kobak
Neighbor embeddings are a family of methods for visualizing complex high-dimensional datasets using kNN graphs. To find the low-dimensional embedding, these algorithms combine an attractive force between neighboring pairs of points with a repulsive force between all points. One of the most popular examples of such algorithms is t-SNE. Here we empirically show that changing the balance between the attractive and the repulsive forces in t-SNE yields a spectrum of embeddings, which is characterized by a simple trade-off: stronger attraction can better represent continuous manifold structures, while stronger repulsion can better represent discrete cluster structures. We find that UMAP embeddings correspond to t-SNE with increased attraction; mathematical analysis shows that this is because the negative sampling optimisation strategy employed by UMAP strongly lowers the effective repulsion. Likewise, ForceAtlas2, commonly used for visualizing developmental single-cell transcriptomic data, yields embeddings corresponding to t-SNE with the attraction increased even more. At the extreme of this spectrum lies Laplacian Eigenmaps, corresponding to zero repulsion. Our results demonstrate that many prominent neighbor embedding algorithms can be placed onto this attraction-repulsion spectrum, and highlight the inherent trade-offs between them.
****

Analyzing the Expressive Power of Graph Neural Networks in a Spectral Perspective
Muhammet Balcilar,Guillaume Renton,Pierre Héroux,Benoit Gaüzère,Sébastien Adam,Paul Honeine

In the recent literature of Graph Neural Networks (GNN), the expressive power of models has been studied through their capability to distinguish if two given graphs are isomorphic or not. Since the graph isomorphism problem is NP-intermediate, and Weisfeiler-Lehman (WL) test can give sufficient but not enough evidence in polynomial time, the theoretical power of GNNs is usually evaluated by the equivalence of WL-test order, followed by an empirical analysis of the models on some reference inductive and transductive datasets. However, such analysis does not account the signal processing pipeline, whose capability is generally evaluated in the spectral domain. In this paper, we argue that a spectral analysis of GNNs behavior can provide a complementary point of view to go one step further in the understanding of GNNs. By bridging the gap between the spectral and spatial design of graph convolutions, we theoretically demonstrate some equivalence of the graph convolution process regardless it is designed in the spatial or the spectral domain. Using this connection, we managed to re-formulate most of the state-of-the-art graph neural networks into one common framework. This general framework allows to lead a spectral analysis of the most popular GNNs, explaining their performance and showing their limits according to spectral point of view. Our theoretical spectral analysis is confirmed by experiments on various graph databases. Furthermore, we demonstrate the necessity of high and/or band-pass filters on a graph dataset, while the majority of GNN is limited to only low-pass and inevitably it fails.
****

Universal Value Density Estimation for Imitation Learning and Goal-Conditioned Reinforcement Learning
Yannick Schroecker,Charles Lee Isbell
This work considers two distinct settings: imitation learning and goal-conditioned reinforcement learning. In either case, effective solutions require the agent to reliably reach a specified state (a goal), or set of states (a demonstration). Drawing a connection between probabilistic long-term dynamics and the desired value function, this work introduces an approach that utilizes recent advances in density estimation to effectively learn to reach a given state. We develop a unified view on the two settings and show that the approach can be applied to both. In goal-conditioned reinforcement learning, we show it to circumvent the problem of sparse rewards while addressing hindsight bias in stochastic domains. In imitation learning, we show that the approach can learn from extremely sparse amounts of expert data and achieves state-of-the-art results on a common benchmark.
****

Learning and Generalization in Univariate Overparameterized Normalizing Flows
Kulin Shah,Amit Deshpande,Navin Goyal
In supervised learning, it is known that overparameterized neural networks with one hidden layer provably and efficiently learn and generalize, when trained using Stochastic Gradient Descent (SGD). In contrast, the benefit of overparameterization in unsupervised learning is not well understood. Normalizing flows (NFs) learn to map complex real-world distributions into simple base distributions, and constitute an important class of models in unsupervised learning for sampling and density estimation. In this paper, we theoretically and empirically analyze these models when the underlying neural network is one hidden layer overparametrized network. On the one hand we provide evidence that for a class of NFs, overparametrization hurts training. On the other, we prove that another class of NFs, with similar underlying networks can efficiently learn any reasonable data distribution under minimal assumptions. We extend theoretical ideas on learning and generalization from overparameterized neural networks in supervised learning to overparameterized normalizing flows in unsupervised learning. We also provide experimental validation to support our theoretical analysis in practice.
****

End-to-end Adversarial Text-to-Speech
Jeff Donahue,Sander Dieleman,Mikolaj Binkowski,Erich Elsen,Karen Simonyan
Modern text-to-speech synthesis pipelines typically involve multiple processing stages, each of which is designed or learnt independently from the rest. In this

work, we take on the challenging task of learning to synthesise speech from normalised text or phonemes in an end-to-end manner, resulting in models which operate directly on character or phoneme input sequences and produce raw speech audio outputs. Our proposed generator is feed-forward and thus efficient for both training and inference, using a differentiable alignment scheme based on token length prediction. It learns to produce high fidelity audio through a combination of adversarial feedback and prediction losses constraining the generated audio to roughly match the ground truth in terms of its total duration and mel-spectrogram. To allow the model to capture temporal variation in the generated audio, we employ soft dynamic time warping in the spectrogram-based prediction loss. The resulting model achieves a mean opinion score exceeding 4 on a 5 point scale, which is comparable to the state-of-the-art models relying on multi-stage training and additional supervision.
****

A unifying view on implicit bias in training linear neural networks
Chulhee Yun,Shankar Krishnan,Hossein Mobahi
We study the implicit bias of gradient flow (i.e., gradient descent with infinitesimal step size) on linear neural network training. We propose a tensor formulation of neural networks that includes fully-connected, diagonal, and convolutional networks as special cases, and investigate the linear version of the formulation called linear tensor networks. With this formulation, we can characterize the convergence direction of the network parameters as singular vectors of a tensor defined by the network. For L-layer linear tensor networks that are orthogonally decomposable, we show that gradient flow on separable classification finds a stationary point of the $\ell_{2/L}$ max-margin problem in a "transformed" input space defined by the network. For underdetermined regression, we prove that gradient flow finds a global minimum which minimizes a norm-like function that interpolates between weighted ℓ_1 and ℓ_2 norms in the transformed input space. Our theorems subsume existing results in the literature while removing standard convergence assumptions. We also provide experiments that corroborate our analysis.
****

A Gradient-based Kernel Approach for Efficient Network Architecture Search
Jingjing Xu,Liang Zhao,Junyang Lin,Xu Sun,Hongxia Yang
It is widely accepted that vanishing and exploding gradient values are the main reason behind the difficulty of deep network training.
In this work, we take a further step to understand the optimization of deep networks and find that both gradient correlations and gradient values have strong impacts on model training.
Inspired by our new finding, we explore a simple yet effective network architecture search (NAS) approach that leverages gradient correlation and gradient values to find well-performing architectures. To be specific, we first formulate these two terms into a unified gradient-based kernel and then select architectures with the largest kernels at initialization as the final networks.
The new approach replaces the expensive ``train-then-test'' evaluation paradigm with a new lightweight function according to the gradient-based kernel at initialization.
Experiments show that our approach achieves competitive results with orders of magnitude faster than ``train-then-test'' paradigms on image classification tasks. Furthermore, the extremely low search cost enables its wide applications. It also obtains performance improvements on two text classification tasks.
****

Deep Kernel Processes
Laurence Aitchison,Adam X. Yang,Sebastian W. Ober
We define deep kernel processes in which positive definite Gram matrices are progressively transformed by nonlinear kernel functions and by sampling from (inverse) Wishart distributions. Remarkably, we find that deep Gaussian processes (DGPs), Bayesian neural networks (BNNs), infinite BNNs, and infinite BNNs with bottlenecks can all be written as deep kernel processes. For DGPs the equivalence arises because the Gram matrix formed by the inner product of features is Wishart d

istributed, and as we show, standard isotropic kernels can be written entirely i
n terms of this Gram matrix (we do not need knowledge of the underlying features
). We define a tractable deep kernel process, the deep inverse Wishart process a
nd give a doubly-stochastic inducing-point variational inference scheme that ope
rates on the Gram matrices, not on the features (as in DGPs). We show that the d
eep inverse Wishart process gives superior performance to DGPs and infinite BNN
on standard fully-connected baselines.
****
Robustness to Pruning Predicts Generalization in Deep Neural Networks
Lorenz Kuhn,Clare Lyle,Aidan Gomez,Jonas Rothfuss,Yarin Gal
Why over-parameterized neural networks generalize as well as they do is a centra
l concern of theoretical analysis in machine learning today. Following Occam's r
azor, it has long been suggested that simpler networks generalize better than mo
re complex ones. Successfully quantifying this principle has proved difficult gi
ven that many measures of simplicity, such as parameter norms, grow with the siz
e of the network and thus fail to capture the observation that larger networks t
end to generalize better in practice.
In this paper, we introduce a new, theoretically motivated measure of a network'
s simplicity: the smallest fraction of the network's parameters that can be kept
 while pruning without adversely affecting its training loss. We show that this
measure is highly predictive of a model's generalization performance across a la
rge set of convolutional networks trained on CIFAR-10. Lastly, we study the mutu
al information between the predictions of our new measure and strong existing me
asures based on models' margin, flatness of minima and optimization speed. We sh
ow that our new measure is similar to -- but more predictive than -- existing fl
atness-based measures.
****
Adversarial Environment Generation for Learning to Navigate the Web
Izzeddin Gur,Natasha Jaques,Kevin Malta,Manoj Tiwari,Honglak Lee,Aleksandra Faus
t
Learning to autonomously navigate the web is a difficult sequential decision mak
ing task. The state and action spaces are large and combinatorial in nature, and
 successful navigation may require traversing several partially-observed pages.
One of the bottlenecks of training web navigation agents is providing a learnabl
e curriculum of training environments that can cover the large variety of real-w
orld websites. Therefore, we propose using Adversarial Environment Generation (A
EG) to generate challenging web environments in which to train reinforcement lea
rning (RL) agents. We introduce a new benchmarking environment, gMiniWoB, which
enables an RL adversary to use compositional primitives to learn to generate com
plex websites. To train the adversary, we present a new decoder-like architectur
e that can directly control the difficulty of the environment, and a new trainin
g technique Flexible b-PAIRED. Flexible b-PAIRED jointly trains the adversary an
d a population of navigator agents and incentivizes the adversary to generate ”j
ust-the-right-challenge” environments by simultaneously learning two policies en
coded in the adversary’s architecture. First, for its environment complexity cho
ice (difficulty budget), the adversary is rewarded with the performance of the b
est-performing agent in the population. Second, for selecting the design element
s the adversary learns to maximize the regret using the difference in capabiliti
es of navigator agents in population (flexible regret). The results show that th
e navigator agent trained with Flexible b-PAIRED generalizes to new environments
, significantly outperforms competitive automatic curriculum generation baseline
s—including a state-of-the-art RL web navigation approach and prior methods for
minimax regret AEG—on a set of challenging unseen test environments that are ord
er of magnitude more complex than the previous benchmarks. The navigator agent a
chieves more than 75% success rate on all tasks, yielding 4x higher success rate
 that the strongest baseline.
****
Coverage as a Principle for Discovering Transferable Behavior in Reinforcement L
earning
Víctor Campos,Pablo Sprechmann,Steven Stenberg Hansen,Andre Barreto,Charles Blun

dell,Alex Vitvitskyi,Steven Kapturowski,Adria Puigdomenech Badia
Designing agents that acquire knowledge autonomously and use it to solve new tas
ks efficiently is an important challenge in reinforcement learning. Unsupervised
 learning provides a useful paradigm for autonomous acquisition of task-agnostic
 knowledge. In supervised settings, representations discovered through unsupervi
sed pre-training offer important benefits when transferred to downstream tasks.
Given the nature of the reinforcement learning problem, we explore how to transf
er knowledge through behavior instead of representations. The behavior of pre-tr
ained policies may be used for solving the task at hand (exploitation), as well
as for collecting useful data to solve the problem (exploration). We argue that
pre-training policies to maximize coverage will result in behavior that is usefu
l for both strategies. When using these policies for both exploitation and explo
ration, our agents discover solutions that lead to larger returns. The largest g
ains are generally observed in domains requiring structured exploration, includi
ng settings where the behavior of the pre-trained policies is misaligned with th
e downstream task.
****

A straightforward line search approach on the expected empirical loss for stocha
stic deep learning problems
Maximus Mutschler,Andreas Zell
   A fundamental challenge in deep learning is that the optimal step sizes for u
pdate steps of stochastic gradient descent are unknown. In traditional optimizat
ion, line searches are used to determine good step sizes, however, in deep learn
ing, it is too costly to search for good step sizes on the expected empirical lo
ss due to noisy losses. This empirical work shows that it is possible to approxi
mate the expected empirical loss on vertical cross sections for common deep lear
ning tasks considerably cheaply. This is achieved by applying traditional one-di
mensional function fitting to measured noisy losses of such cross sections. The
step to a minimum of the resulting approximation is then used as step size for t
he optimization. This approach leads to a robust and straightforward optimizatio
n method which performs well across datasets and architectures without the need
of hyperparameter tuning.

****

What to Prune and What Not to Prune at Initialization
Maham Haroon
Post-training dropout based approaches achieve high sparsity and are well establ
ished means of deciphering problems relating to computational cost and overfitti
ng in Neural Network architectures. Contrastingly, pruning at initialization is
still far behind. Initialization pruning is more efficacious when it comes to sc
aling computation cost of the network. Furthermore, it handles overfitting just
as well as post training dropout. It is also averse to retraining losses.

In approbation of the above reasons, the paper presents two approaches to prune
at initialization. The goal is to achieve higher sparsity while preserving perfo
rmance. 1) K-starts, begins with k random p-sparse matrices at initialization. I
n the first couple of epochs the network then determines the "fittest" of these
p-sparse matrices in an attempt to find the "lottery ticket" p-sparse network. T
he approach is adopted from how evolutionary algorithms find the best individual
. Depending on the Neural Network architecture, fitness criteria can be based on
 magnitude of network weights, magnitude of gradient accumulation over an epoch
or a combination of both. 2) Dissipating gradients approach, aims at eliminating
 weights that remain within a fraction of their initial value during the first c
ouple of epochs. Removing weights in this manner despite their magnitude best pr
eserves performance of the network. Contrarily, the approach also takes the most
 epochs to achieve higher sparsity. 3) Combination of dissipating gradients and
kstarts outperforms either methods and random dropout consistently.

The benefits of using the provided pertaining approaches are: 1) They do not req
uire specific knowledge of the classification task, fixing of dropout threshold

or regularization parameters 2) Retraining of the model is neither necessary nor affects the performance of the p-sparse network.

We evaluate the efficacy of the said methods on Autoencoders and Fully Connected Multilayered Perceptrons. The datasets used are MNIST and Fashion MNIST.
****
Balancing Constraints and Rewards with Meta-Gradient D4PG
Dan A. Calian,Daniel J Mankowitz,Tom Zahavy,Zhongwen Xu,Junhyuk Oh,Nir Levine,Timothy Mann
Deploying Reinforcement Learning (RL) agents to solve real-world applications often requires satisfying complex system constraints. Often the constraint thresholds are incorrectly set due to the complex nature of a system or the inability to verify the thresholds offline (e.g, no simulator or reasonable offline evaluation procedure exists). This results in solutions where a task cannot be solved without violating the constraints. However, in many real-world cases, constraint violations are undesirable yet they are not catastrophic, motivating the need for soft-constrained RL approaches. We present two soft-constrained RL approaches that utilize meta-gradients to find a good trade-off between expected return and minimizing constraint violations. We demonstrate the effectiveness of these approaches by showing that they consistently outperform the baselines across four different Mujoco domains.
****
Warpspeed Computation of Optimal Transport, Graph Distances, and Embedding Alignment
Johannes Klicpera,Marten Lienen,Stephan Günnemann
Optimal transport (OT) is a cornerstone of many machine learning tasks. The current best practice for computing OT is via entropy regularization and Sinkhorn iterations. This algorithm runs in quadratic time and requires calculating the full pairwise cost matrix, which is prohibitively expensive for large sets of objects. To alleviate this limitation we propose to instead use a sparse approximation of the cost matrix based on locality sensitive hashing (LSH). Moreover, we fuse this sparse approximation with the Nyström method, resulting in the locally corrected Nyström method (LCN). These approximations enable general log-linear time algorithms for entropy-regularized OT that perform well even in complex, high-dimensional spaces. We thoroughly demonstrate these advantages via a theoretical analysis and by evaluating multiple approximations both directly and as a component of two real-world models. Using approximate Sinkhorn for unsupervised word embedding alignment enables us to train the model full-batch in a fraction of the time while improving upon the original on average by 3.1 percentage points without any model changes. For graph distance regression we propose the graph transport network (GTN), which combines graph neural networks (GNNs) with enhanced Sinkhorn and outcompetes previous models by 48%. LCN-Sinkhorn enables GTN to achieve this while still scaling log-linearly in the number of nodes.
****
Disentangling Action Sequences: Discovering Correlated Samples
Jiantao Wu,Chunxiuzi Liu,Lin Wang
Disentanglement is a highly desirable property of representation due to its similarity with human’s understanding and reasoning. This improves interpretability, enables the performance of down-stream tasks, and enables controllable generative models.However, this domain is challenged by the abstract notion and incomplete theories to support unsupervised disentanglement learning. We demonstrate the data itself, such as the orientation of images, plays a crucial role in disentanglement and instead of the factors, and the disentangled representations align the latent variables with the action sequences. We further introduce the concept of disentangling action sequences which facilitates the description of the behaviours of the existing disentangling approaches. An analogy for this process is to discover the commonality between the things and categorizing them.

Furthermore, we analyze the inductive biases on the data and find that the latent information thresholds are correlated with the significance of the actions. Fo

r the supervised and unsupervised settings, we respectively introduce two method s to measure the thresholds. We further propose a novel framework, fractional va riational autoencoder (FVAE), to disentangle the action sequences with different significance step-by-step. Experimental results on dSprites and 3D Chairs show that FVAE improves the stability of disentanglement.
****

Improving Model Robustness with Latent Distribution Locally and Globally
Zhuang QIAN,Shufei Zhang,Kaizhu Huang,Qiufeng Wang,Rui Zhang,Xinping Yi
We propose a novel adversarial training method which leverages both the local an d global information to defend adversarial attacks. Existing adversarial traini ng methods usually generate adversarial perturbations locally in a supervised m anner and fail to consider the data manifold information in a global way. Conseq uently, the resulting adversarial examples may corrupt the underlying data stru cture and are typically biased towards the decision boundary. In this work, we e xploit both the local and global information of data manifold to generate advers arial examples in an unsupervised manner. Specifically, we design our novel fram ework via an adversarial game between a discriminator and a classifier: the di scriminator is learned to differentiate the latent distributions of the natura l data and the perturbed counterpart, while the classifier is trained to recogni ze accurately the perturbed examples as well as enforcing the invariance between the two latent distributions. We conduct a series of analysis on the model robu stness and also verify the effectiveness of our proposed method empirically. Exp erimental results show that our method substantially outperforms the recent sta te-of-the-art (i.e. Feature Scattering) in defending adversarial attacks by a l arge accuracy margin (e.g. 17.0% and 18.1% on SVHN dataset, 9.3% and $ 17.4\%$ on CIFAR-10 dataset, 6.0% and 16.2% on CIFAR-100 dataset for defen ding PGD20 and CW20 attacks respectively).
****

Decentralized Deterministic Multi-Agent Reinforcement Learning
Antoine Grosnit,Desmond Cai,Laura Wynter
Recent work in multi-agent reinforcement learning (MARL) by [Zhang, ICML12018] p rovided the first decentralized actor-critic algorithm to offer convergence guar antees. In that work, policies are stochastic and are defined on finite action s paces. We extend those results to develop a provably-convergent decentralized ac tor-critic algorithm for learning deterministic policies on continuous action sp aces. Deterministic policies are important in many real-world settings. To handl e the lack of exploration inherent in deterministic policies we provide results for the off-policy setting as well as the on-policy setting. We provide the main ingredients needed for this problem: the expression of a local deterministic po licy gradient, a decentralized deterministic actor-critic algorithm, and converg ence guarantees when the value functions are approximated linearly. This work en ables decentralized MARL in high-dimensional action spaces and paves the way for more widespread application of MARL.
****

Robust Curriculum Learning: from clean label detection to noisy label self-corre ction
Tianyi Zhou,Shengjie Wang,Jeff Bilmes
Neural network training can easily overfit noisy labels resulting in poor genera lization performance. Existing methods address this problem by (1) filtering out the noisy data and only using the clean data for training or (2) relabeling the noisy data by the model during training or by another model trained only on a c lean dataset. However, the former does not leverage the features' information of wrongly-labeled data, while the latter may produce wrong pseudo-labels for some data and introduce extra noises. In this paper, we propose a smooth transition and interplay between these two strategies as a curriculum that selects training samples dynamically. In particular, we start with learning from clean data and then gradually move to learn noisy-labeled data with pseudo labels produced by a time-ensemble of the model and data augmentations. Instead of using the instant aneous loss computed at the current step, our data selection is based on the dyn amics of both the loss and output consistency for each sample across historical

steps and different data augmentations, resulting in more precise detection of b
oth clean labels and correct pseudo labels. On multiple benchmarks of noisy labe
ls, we show that our curriculum learning strategy can significantly improve the
test accuracy without any auxiliary model or extra clean data.
****
SkillBERT: “Skilling” the BERT to classify skills!
Amber Nigam,Shikha Tyagi,Kuldeep Tyagi,Arpan Saxena
In the age of digital recruitment, job posts can attract a large number of appli
cations, and screening them manually can become a very tedious task. These recru
itment records are stored in the form of tables in our recruitment database (Ele
ctronic Recruitment Records, referred to as ERRs). We have released a de-identif
ied ERR dataset to the public domain. We also propose a BERT-based model, SkillB
ERT, the embeddings of which are used as features for classifying skills present
 in the ERRs into groups referred to as "competency groups". A competency group
is a group of similar skills and it is used as matching criteria (instead of mat
ching on skills) for finding the overlap of skills between the candidates and th
e jobs. This proxy match takes advantage of the BERT's capability of deriving me
aning from the structure of competency groups present in the skill dataset. In o
ur experiments, the SkillBERT, which is trained from scratch on the skills prese
nt in job requisitions, is shown to be better performing than the pre-trained BE
RT and the Word2Vec. We have also explored K-means clustering and spectral clust
ering on SkillBERT embeddings to generate cluster-based features. Both algorithm
s provide similar performance benefits. Last, we have experimented with differen
t machine learning algorithms like Random Forest, XGBoost, and a deep learning a
lgorithm Bi-LSTM . We did not observe a significant performance difference among
 the algorithms, although XGBoost and Bi-LSTM perform slightly better than Rando
m Forest. The features created using SkillBERT are most predictive in the classi
fication task, which demonstrates that the SkillBERT is able to capture informat
ion about the skills' ontology from the data. We have made the source code and t
he trained models of our experiments publicly available.
****
Clairvoyance: A Pipeline Toolkit for Medical Time Series
Daniel Jarrett,Jinsung Yoon,Ioana Bica,Zhaozhi Qian,Ari Ercole,Mihaela van der S
chaar
Time-series learning is the bread and butter of data-driven *clinical decision s
upport*, and the recent explosion in ML research has demonstrated great potentia
l in various healthcare settings. At the same time, medical time-series problems
 in the wild are challenging due to their highly *composite* nature: They entail
 design choices and interactions among components that preprocess data, impute m
issing values, select features, issue predictions, estimate uncertainty, and int
erpret models. Despite exponential growth in electronic patient data, there is a
 remarkable gap between the potential and realized utilization of ML for clinica
l research and decision support. In particular, orchestrating a real-world proje
ct lifecycle poses challenges in engineering (i.e. hard to build), evaluation (i
.e. hard to assess), and efficiency (i.e. hard to optimize). Designed to address
 these issues simultaneously, Clairvoyance proposes a unified, end-to-end, autoM
L-friendly pipeline that serves as a (i) software toolkit, (ii) empirical standa
rd, and (iii) interface for optimization. Our ultimate goal lies in facilitating
 transparent and reproducible experimentation with complex inference workflows,
providing integrated pathways for (1) personalized prediction, (2) treatment-eff
ect estimation, and (3) information acquisition. Through illustrative examples o
n real-world data in outpatient, general wards, and intensive-care settings, we
illustrate the applicability of the pipeline paradigm on core tasks in the healt
hcare journey. To the best of our knowledge, Clairvoyance is the first to demons
trate viability of a comprehensive and automatable pipeline for clinical time-se
ries ML.
****
Machine Learning Algorithms for Data Labeling: An Empirical Evaluation
Teodor Anders Fredriksson,David Issa Mattos,Jan Bosch,Helena Holmström Olsson
The lack of labeled data is a major problem in both research and industrial sett

ings since obtaining labels is often an expensive and time-consuming activity. In the past years, several machine learning algorithms were developed to assist and perform automated labeling in partially labeled datasets. While many of these algorithms are available in open-source packages, there is no research that investigates how these algorithms compare to each other in different types of datasets and with different percentages of available labels. To address this problem, this paper empirically evaluates and compares seven algorithms for automated labeling in terms of accuracy. We investigate how these algorithms perform in six different and well-known datasets with three different types of data, images, texts, and numerical values. We evaluate these algorithms under two different experimental conditions, with 10\% and 50\% labels of available labels in the dataset. Each algorithm, in each dataset for each experimental condition, is evaluated independently ten times with different random seeds. The results are analyzed and the algorithms are compared utilizing a Bayesian Bradley-Terry model. The results indicate that while the algorithms label spreading with K-nearest neighbors perform better in the aggregated results, the active learning algorithms query by instance QBC and query instance uncertainty sample perform better when there is only 10\% of labels available. These results can help machine learning practitioners in choosing optimal machine learning algorithms to label their data.
****

Regularization Cocktails for Tabular Datasets
Arlind Kadra,Marius Lindauer,Frank Hutter,Josif Grabocka
The regularization of prediction models is arguably the most crucial ingredient that allows Machine Learning solutions to generalize well on unseen data. Several types of regularization are popular in the Deep Learning community (e.g., weight decay, drop-out, early stopping, etc.), but so far these are selected on an ad-hoc basis, and there is no systematic study as to how different regularizers should be combined into the best “cocktail”. In this paper, we fill this gap, by considering the cocktails of 13 different regularization methods and framing the question of how to best combine them as a standard hyperparameter optimization problem. We perform a large-scale empirical study on 40 tabular datasets, concluding that, firstly, regularization cocktails substantially outperform individual regularization methods, even if the hyperparameters of the latter are carefully tuned; secondly, the optimal regularization cocktail depends on the dataset; and thirdly, regularization cocktails yield the state-of-the-art in classifying tabular datasets by outperforming Gradient-Boosted Decision Trees.
****

How Important is Importance Sampling for Deep Budgeted Training?
Eric Arazo,Diego Ortego,Paul Albert,Noel O'Connor,Kevin McGuinness
Long iterative training processes for Deep Neural Networks (DNNs) are commonly required to achieve state-of-the-art performance in many computer vision tasks. Core-set selection and importance sampling approaches might play a key role in budgeted training regimes, i.e.~when limiting the number of training iterations. The former demonstrate that retaining informative samples is important to avoid large drops in accuracy, and the later aim at dynamically estimating the sample importance to speed-up convergence. This work explores this paradigm and how a budget constraint interacts with importance sampling approaches and data augmentation techniques. We show that under budget restrictions, importance sampling approaches do not provide a consistent improvement over uniform sampling. We suggest that, given a specific budget, the best course of action is to disregard the importance and introduce adequate data augmentation. For example, training in CIFAR-10/100 with 30% of the full training budget, a uniform sampling strategy with certain data augmentation surpasses the performance of 100% budget models trained with standard data augmentation. We conclude from our work that DNNs under budget restrictions benefit greatly from variety in the samples and that finding the right samples to train is not the most effective strategy when balancing high performance with low computational requirements. The code will be released after the review process.
****

Generative Adversarial User Privacy in Lossy Single-Server Information Retrieval

Chung-Wei Weng,Yauhen Yakimenka,Hsuan-Yin Lin,Eirik Rosnes,Joerg Kliewer
We consider the problem of information retrieval from a dataset of files stored on a single server under both a user distortion and a user privacy constraint. Specifically, a user requesting a file from the dataset should be able to reconstruct the requested file with a prescribed distortion, and in addition, the identity of the requested file should be kept private from the server with a prescribed privacy level. The proposed model can be seen as an extension of the well-known concept of private information retrieval by allowing for distortion in the retrieval process and relaxing the perfect privacy requirement. We initiate the study of the tradeoff between download rate, distortion, and user privacy leakage, and show that the optimal rate-distortion-leakage tradeoff is convex and that it allows for a concise information-theoretical formulation in terms of mutual information in the limit of large file sizes. Moreover, we propose a new data-driven framework by leveraging recent advancements in generative adversarial models which allows a user to learn efficient schemes in terms of download rate from the data itself. Learning the scheme is formulated as a constrained minimax game between a user which desires to keep the identity of the requested file private and an adversary that tries to infer which file the user is interested in under a distortion constraint. In general, guaranteeing a certain privacy level leads to a higher rate-distortion tradeoff curve, and hence a sacrifice in either download rate or distortion. We evaluate the performance of the scheme on a synthetic Gaussian dataset as well as on the MNIST dataset. For the MNIST dataset, the data-driven approach significantly outperforms a proposed general achievable scheme combining source coding with the download of multiple files.
****

Monotonic neural network: combining deep learning with domain knowledge for chiller plants energy optimization
Fanhe Ma,Faen Zhang,Shenglan Ben,Shuxin Qin,pengcheng Zhou,Changsheng Zhou,Fengyi Xu
In this paper, we are interested in building a domain knowledge based deep learning framework to solve the chiller plants energy optimization problems. Compared to the hotspot applications of deep learning (e.g. image classification and NLP), it is difficult to collect enormous data for deep network training in real-world physical systems. Most existing methods reduce the complex systems into linear model to facilitate the training on small samples. To tackle the small sample size problem, this paper considers domain knowledge in the structure and loss design of deep network to build a nonlinear model with lower redundancy function space. Specifically, the energy consumption estimation of most chillers can be physically viewed as an input-output monotonic problem. Thus, we can design a Neural Network with monotonic constraints to mimic the physical behavior of the system. We verify the proposed method in a cooling system of a data center, experimental results show the superiority of our framework in energy optimization compared to the existing ones.
****

A Probabilistic Model for Discriminative and Neuro-Symbolic Semi-Supervised Learning
Carl Allen,Ivana Balazevic,Timothy Hospedales
Strong progress has been achieved in semi-supervised learning (SSL) by combining several methods, some of which relate to properties of the data distribution p(x), others to the model outputs p(y|x), e.g. minimising the entropy of unlabelled predictions. Focusing on the latter, we fill a gap in the standard text by introducing a probabilistic model for discriminative semi-supervised learning, mirroring the classical generative model. Several SSL methods are theoretically explained by our model as inducing (approximate) strong priors over parameters of p(y|x). Applying this same probabilistic model to tasks in which labels represent binary attributes, we theoretically justify a family of neuro-symbolic SSL approaches, taking a step towards bridging the divide between statistical learning and logical reasoning.
****

Solving NP-Hard Problems on Graphs with Extended AlphaGo Zero

Kenshin Abe,Zijian Xu,Issei Sato,Masashi Sugiyama
There have been increasing challenges to solve combinatorial optimization problems by machine learning.
Khalil et al. (NeurIPS 2017) proposed an end-to-end reinforcement learning framework, which automatically learns graph embeddings to construct solutions to a wide range of problems.
However, it sometimes performs poorly on graphs having different characteristics than training graphs.
To improve its generalization ability to various graphs, we propose a novel learning strategy based on AlphaGo Zero, a Go engine that achieved a superhuman level without the domain knowledge of the game.
We redesign AlphaGo Zero for combinatorial optimization problems, taking into account several differences from two-player games.
In experiments on five NP-hard problems such as {\sc MinimumVertexCover} and {\sc MaxCut}, our method, with only a policy network, shows better generalization than the previous method to various instances that are not used for training, including random graphs, synthetic graphs, and real-world graphs.
Furthermore, our method is significantly enhanced by a test-time Monte Carlo Tree Search which makes full use of the policy network and value network.
We also compare recently-developed graph neural network (GNN) models, with an interesting insight into a suitable choice of GNN models for each task.
****
Co-complexity: An Extended Perspective on Generalization Error
Rohan Ghosh,Mehul Motani
It is well known that the complexity of a classifier's function space controls its generalization gap, with two important examples being VC-dimension and Rademacher complexity (R-Complexity). We note that these traditional generalization error bounds consider the ground truth label generating function (LGF) to be fixed. However, if we consider a scenario where the LGF has no constraints at all, then the true generalization error can be large, irrespective of training performance, as the values of the LGF on unseen data points can be largely independent of the values on the training data. To account for this, in this work, we consider an extended characterization of the problem, where the ground truth labels are generated by a function within another function space, which we call the \textit{generator} space. We find that the generalization gap in this scenario depends on the R-Complexity of both the classifier and the generator function spaces. Thus, we find that, even if the R-Complexity of the classifier is low and it has a good training fit, a highly complex generator space could worsen generalization performance, in accordance with the no free lunch theorem. Furthermore, the characterization of a generator space allows us to model constraints, such as invariances (translation and scale in vision) or local smoothness. Subsequently, we propose a joint entropy-like measure of complexity between function spaces (classifier and generator), called co-complexity, which leads to tighter bounds on the generalization error in this setting. Co-complexity captures the similarities between the classifier and generator spaces. It can be decomposed into an invariance co-complexity term, which measures the extent to which the classifier respects the invariant transformations in the generator, and a dissociation co-complexity term, which measures the ability of the classifier to differentiate separate categories in the generator. Our major finding is that reducing the invariance co-complexity of a classifier, while maintaining its dissociation co-complexity, improves the training error and reduces the generalization gap. Furthermore, our results, when specialized to the previous setting where the LGF is fixed, lead to tighter generalization error bounds. Theoretical results are supported by empirical validation on the CNN architecture and its transformation-equivariant extensions. Co-complexity showcases a new side to the generalization abilities of classifiers and can potentially be used to improve their design.
****
Thinking Like Transformers
Gail Weiss,Yoav Goldberg,Eran Yahav
What is the computational model behind a transformer? Where recurrent neural net

works have direct parallels in finite state machines, allowing clear discussion and thought around architecture variants or trained models, transformers have no such familiar parallel. In this paper we aim to change that, proposing a computational model for the transformer-encoder in the form of a programming language. We map the basic components of a transformer-encoder – attention and feed-forward computation – into the simple primitives of select, aggregate, and zipmap, around which we form a programming language: the Restricted Access Sequence Processing Language (RASP). We show how RASP can be used to program solutions to tasks that could conceivably be learned by a transformer, augmenting it with tools we discover in our work. In particular, we provide RASP programs for histograms, sorting, and even logical inference similar to that of Clark et al. (2020). We further use our model to relate their difficulty in terms of the number of required layers and attention heads. Finally, we see how insights gained from our abstraction might be used to explain phenomena seen in recent works.
****
Q-Value Weighted Regression: Reinforcement Learning with Limited Data
Piotr Kozakowski,Lukasz Kaiser,Henryk Michalewski,Afroz Mohiuddin,Katarzyna Ka■s ka
Sample efficiency and performance in the offline setting have emerged as among the main
challenges of deep reinforcement learning. We introduce Q-Value Weighted Regression (QWR),
a simple RL algorithm that excels in these aspects.

QWR is an extension of Advantage Weighted Regression (AWR), an off-policy actor-critic algorithm
that performs very well on continuous control tasks, also in the offline setting, but struggles
on tasks with discrete actions and in sample efficiency. We perform a theoretical analysis
of AWR that explains its shortcomings and use the insights to motivate QWR theoretically.

We show experimentally that QWR matches state-of-the-art algorithms both on tasks with
continuous and discrete actions. We study the main hyperparameters of QWR
and find that it is stable in a wide range of their choices and on different tasks.
In particular, QWR yields results on par with SAC on the MuJoCo suite and - with the same set of hyperparameters -- yields results on par with a highly tuned Rainbow
implementation on a set of Atari games. We also verify that QWR performs well in the
offline RL setting, making it a compelling choice for reinforcement learning in domains
with limited data.
****
Untangle: Critiquing Disentangled Recommendations
Preksha Nema,Alexandros Karatzoglou,Filip Radlinski
The core principle behind most collaborative filtering methods is to embed users and items in latent spaces, where individual dimensions are learned independently of any particular item attributes. It is thus difficult for users to control their recommendations based on particular aspects (critiquing). In this work, we propose Untangle: a recommendation model that gives users control over the recommendation list with respect to specific item attributes, (e.g.:less violent, funnier movies) that have a causal relationship in user preferences. Untangle uses a refined training procedure by training (i) a (partially) supervised β-VAE that disentangles the item representations and (ii) a second phase which optimized to generate recommendations for users. Untangle gives control on critiquing recommendations based on users preferences, without sacrificing on recommendation ac

curacy. Moreover only a tiny fraction of labeled items is needed to create disen tangled preference representations over attributes.

****
A spherical analysis of Adam with Batch Normalization
Simon Roburin,Yann Dubois de Mont-Marin,Andrei Bursuc,Renaud Marlet,Patrick Pere z,Mathieu Aubry
Batch Normalization (BN) is a prominent deep learning technique. In spite of its apparent simplicity, its implications over optimization are yet to be fu lly understood. While previous studies mostly focus on the interaction between BN and stochastic gradient descent (SGD), we develop a geometric perspective whi ch allows us to precisely characterize the relation between BN and Adam. More pr ecisely, we leverage the radial invariance of groups of parameters, such as filt ers for convolutional neural networks, to translate the optimization steps on th e L_2 unit hypersphere. This formulation and the associated geometric interpre tation shed new light on the training dynamics. Firstly, we use it to derive the first effective learning rate expression of Adam. Then we show that, in the pre sence of BN layers, performing SGD alone is actually equivalent to a variant of Adam constrained to the unit hypersphere. Finally, our analysis outlines phenome na that previous variants of Adam act on and we experimentally validate their im portance in the optimization process.
****
Sparse Gaussian Process Variational Autoencoders
Matthew Ashman,Jonathan So,William Tebbutt,Vincent Fortuin,Michael Arthur Leopol d Pearce,Richard E Turner
Large, multi-dimensional spatio-temporal datasets are omnipresent in modern scie nce and engineering. An effective framework for handling such data are Gaussian process deep generative models (GP-DGMs), which employ GP priors over the latent variables of DGMs. Existing approaches for performing inference in GP-DGMs do n ot support sparse GP approximations based on inducing points, which are essentia l for the computational efficiency of GPs, nor do they handle missing data -- a natural occurrence in many spatio-temporal datasets -- in a principled manner. W e address these shortcomings with the development of the sparse Gaussian process variational autoencoder (SGP-VAE), characterised by the use of partial inferenc e networks for parameterising sparse GP approximations. Leveraging the benefits of amortised variational inference, the SGP-VAE enables inference in multi-outpu t sparse GPs on previously unobserved data with no additional training. The SGP-VAE is evaluated in a variety of experiments where it outperforms alternative ap proaches including multi-output GPs and structured VAEs.
****
Learning a Transferable Scheduling Policy for Various Vehicle Routing Problems b ased on Graph-centric Representation Learning
Inwook Kim,Jinkyoo Park
Reinforcement learning has been used to learn to solve various routing problems. however, most of the algorithm is restricted to finding an optimal routing stra tegy for only a single vehicle. In addition, the trained policy under a specific target routing problem is not able to solve different types of routing problems with different objectives and constraints. This paper proposes an reinforcement learning approach to solve the min-max capacitated multi vehicle routing proble m (mCVRP), the problem seeks to minimize the total completion time for multiple vehicles whose one-time traveling distance is constrained by their fuel levels t o serve the geographically distributed customer nodes. The method represents the relationships among vehicles, customers, and fuel stations using relationship-s pecific graphs to consider their topological relationships and employ graph neur al network (GNN) to extract the graph's embedding to be used to make a routing a ction. We train the proposed model using the random mCVRP instance with differen t numbers of vehicles, customers, and refueling stations. We then validate that the trained policy solve not only new mCVRP problems with different complexity (weak transferability but also different routing problems (CVRP, mTSP, TSP) with

different objectives and constraints (storing transferability).
****

Plan-Based Relaxed Reward Shaping for Goal-Directed Tasks
Ingmar Schubert,Ozgur S Oguz,Marc Toussaint
In high-dimensional state spaces, the usefulness of Reinforcement Learning (RL) is limited by the problem of exploration. This issue has been addressed using potential-based reward shaping (PB-RS) previously. In the present work, we introduce Final-Volume-Preserving Reward Shaping (FV-RS). FV-RS relaxes the strict optimality guarantees of PB-RS to a guarantee of preserved long-term behavior. Being less restrictive, FV-RS allows for reward shaping functions that are even better suited for improving the sample efficiency of RL algorithms. In particular, we consider settings in which the agent has access to an approximate plan. Here, we use examples of simulated robotic manipulation tasks to demonstrate that plan-based FV-RS can indeed significantly improve the sample efficiency of RL over plan-based PB-RS.
****

VideoGen: Generative Modeling of Videos using VQ-VAE and Transformers
Yunzhi Zhang,Wilson Yan,Pieter Abbeel,Aravind Srinivas
We present VideoGen: a conceptually simple architecture for scaling likelihood based generative modeling to natural videos. VideoGen uses VQ-VAE that learns learns downsampled discrete latent representations of a video by employing 3D convolutions and axial self-attention. A simple GPT-like architecture is then used to autoregressively model the discrete latents using spatio-temporal position encodings. Despite the simplicity in formulation, ease of training and a light compute requirement, our architecture is able to generate samples competitive with state-of-the-art GAN models for video generation on the BAIR Robot dataset, and generate coherent action-conditioned samples based on experiences gathered from the VizDoom simulator. We hope our proposed architecture serves as a reproducible reference for a minimalistic implementation of transformer based video generation models without requiring industry scale compute resources. Samples are available at https://sites.google.com/view/videogen
****

Learning without Forgetting: Task Aware Multitask Learning for Multi-Modality Tasks
Sathish Reddy Indurthi,Mohd Abbas Zaidi,Nikhil Kumar Lakumarapu,Beomseok Lee,Hyo Jung Han,Sangha Kim,Inchul Hwang
Existing joint learning strategies like multi-task or meta-learning focus more on shared learning and have little to no scope for task-specific learning. This creates the need for a distinct shared pretraining phase and a task-specific fine tuning phase. The fine-tuning phase creates separate systems for each task, where improving the performance of a particular task necessitates forgetting some of the knowledge garnered in other tasks. Humans, on the other hand, perform task-specific learning in synergy with general domain-based learning. Inspired by these learning patterns in humans, we suggest a simple yet generic task aware framework to incorporate into existing joint learning strategies. The proposed framework computes task-specific representations to modulate model parameters during joint learning. Hence, it performs both shared and task-specific learning in a single-phase resulting in a single model for all the tasks. The single model itself achieves significant performance gains over the existing joint learning strategies. For example, we train a model on Speech Translation (ST), Automatic Speech Recognition (ASR), and Machine Translation (MT) tasks using the proposed task aware joint learning approach. This single model achieves a performance of 28.64 BLEU score on ST MuST-C English-German, WER of 11.61 on ASR TEDLium v3, and BLEU score of 23.35 on MT WMT14 English-German tasks. This sets a new state-of-the-art performance (SOTA) on the ST task while outperforming the existing end-to-end ASR systems with a competitive performance on the MT task.
****

Sample efficient Quality Diversity for neural continuous control
Thomas PIERROT,Valentin Macé,Geoffrey Cideron,Nicolas Perrin,Karim Beguir,Olivier Sigaud

We propose a novel Deep Neuroevolution algorithm, QD-RL, that combines the stren gths of off-policy reinforcement learning (RL) algorithms and Quality Diversity (QD) approaches to solve continuous control problems with neural controllers. Th e QD part contributes structural biases by decoupling the search for diversity f rom the search for high return, resulting in efficient management of the explora tion-exploitation trade-off. The RL part contributes sample efficiency by relyin g on off-policy gradient-based updates of the agents. More precisely, we train a population of off-policy deep RL agents to simultaneously maximize diversity wi thin the population and the return of each individual agent. QD-RL selects agent s interchangeably from a Pareto front or from a Map-Elites grid, resulting in st able and efficient population updates. Our experiments in the Ant-Maze and Ant-T rap environments show that QD-RL can solve challenging exploration and control p roblems with deceptive rewards while being two orders of magnitude more sample e fficient than the evolutionary counterpart.
****

On the Geometry of Deep Bayesian Active Learning
Xiaofeng Cao,Ivor Tsang
We present geometric Bayesian active learning by disagreements (GBALD), a frame work that performs BALD on its geometric interpretation interacting with a deep learning model. There are two main components in GBALD: initial acquisitions bas ed on core-set construction and model uncertainty estimation with those initial acquisitions. Our key innovation is to construct the core-set on an ellipsoid, n ot typical sphere, preventing its updates towards the boundary regions of the di stributions. Main improvements over BALD are twofold: relieving sensitivity to u ninformative prior and reducing redundant information of model uncertainty. To g uarantee the improvements, our generalization analysis proves that, compared to typical Bayesian spherical interpretation, geodesic search with ellipsoid can d erive a tighter lower error bound and achieve higher probability to obtain a ne arly zero error. Experiments on acquisitions with several scenarios demonstrate that, yielding slight perturbations to noisy and repeated samples, GBALD furthe r achieves significant accuracy improvements than BALD, BatchBALD and other bas elines.
****

Improving VAEs' Robustness to Adversarial Attack
Matthew JF Willetts,Alexander Camuto,Tom Rainforth,S Roberts,Christopher C Holme s
Variational autoencoders (VAEs) have recently been shown to be vulnerable to adv ersarial attacks, wherein they are fooled into reconstructing a chosen target im age. However, how to defend against such attacks remains an open problem. We mak e significant advances in addressing this issue by introducing methods for produ cing adversarially robust VAEs. Namely, we first demonstrate that methods propos ed to obtain disentangled latent representations produce VAEs that are more robu st to these attacks. However, this robustness comes at the cost of reducing the quality of the reconstructions. We ameliorate this by applying disentangling met hods to hierarchical VAEs. The resulting models produce high--fidelity autoencod ers that are also adversarially robust. We confirm their capabilities on several different datasets and with current state-of-the-art VAE adversarial attacks, a nd also show that they increase the robustness of downstream tasks to attack.
****

ZCal: Machine learning methods for calibrating radio interferometric data
Simphiwe Zitha,Arun aniyan,Oleg Smirnov,Risuna Nkolele
Calibration is the most critical data processing step needed for generating imag es of high dynamic range \citep{editioncasa}. With ever-increasing data volumes produced by modern radio telescopes \cite{aniyan2017classifying}, astronomers ar e overwhelmed by the amount of data that needs to be manually processed and anal yzed using limited computational resources \citep{yatawatta2020stochastic}. Ther efore, intelligent and automated systems are required to overcome these challeng es. Traditionally, astronomers use a package such as Common Astronomy Software A pplications (CASA) to compute the gain solutions based on regular observations o f a known calibrator source \citep{thompson2017interferometry} \citep{abebe2015s

tudy} \citep{grobler2016calibration} \citep{editioncasa}. The traditional approa
ch to calibration is iterative and time-consuming \citep{jajarmizadeh2017optimal
}, thus, the proposal of machine learning techniques. The applications of machin
e learning have created an opportunity to deal with complex problems currently e
ncountered in radio astronomy data processing \citep{aniyan2017classifying}. In
this work, we propose the use of supervised machine learning models to first gen
eration calibration (1GC), using the KAT-7 telescope environmental and pointing
sensor data recorded during observations. Applying machine learning to 1GC, as o
pposed to calculating the gain solutions in CASA, has shown evidence of reducing
 computation, as well as accurately predicting the 1GC gain solutions and antenn
a behaviour. These methods are computationally less expensive, however they have
 not fully learned to generalise in predicting accurate 1GC solutions by looking
 at environmental and pointing sensors. We use an ensemble multi-output regressi
on models based on random forest, decision trees, extremely randomized trees and
 K-nearest neighbor algorithms. The average prediction error obtained during the
 testing of our models on testing data is $ \approx 0.01 < rmse < 0.09$ for gain
 amplitude per antenna, and $0.2 rad < rmse <0.5 rad$ for gain phase. This shows
 that the instrumental parameters used to train our model strongly correlate wit
h gain amplitude effects than a phase.

****
Reinforcement Learning for Control with Probabilistic Stability Guarantee
Minghao Han,Zhipeng Zhou,Lixian Zhang,Jun Wang,Wei Pan
Reinforcement learning is promising to control dynamical systems for which the t
raditional control methods are hardly applicable. However, in control theory, th
e stability of a closed-loop system can be hardly guaranteed using the policy/co
ntroller learned solely from samples. In this paper, we will combine Lyapunov's
method in control theory and stochastic analysis to analyze the mean square stab
ility of MDP in a model-free manner. Furthermore, the finite sample bounds on th
e probability of stability are derived as a function of the number M and length
T of the sampled trajectories. And we show that there is a lower bound on T and
the probability is much more demanding for M than T. Based on the theoretical re
sults, a REINFORCE like algorithm is proposed to learn the controller and the Ly
apunov function simultaneously.
****
Gauge Equivariant Mesh CNNs: Anisotropic convolutions on geometric graphs
Pim De Haan,Maurice Weiler,Taco Cohen,Max Welling
A common approach to define convolutions on meshes is to interpret them as a gra
ph and apply graph convolutional networks (GCNs). Such GCNs utilize isotropic k
ernels and are therefore insensitive to the relative orientation of vertices and
 thus to the geometry of the mesh as a whole. We propose Gauge Equivariant Mesh
CNNs which generalize GCNs to apply anisotropic gauge equivariant kernels. Since
 the resulting features carry orientation information, we introduce a geometric
message passing scheme defined by parallel transporting features over mesh edges
. Our experiments validate the significantly improved expressivity of the propos
ed model over conventional GCNs and other methods.
****
Semi-Relaxed Quantization with DropBits: Training Low-Bit Neural Networks via Bi
twise Regularization
Jung Hyun Lee,Jihun Yun,Sung Ju Hwang,Eunho Yang
Network quantization, which aims to reduce the bit-lengths of the network weight
s and activations, has emerged as one of the key ingredients to reduce the size
of neural networks for their deployments to resource-limited devices. In order t
o overcome the nature of transforming continuous activations and weights to disc
rete ones, recent study called Relaxed Quantization (RQ) [Louizos et al. 2019] s
uccessfully employ the popular Gumbel-Softmax that allows this transformation wi
th efficient gradient-based optimization. However, RQ with this Gumbel-Softmax r
elaxation still suffers from bias-variance trade-off depending on the temperatur
e parameter of Gumbel-Softmax. To resolve the issue, we propose a novel method,
Semi-Relaxed Quantization (SRQ) that uses multi-class straight-through estimator

to effectively reduce the bias and variance, along with a new regularization technique, DropBits that replaces dropout regularization to randomly drop the bits instead of neurons to further reduce the bias of the multi-class straight-through estimator in SRQ. As a natural extension of DropBits, we further introduce the way of learning heterogeneous quantization levels to find proper bit-length for each layer using DropBits. We experimentally validate our method on various benchmark datasets and network architectures, and also support the quantized lottery ticket hypothesis: learning heterogeneous quantization levels outperforms the case using the same but fixed quantization levels from scratch.
****
Generative Fairness Teaching
Rongmei Lin,Hanjun Dai,Li Xiong,Wei Wei
Increasing evidences has shown that data biases towards sensitive features such as gender or race are often inherited or even amplified by machine learning models. Recent advancements in fairness mitigate such biases by adjusting the predictions across sensitive groups during the training. Such a correction, however, can only take advantage of samples in a fixed dataset, which usually has limited amount of samples for the minority groups. We propose a generative fairness teaching framework that provides a model with not only real samples but also synthesized samples to compensate the data biases during training. We employ such a teaching strategy by implementing a Generative Fairness Teacher (GFT) that dynamically adjust the proportion of training data for a biased student model. Experimental results indicated that our teacher model is capable of guiding a wide range of biased models by improving the fairness and performance trade-offs significantly.
****
Least Probable Disagreement Region for Active Learning
Seong Jin Cho,Gwangsu Kim,Chang D. Yoo
Active learning strategy to query unlabeled samples nearer the estimated decision boundary at each step has been known to be effective when the distance from the sample data to the decision boundary can be explicitly evaluated; however, in numerous cases in machine learning, especially when it involves deep learning, conventional distance such as the ℓ_p from sample to decision boundary is not readily measurable. This paper defines a theoretical distance of unlabeled sample to the decision boundary as the least probable disagreement region (LPDR) containing the unlabeled sample, and it discusses how this theoretical distance can be empirically evaluated with a lower order of time complexity. Monte Carlo sampling of the hypothesis is performed in approximating the theoretically defined distance. Experimental results on various datasets show that the proposed algorithm consistently outperforms all other high performing uncertainty based active learning algorithms and leads to state-of-the-art active learning performance on CIFAR10, CIFAR100, Tiny ImageNet and Food101 datasets. Only the proposed algorithm outperforms random sampling on CIFAR100 dataset using K-CNN while all other algorithms fail to do so.
****
Differentiable Segmentation of Sequences
Erik Scharwächter,Jonathan Lennartz,Emmanuel Müller
Segmented models are widely used to describe non-stationary sequential data with discrete change points. Their estimation usually requires solving a mixed discrete-continuous optimization problem, where the segmentation is the discrete part and all other model parameters are continuous. A number of estimation algorithms have been developed that are highly specialized for their specific model assumptions. The dependence on non-standard algorithms makes it hard to integrate segmented models in state-of-the-art deep learning architectures that critically depend on gradient-based optimization techniques. In this work, we formulate a relaxed variant of segmented models that enables joint estimation of all model parameters, including the segmentation, with gradient descent. We build on recent advances in learning continuous warping functions and propose a novel family of warping functions based on the two-sided power (TSP) distribution. TSP-based warping functions are differentiable, have simple closed-form expressions, and can re

present segmentation functions exactly. Our formulation includes the important c lass of segmented generalized linear models as a special case, which makes it hi ghly versatile. We use our approach to model the spread of COVID-19 with Poisson regression, apply it on a change point detection task, and learn classification models with concept drift. The experiments show that our approach effectively l earns all these tasks with standard algorithms for gradient descent.
****

Symmetry Control Neural Networks
Marc Syvaeri,Sven Krippendorf
This paper continues the quest for designing the optimal physics bias for neural networks predicting the dynamics of systems when the underlying dynamics shall be inferred from the data directly. The description of physical systems is great ly simplified when the underlying symmetries of the system are taken into accoun t. In classical systems described via Hamiltonian dynamics this is achieved by u sing appropriate coordinates, so-called cyclic coordinates, which reveal conserv ed quantities directly. Without changing the Hamiltonian, these coordinates can be obtained via canonical transformations. We show that such coordinates can be searched for automatically with appropriate loss functions which naturally arise from Hamiltonian dynamics. As a proof of principle, we test our method on stand ard classical physics systems using synthetic and experimental data where our ne twork identifies the conserved quantities in an unsupervised way and find improv ed performance on predicting the dynamics of the system compared to networks bia sing just to the Hamiltonian. Effectively, these new coordinates guarantee that motion takes place on symmetry orbits in phase space, i.e.~appropriate lower dim ensional sub-spaces of phase space. By fitting analytic formulae we recover that our networks are utilising conserved quantities such as (angular) momentum.
****

Generalization bounds via distillation
Daniel Hsu,Ziwei Ji,Matus Telgarsky,Lan Wang
This paper theoretically investigates the following empirical phenomenon: given a high-complexity network with poor generalization bounds, one can distill it in to a network with nearly identical predictions but low complexity and vastly sma ller generalization bounds. The main contribution is an analysis showing that t he original network inherits this good generalization bound from its distillatio n, assuming the use of well-behaved data augmentation. This bound is presented both in an abstract and in a concrete form, the latter complemented by a reducti on technique to handle modern computation graphs featuring convolutional layers, fully-connected layers, and skip connections, to name a few. To round out the story, a (looser) classical uniform convergence analysis of compression is also presented, as well as a variety of experiments on cifar and mnist demonstrating similar generalization performance between the original network and its distilla tion.

****

Learning Mesh-Based Simulation with Graph Networks
Tobias Pfaff,Meire Fortunato,Alvaro Sanchez-Gonzalez,Peter Battaglia
Mesh-based simulations are central to modeling complex physical systems in many disciplines across science and engineering. Mesh representations support powerfu l numerical integration methods and their resolution can be adapted to strike fa vorable trade-offs between accuracy and efficiency. However, high-dimensional sc ientific simulations are very expensive to run, and solvers and parameters must often be tuned individually to each system studied.
Here we introduce MeshGraphNets, a framework for learning mesh-based simulations using graph neural networks. Our model can be trained to pass messages on a mes h graph and to adapt the mesh discretization during forward simulation. Our resu lts show it can accurately predict the dynamics of a wide range of physical syst ems, including aerodynamics, structural mechanics, and cloth. The model's adapti vity supports learning resolution-independent dynamics and can scale to more com plex state spaces at test time. Our method is also highly efficient, running 1-2 orders of magnitude faster than the simulation on which it is trained. Our appr

oach broadens the range of problems on which neural network simulators can opera te and promises to improve the efficiency of complex, scientific modeling tasks.
****

Wat zei je? Detecting Out-of-Distribution Translations with Variational Transfor mers

Tim Z. Xiao,Aidan Gomez,Yarin Gal

We detect out-of-training-distribution sentences in Neural Machine Translation u sing the Bayesian Deep Learning equivalent of Transformer models. For this we de velop a new measure of uncertainty designed specifically for long sequences of d iscrete random variables—i.e. words in the output sentence. Our new measure of u ncertainty solves a major intractability in the naive application of existing ap proaches on long sentences. We use our new measure on a Transformer model traine d with dropout approximate inference. On the task of German-English tr anslation using WMT13 and Europarl, we show that with dropout uncertainty our me asure is able to identify when Dutch source sentences, sentences which use the s ame word types as German, are given to the model instead of German.
****

GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing

Tao Yu,Chien-Sheng Wu,Xi Victoria Lin,bailin wang,Yi Chern Tan,Xinyi Yang,Dragom ir Radev,richard socher,Caiming Xiong

We present GraPPa, an effective pre-training approach for table semantic parsing that learns a compositional inductive bias in the joint representations of text ual and tabular data. We construct synthetic question-SQL pairs over high-qualit y tables via a synchronous context-free grammar (SCFG). We pre-train our model o n the synthetic data to inject important structural properties commonly found in semantic parsing into the pre-training language model. To maintain the model's ability to represent real-world data, we also include masked language modeling (MLM) on several existing table-related datasets to regularize our pre-training p rocess. Our proposed pre-training strategy is much data-efficient. When incorpo rated with strong base semantic parsers, GraPPa achieves new state-of-the-art re sults on four popular fully supervised and weakly supervised table semantic pars ing tasks.
****

Sliced Kernelized Stein Discrepancy

Wenbo Gong,Yingzhen Li,José Miguel Hernández-Lobato

Kernelized Stein discrepancy (KSD), though being extensively used in goodness-of -fit tests and model learning, suffers from the curse-of-dimensionality. We addr ess this issue by proposing the sliced Stein discrepancy and its scalable and ke rnelized variants, which employs kernel-based test functions defined on the opti mal one-dimensional projections. When applied to goodness-of-fit tests, extensiv e experiments show the proposed discrepancy significantly outperforms KSD and va rious baselines in high dimensions. For model learning, we show its advantages b y training an independent component analysis when compared with existing Stein d iscrepancy baselines. We further propose a novel particle inference method calle d sliced Stein variational gradient descent (S-SVGD) which alleviates the mode-c ollapse issue of SVGD in training variational autoencoders.
****

Variational Intrinsic Control Revisited

Taehwan Kwon

In this paper, we revisit variational intrinsic control (VIC), an unsupervised r einforcement learning method for finding the largest set of intrinsic options av ailable to an agent. In the original work by Gregor et al. (2016), two VIC algor ithms were proposed: one that represents the options explicitly, and the other t hat does it implicitly. We show that the intrinsic reward used in the latter is subject to bias in stochastic environments, causing convergence to suboptimal so lutions. To correct this behavior, we propose two methods respectively based on the transitional probability model and Gaussian Mixture Model. We substantiate o ur claims through rigorous mathematical derivations and experimental analyses.
****

Federated Averaging as Expectation Maximization

Christos Louizos,Matthias Reisser,Joseph Soriaga,Max Welling
Federated averaging (FedAvg), despite its simplicity, has been the main approach in training neural networks in the federated learning setting. In this work, we show that the algorithmic choices of the FedAvg algorithm correspond to optimizing a single objective function that involves the global and all of the shard specific models using a hard version of the well known Expectation-Maximization (EM) algorithm. As a result, we gain a better understanding of the behavior and design choices of federated averaging while being able to provide interesting connections to recent literature. Based on this view, we further propose FedSparse, a version of federated averaging that employs prior distributions to promote model sparsity. In this way, we obtain a procedure that leads to reductions in both server-client and client-server communication costs as well as more efficient models.
****

On Statistical Bias In Active Learning: How and When to Fix It
Sebastian Farquhar,Yarin Gal,Tom Rainforth
Active learning is a powerful tool when labelling data is expensive, but it introduces a bias because the training data no longer follows the population distribution. We formalize this bias and investigate the situations in which it can be harmful and sometimes even helpful. We further introduce novel corrective weights to remove bias when doing so is beneficial. Through this, our work not only provides a useful mechanism that can improve the active learning approach, but also an explanation for the empirical successes of various existing approaches which ignore this bias. In particular, we show that this bias can be actively helpful when training overparameterized models---like neural networks---with relatively modest dataset sizes.
****

Differentiable Approximations for Multi-resource Spatial Coverage Problems
Nitin Kamra,Yan Liu
Resource allocation for coverage of physical spaces is a challenging problem in robotic surveillance, mobile sensor networks and security domains. Recent gradient-based optimization approaches to this problem estimate utilities of actions by using neural networks to learn a differentiable approximation to spatial coverage objectives. In this work, we empirically show that spatial coverage objectives with multiple-resources are combinatorially hard to approximate for neural networks and lead to sub-optimal policies. As our major contribution, we propose a tractable framework to approximate a general class of spatial coverage objectives and their gradients using a combination of Newton-Leibniz theorem, spatial discretization and implicit boundary differentiation. We empirically demonstrate the efficacy of our proposed framework on single and multi-agent spatial coverage problems.
****

Winning the L2RPN Challenge: Power Grid Management via Semi-Markov Afterstate Actor-Critic
Deunsol Yoon,Sunghoon Hong,Byung-Jun Lee,Kee-Eung Kim
Safe and reliable electricity transmission in power grids is crucial for modern society. It is thus quite natural that there has been a growing interest in the automatic management of power grids, exempli■ed by the Learning to Run a Power Network Challenge (L2RPN), modeling the problem as a reinforcement learning (RL) task. However, it is highly challenging to manage a real-world scale power grid, mostly due to the massive scale of its state and action space. In this paper, we present an off-policy actor-critic approach that effectively tackles the unique challenges in power grid management by RL, adopting the hierarchical policy together with the afterstate representation. Our agent ranked ■rst in the latest challenge (L2RPN WCCI 2020), being able to avoid disastrous situations while maintaining the highest level of operational ef■ciency in every test scenarios. This paper provides a formal description of the algorithmic aspect of our approach, as well as further experimental studies on diverse power grids.
****

What's in the Box? Exploring the Inner Life of Neural Networks with Robust Rules

Jonas Fischer,Anna Oláh,Jilles Vreeken
We propose a novel method for exploring how neurons within a neural network interact. In particular, we consider activation values of a network for given data, and propose to mine noise-robust rules of the form $X \rightarrow Y$, where X and Y are sets of neurons in different layers. To ensure we obtain a small and non-redundant set of high quality rules, we formalize the problem in terms of the Minimum Description Length principle, by which we identify the best set of rules as the one that best compresses the activation data. To discover good rule sets, we propose the unsupervised ExplaiNN algorithm. Extensive evaluation shows that our rules give clear insight in how networks perceive the world: they identify shared, resp. class-specific traits, compositionality within the network, as well as locality in convolutional layers. Our rules are easily interpretable, but also super-charge prototyping as they identify which groups of neurons to consider in unison.
****

HyperDynamics: Meta-Learning Object and Agent Dynamics with Hypernetworks
Zhou Xian,Shamit Lal,Hsiao-Yu Tung,Emmanouil Antonios Platanios,Katerina Fragkiadaki
We propose HyperDynamics, a dynamics meta-learning framework that conditions on an agent’s interactions with the environment and optionally its visual observations, and generates the parameters of neural dynamics models based on inferred properties of the dynamical system. Physical and visual properties of the environment that are not part of the low-dimensional state yet affect its temporal dynamics are inferred from the interaction history and visual observations, and are implicitly captured in the generated parameters. We test HyperDynamics on a set of object pushing and locomotion tasks. It outperforms existing dynamics models in the literature that adapt to environment variations by learning dynamics over high dimensional visual observations, capturing the interactions of the agent in recurrent state representations, or using gradient-based meta-optimization. We also show our method matches the performance of an ensemble of separately trained experts, while also being able to generalize well to unseen environment variations at test time. We attribute its good performance to the multiplicative interactions between the inferred system properties—captured in the generated parameters—and the low-dimensional state representation of the dynamical system.
****

Towards Resolving the Implicit Bias of Gradient Descent for Matrix Factorization: Greedy Low-Rank Learning
Zhiyuan Li,Yuping Luo,Kaifeng Lyu
Matrix factorization is a simple and natural test-bed to investigate the implicit regularization of gradient descent. Gunasekar et al. (2017) conjectured that gradient flow with infinitesimal initialization converges to the solution that minimizes the nuclear norm, but a series of recent papers argued that the language of norm minimization is not sufficient to give a full characterization for the implicit regularization. In this work, we provide theoretical and empirical evidence that for depth-2 matrix factorization, gradient flow with infinitesimal initialization is mathematically equivalent to a simple heuristic rank minimization algorithm, Greedy Low-Rank Learning, under some reasonable assumptions. This generalizes the rank minimization view from previous works to a much broader setting and enables us to construct counter-examples to refute the conjecture from Gunasekar et al. (2017). We also extend the results to the case where depth >= 3, and we show that the benefit of being deeper is that the above convergence has a much weaker dependence over initialization magnitude so that this rank minimization is more likely to take effect for initialization with practical scale.
****

Probabilistic Meta-Learning for Bayesian Optimization
Felix Berkenkamp,Anna Eivazi,Lukas Grossberger,Kathrin Skubch,Jonathan Spitz,Christian Daniel,Stefan Falkner
Transfer and meta-learning algorithms leverage evaluations on related tasks in order to significantly speed up learning or optimization on a new problem. For applications that depend on uncertainty estimates, e.g., in Bayesian optimization,

recent probabilistic approaches have shown good performance at test time, but either scale poorly with the number of data points or under-perform with little data on the test task. In this paper, we propose a novel approach to probabilistic transfer learning that uses a generative model for the underlying data distribution and simultaneously learns a latent feature distribution to represent unknown task properties. To enable fast and accurate inference at test-time, we introduce a novel meta-loss that structures the latent space to match the prior used for inference. Together, these contributions ensure that our probabilistic model exhibits high sample-efficiency and provides well-calibrated uncertainty estimates. We evaluate the proposed approach and compare its performance to probabilistic models from the literature on a set of Bayesian optimization transfer-learning tasks.

****

Multimodal Variational Autoencoders for Semi-Supervised Learning: In Defense of Product-of-Experts

Svetlana Kutuzova,Oswin Krause,Douglas McCloskey,Mads Nielsen,Christian Igel

Multimodal generative models should be able to learn a meaningful latent representation that enables a coherent joint generation of all modalities (e.g., images and text). Many applications also require the ability to accurately sample modalities conditioned on observations of a subset of the modalities. Often not all modalities may be observed for all training data points, so semi-supervised learning should be possible.

In this study, we evaluate a family of product-of-experts (PoE) based variational autoencoders that have these desired properties. We include a novel PoE based architecture and training procedure. An empirical evaluation shows that the PoE based models can outperform an additive mixture-of-experts (MoE) approach.

Our experiments support the intuition that PoE models are more suited for a conjunctive combination of modalities while MoEs are more suited for a disjunctive fusion.

****

Private Post-GAN Boosting

Marcel Neunhoeffer,Steven Wu,Cynthia Dwork

Differentially private GANs have proven to be a promising approach for generating realistic synthetic data without compromising the privacy of individuals. Due to the privacy-protective noise introduced in the training, the convergence of GANs becomes even more elusive, which often leads to poor utility in the output generator at the end of training. We propose Private post-GAN boosting (Private PGB), a differentially private method that combines samples produced by the sequence of generators obtained during GAN training to create a high-quality synthetic dataset. To that end, our method leverages the Private Multiplicative Weights method (Hardt and Rothblum, 2010) to reweight generated samples. We evaluate Private PGB on two dimensional toy data, MNIST images, US Census data and a standard machine learning prediction task. Our experiments show that Private PGB improves upon a standard private GAN approach across a collection of quality measures. We also provide a non-private variant of PGB that improves the data quality of standard GAN training.

****

Wide-minima Density Hypothesis and the Explore-Exploit Learning Rate Schedule

Nikhil Iyer,V Thejas,Nipun Kwatra,Ramachandran Ramjee,Muthian Sivathanu

Several papers argue that wide minima generalize better than narrow minima. In this paper, through detailed experiments that not only corroborate the generalization properties of wide minima, we also provide empirical evidence for a new hypothesis that the density of wide minima is likely lower than the density of narrow minima. Further, motivated by this hypothesis, we design a novel explore-exploit learning rate schedule. On a variety of image and natural language datasets, compared to their original hand-tuned learning rate baselines, we show that our explore-exploit schedule can result in either up to 0.84% higher absolute accuracy using the original training budget or up to 57% reduced training time while achieving the original reported accuracy. For example, we achieve state-of-the-art (SOTA) accuracy for IWSLT'14 (DE-EN) and WMT'14 (DE-EN) datasets by just modi

fying the learning rate schedule of a high performing model.
****
Learning from deep model via exploring local targets
Wenxian Shi,Yuxuan Song,Hao Zhou,Bohan Li,Lei Li
Deep neural networks often have huge number of parameters, which posts challenges in deployment in application scenarios with limited memory and computation capacity. Knowledge distillation is one approach to derive compact models from bigger ones.
However, it has been observed that a converged heavy teacher model is strongly constrained for learning a compact student network and could make the optimization subject to poor local optima. In this paper, we propose proKT, a new model-agnostic method by projecting the supervision signals of a teacher model into the student's parameter space. Such projection is implemented by decomposing the training objective into local intermediate targets with approximate mirror descent technique. The proposed method could be less sensitive with the quirks during optimization which could result in a better local optima. Experiments on both image and text datasets show that our proposed proKT consistently achieves the state-of-the-art performance comparing to all existing knowledge distillation methods.
****
Characterizing signal propagation to close the performance gap in unnormalized ResNets
Andrew Brock,Soham De,Samuel L Smith
Batch Normalization is a key component in almost all state-of-the-art image classifiers, but it also introduces practical challenges: it breaks the independence between training examples within a batch, can incur compute and memory overhead, and often results in unexpected bugs. Building on recent theoretical analyses of deep ResNets at initialization, we propose a simple set of analysis tools to characterize signal propagation on the forward pass, and leverage these tools to design highly performant ResNets without activation normalization layers. Crucial to our success is an adapted version of the recently proposed Weight Standardization. Our analysis tools show how this technique preserves the signal in ReLU networks by ensuring that the per-channel activation means do not grow with depth. Across a range of FLOP budgets, our networks attain performance competitive with state-of-the-art EfficientNets on ImageNet.
****
Local Clustering Graph Neural Networks
Jiezhong Qiu,Yukuo Cen,Qibin Chen,Chang Zhou,Jingren Zhou,Hongxia Yang,Jie Tang
Graph Neural Networks (GNNs), which benefit various real-world problems and applications, have emerged as a powerful technique for learning graph representations. The depth of a GNN model, denoted by K, restricts the receptive field of a node to its K-hop neighbors and plays a subtle role in the performance of GNNs. Recent works demonstrate how different choices of K produce a trade-off between increasing representation capacity and avoiding over-smoothing. We establish a theoretical connection between GNNs and local clustering, showing that short random-walks in GNNs have a high probability to be stuck at a local cluster. Based on the theoretical analysis, we propose Local Clustering Graph Neural Networks (LCGNN), a GNN learning paradigm that utilizes local clustering to efficiently search for small but compact subgraphs for GNN training and inference. Compared to full-batch GNNs, sampling-based GNNs and graph partition-based GNNs, LCGNN performs comparably or even better, achieving state-of-the-art results on four Open Graph Benchmark (OGB) datasets. The locality of LCGNN allows it to scale to graphs with 100M nodes and 1B edges on a single GPU.
****
Prototypical Contrastive Learning of Unsupervised Representations
Junnan Li,Pan Zhou,Caiming Xiong,Steven Hoi
This paper presents Prototypical Contrastive Learning (PCL), an unsupervised representation learning method that bridges contrastive learning with clustering. PCL not only learns low-level features for the task of instance discrimination, but more importantly, it implicitly encodes semantic structures of the data into

the learned embedding space. Specifically, we introduce prototypes as latent variables to help find the maximum-likelihood estimation of the network parameters in an Expectation-Maximization framework. We iteratively perform E-step as finding the distribution of prototypes via clustering and M-step as optimizing the network via contrastive learning. We propose ProtoNCE loss, a generalized version of the InfoNCE loss for contrastive learning, which encourages representations to be closer to their assigned prototypes. PCL outperforms state-of-the-art instance-wise contrastive learning methods on multiple benchmarks with substantial improvement in low-resource transfer learning. Code and pretrained models are available at https://github.com/salesforce/PCL.
****

Truly Deterministic Policy Optimization
Ehsan Saleh,Saba Ghaffari,Matthew West,Tim Bretl
In this paper, we present a policy gradient method that avoids exploratory noise injection and performs policy search over the deterministic landscape. By avoiding noise injection all sources of estimation variance can be eliminated in systems with deterministic dynamics (up to the initial state distribution). Since deterministic policy regularization is impossible using traditional non-metric measures such as the KL divergence, we derive a Wasserstein-based quadratic model for our purposes. We state conditions on the system model under which it is possible to establish a monotonic policy improvement guarantee, propose a surrogate function for policy gradient estimation, and show that it is possible to compute exact advantage estimates if both the state transition model and the policy are deterministic. Finally, we describe two novel robotic control environments---one with non-local rewards in the frequency domain and the other with a long horizon (8000 time-steps)---for which our policy gradient method (TDPO) significantly outperforms existing methods (PPO, TRPO, DDPG, and TD3).
****

Hyperbolic Neural Networks++
Ryohei Shimizu,YUSUKE Mukuta,Tatsuya Harada
Hyperbolic spaces, which have the capacity to embed tree structures without distortion owing to their exponential volume growth, have recently been applied to machine learning to better capture the hierarchical nature of data. In this study, we generalize the fundamental components of neural networks in a single hyperbolic geometry model, namely, the Poincaré ball model. This novel methodology constructs a multinomial logistic regression, fully-connected layers, convolutional layers, and attention mechanisms under a unified mathematical interpretation, without increasing the parameters. Experiments show the superior parameter efficiency of our methods compared to conventional hyperbolic components, and stability and outperformance over their Euclidean counterparts.
****

Continual Invariant Risk Minimization
Francesco Alesiani,Shujian Yu,Mathias Niepert
Empirical risk minimization can lead to poor generalization behaviour on unseen environments if the learned model does not capture invariant feature represen- tations. Invariant risk minimization (IRM) is a recent proposal for discovering environment-invariant representations. It was introduced by Arjovsky et al. (2019) and extended by Ahuja et al. (2020). The assumption of IRM is that all environ - ments are available to the learning system at the same time. With this work, we generalize the concept of IRM to scenarios where environments are observed se- quentially. We show that existing approaches, including those designed for contin- ual learning, fail to identify the invariant features and models across sequentially presented environments. We extend IRM under a variational Bayesian and bilevel framework, creating a general approach to continual invariant risk minimization. We also describe a strategy to solve the optimization problems using a variant of the alternating direction method of multiplier (ADMM). We show empirically us- ing multiple datasets and with multiple sequential environments that the proposed methods outperforms or is competitive with prior approaches.
****

Early Stopping by Gradient Disparity

mahsa forouzesh,Patrick Thiran
Validation-based early-stopping methods are one of the most popular techniques u sed to avoid over-training deep neural networks. They require to set aside a rel iable unbiased validation set, which can be expensive in applications offering l imited amounts of data. In this paper, we propose to use \emph{gradient disparit y}, which we define as the ℓ_2 norm distance between the gradient vectors o f two batches drawn from the training set. It comes from a probabilistic upper b ound on the difference between the classification errors over a given batch, whe n the network is trained on this batch and when the network is trained on anothe r batch of points sampled from the same dataset. We empirically show that gradie nt disparity is a very promising early-stopping criterion when data is limited, because it uses all the training samples during training. Furthermore, we show i n a wide range of experimental settings that gradient disparity is not only stro ngly related to the usual generalization error between the training and test set s, but that it is also much more informative about the level of label noise.
****

Sparse Binary Neural Networks
Riccardo Schiavone,Maria A Zuluaga
Quantized neural networks are gaining popularity thanks to their ability to solv e complex tasks with comparable accuracy as full-precision Deep Neural Networks (DNNs), while also reducing computational power and storage requirements and inc reasing the processing speed. These properties make them an attractive alternati ve for the development and deployment of DNN-based applications in Internet-Of-T hings (IoT) devices. Among quantized networks, Binary Neural Networks (BNNs) hav e reported the largest speed-up. However, they suffer from a fixed and limited c ompression factor that may result insufficient for certain devices with very lim ited resources. In this work, we propose Sparse Binary Neural Networks, a novel model and training scheme that allows to introduce sparsity in BNNs by using pos itive 0/1 binary weights, instead of the -1/+1 weights used by state-of-the-art binary networks. As a result, our method is able to achieve a high compression f actor and reduces the number of operations and parameters at inference time. We study the properties of our method through experiments on linear and convolution al networks over MNIST and CIFAR-10 datasets. Experiments confirm that SBNNs can achieve high compression rates and good generalization, while further reducing the operations of BNNs, making it a viable option for deploying DNNs in very che ap and low-cost IoT devices and sensors.
****

Accurate and fast detection of copy number variations from short-read whole-geno me sequencing with deep convolutional neural network
Jiajin Li,Stephen Hwang,Luke Zhang,Jae Hoon Sul
A copy number variant (CNV) is a type of genetic mutation where a stretch of DNA is lost or duplicated once or multiple times. CNVs play important roles in the development of diseases and complex traits. CNV detection with short-read DNA se quencing technology is challenging because CNVs significantly vary in size and a re similar to DNA sequencing artifacts. Many methods have been developed but sti ll yield unsatisfactory results with high computational costs. Here, we propose CNV-Net, a novel approach for CNV detection using a six-layer convolutional neur al network. We encode DNA sequencing information into RGB images and train the c onvolutional neural network with these images. The fitted convolutional neural n etwork can then be used to predict CNVs from DNA sequencing data. We benchmark C NV-Net with two high-quality whole-genome sequencing datasets available from the Genome in a Bottle Consortium, considered as gold standard benchmarking dataset s for CNV detection. We demonstrate that CNV-Net is more accurate and efficient in CNV detection than current tools.
****

Distantly supervised end-to-end medical entity extraction from electronic health records with human-level quality
Alexander Nesterov,Dmitry Umerenkov
Medical entity extraction (EE) is a standard procedure used as a firs t stage inmedical texts processing. Usually Medical EE is a two-step process:

named entityrecognition (NER) and named entity normalization (NEN). We propose a novelmethod of doing medical EE from electronic health records (EHR) as a single-step multi-label classification task by fine-tuning a transformer model pretrainedon a large EHR dataset. Our model is trained end-to-end in an distantly supervisedmanner using targets automatically extracted from medical knowledge base. Weshow that our model learns to generalize for entities that are present frequentlyenough, achieving human-level classification quality for most frequent entities.Our work demonstrates that medical entity extraction can be done end-to-endwithout human supervision and with human quality given the availability of alarge enough amount of unlabeled EHR and a medical knowledge base.
****
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.
A step in the Q-function indicates solving a sub-task, where the expectation of the return increases.
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.
Typically the number of demonstrations is small and RUDDER's LSTM model does not learn well.
Hence, 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 inherits the concept of reward redistribution, which speeds up learning by reducing the delay of rewards.
Align-RUDDER outperforms competitors on complex artificial tasks with delayed reward and few demonstrations.
On the MineCraft ObtainDiamond task, Align-RUDDER is able to mine a diamond, though not frequently.
****
NAS-Bench-301 and the Case for Surrogate Benchmarks for Neural Architecture Search
Julien Niklas Siems,Lucas Zimmer,Arber Zela,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. As a remedy, several tabular NAS benchmarks were proposed to simulate runs of NAS methods in seconds. However, all existing tabular NAS benchmarks are limited to extremely small architectural spaces since they rely on exhaustive evaluations of the space. This leads to unrealistic results that do not transfer to larger search spaces. To overcome this fundamental limitation, we propose NAS-Bench-301, the first surrogate NAS benchmark, using a search space containing 10^{18} architectures, many orders of magnitude larger than any previous tabular NAS benchmark. After motivating the benefits of a surrogate benchmark over a tabular one, we fit various regression models on our dataset, which consists of $\sim$60k architecture evaluations, and build surrogates via deep ensembles to model uncertainty. We benchmark a wide range of NAS algorith

ms using NAS-Bench-301 and obtain comparable results to the true benchmark at a fraction of the real cost. Finally, we show how NAS-Bench-301 can be used to generate new scientific insights.
****
Lipschitz Recurrent Neural Networks
N. Benjamin Erichson,Omri Azencot,Alejandro Queiruga,Liam Hodgkinson,Michael W. Mahoney
Viewing recurrent neural networks (RNNs) as continuous-time dynamical systems, we propose a recurrent unit that describes the hidden state's evolution with two parts: a well-understood linear component plus a Lipschitz nonlinearity. This particular functional form facilitates stability analysis of the long-term behavior of the recurrent unit using tools from nonlinear systems theory. In turn, this enables architectural design decisions before experimentation. Sufficient conditions for global stability of the recurrent unit are obtained, motivating a novel scheme for constructing hidden-to-hidden matrices. Our experiments demonstrate that the Lipschitz RNN can outperform existing recurrent units on a range of benchmark tasks, including computer vision, language modeling and speech prediction tasks. Finally, through Hessian-based analysis we demonstrate that our Lipschitz recurrent unit is more robust with respect to input and parameter perturbations as compared to other continuous-time RNNs.
****
Meta Auxiliary Labels with Constituent-based Transformer for Aspect-based Sentiment Analysis
Ling Min Serena Khoo,Hai Leong Chieu
Aspect based sentiment analysis (ABSA) is a challenging natural language processing task that could benefit from syntactic information. Previous work exploit dependency parses to improve performance on the task, but this requires the existence of good dependency parsers. In this paper, we build a constituent-based transformer for ABSA that can induce constituents without constituent parsers. We also apply meta auxiliary learning to generate labels on edges between tokens, supervised by the objective of the ABSA task. Without input from dependency parsers, our models outperform previous work on three Twitter data sets and match previous work closely on two review data sets.
****
Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification
Yunsheng Shi,Zhengjie Huang,shikun feng,Hui Zhong,Wenjin Wang,Yu Sun
Graph neural network (GNN) and label propagation algorithm (LPA) are both message passing algorithms, which have achieved superior performance in semi-supervised classification. GNN performs \emph{feature propagation} by a neural network to make predictions, while LPA uses \emph{label propagation} across graph adjacency matrix to get results. However, there is still no good way to combine these two kinds of algorithms. In this paper, we proposed a new {\bf Uni}fied {\bf M}essage {\bf P}assaging Model (UniMP) that can incorporate \emph{feature propagation} and \emph{label propagation} with a shared message passing network, providing a better performance in semi-supervised classification. First, we adopt a Graph Transformer jointly label embedding to propagate both the feature and label information. Second, to train UniMP without overfitting in self-loop label information, we propose a masked label prediction strategy, in which some percentage of training labels are simply masked at random, and then predicted. UniMP conceptually unifies feature propagation and label propagation and be empirically powerful. It obtains new state-of-the-art semi-supervised classification results in Open Graph Benchmark (OGB).
****
Adversarial Deep Metric Learning
Thomas Kobber Panum,Zi Wang,Pengyu Kan,Earlence Fernandes,Somesh Jha
Learning a distance metric between pairs of examples is widely important for various tasks. Deep Metric Learning (DML) utilizes deep neural network architectures to learn semantic feature embeddings where the distance between similar examples is close and dissimilar examples are far. While the underlying neural network

s produce good accuracy on naturally occurring samples, they are vulnerable to a dversarially-perturbed samples that can reduce their accuracy. To create robust versions of DML models, we introduce a robust training approach. A key challenge is that metric losses are not independent --- they depend on all samples in a m ini-batch. This sensitivity to samples, if not accounted for, can lead to incor rect robust training. To the best of our knowledge, we are the first to systema tically analyze this dependence effect and propose a principled approach for rob ust training of deep metric learning networks that accounts for the nuances of m etric losses. Using experiments on three popular datasets in metric learning, we demonstrate the DML models trained using our techniques display robustness agai nst strong iterative attacks while their performance on unperturbed (natural) sa mples remains largely unaffected.

****

DISE: Dynamic Integrator Selection to Minimize Forward Pass Time in Neural ODEs

Soyoung Kang,Ganghyeon Park,Kwang-Sung Jun,Noseong Park

Neural ordinary differential equations (Neural ODEs) are appreciated for their a bility to significantly reduce the number of parameters when constructing a neur al network. On the other hand, they are sometimes blamed for their long forward-pass inference time, which is incurred by solving integral problems. To improve the model accuracy, they rely on advanced solvers, such as the Dormand--Prince (DOPRI) method. To solve an integral problem, however, it requires at least tens (or sometimes thousands) of steps in many Neural ODE experiments. In this work, we propose to i) directly regularize the step size of DOPRI to make the forward-pass faster and ii) dynamically choose a simpler integrator than DOPRI for a car efully selected subset of input. Because it is not the case that every input req uires the advanced integrator, we design an auxiliary neural network to choose a n appropriate integrator given input to decrease the overall inference time with out significantly sacrificing accuracy. We consider the Euler method, the fourth -order Runge--Kutta (RK4) method, and DOPRI as selection candidates. We found th at 10-30% of cases can be solved with simple integrators in our experiments. The refore, the overall number of functional evaluations (NFE) decreases up to 78% w ith improved accuracy.

****

Learning to generate Wasserstein barycenters

Julien Lacombe,Julie Digne,Nicolas Courty,Nicolas Bonneel

Optimal transport is a notoriously difficult problem to solve numerically, with current approaches often remaining intractable for very large scale applications such as those encountered in machine learning. Wasserstein barycenters -- the p roblem of finding measures in-between given input measures in the optimal transp ort sense -- is even more computationally demanding.

By training a deep convolutional neural network, we improve by a factor of 60 th e computational speed of Wasserstein barycenters over the fastest state-of-the-a rt approach on the GPU, resulting in milliseconds computational times on $512\ti mes512$ regular grids.

We show that our network, trained on Wasserstein barycenters of pairs of measure s, generalizes well to the problem of finding Wasserstein barycenters of more th an two measures. We validate our approach on synthetic shapes generated via Cons tructive Solid Geometry as well as on the ``Quick, Draw'' sketches dataset.

****

Calibrated Adversarial Refinement for Stochastic Semantic Segmentation

Elias Kassapis,Georgi Dikov,Deepak Gupta,Cedric Nugteren

Ambiguities in images or unsystematic annotation can lead to multiple valid solu tions in semantic segmentation. To learn a distribution over predictions, recent work has explored the use of probabilistic networks. However, these do not nece ssarily capture the empirical distribution accurately. In this work, we aim to l earn a calibrated multimodal predictive distribution, where the empirical freque ncy of the sampled predictions closely reflects that of the corresponding labels in the training set. To this end, we propose a novel two-stage, cascaded strate gy for calibrated adversarial refinement. In the first stage, we explicitly mode l the data with a categorical likelihood. In the second, we train an adversarial

network to sample from it an arbitrary number of coherent predictions. The model can be used independently or integrated into any black-box segmentation framework to facilitate learning of calibrated stochastic mappings. We demonstrate the utility and versatility of the approach by attaining state-of-the-art results on the multigrader LIDC dataset and a modified Cityscapes dataset. In addition, we use a toy regression dataset to show that our framework is not confined to semantic segmentation, and the core design can be adapted to other tasks requiring learning a calibrated predictive distribution.
****
Guided Exploration with Proximal Policy Optimization using a Single Demonstration
Gabriele Libardi,Gianni De Fabritiis
Solving sparse reward tasks through exploration is one of the major challenges in deep reinforcement learning, especially in three-dimensional, partially-observable environments. Critically, the algorithm proposed in this article uses a single human demonstration to solve hard-exploration problems. We train an agent on a combination of demonstrations and own experience to solve problems with variable initial conditions. We adapt this idea and integrate it with the proximal policy optimization (PPO). The agent is able to increase its performance and to tackle harder problems by replaying its own past trajectories prioritizing them based on the obtained reward and the maximum value of the trajectory.
We compare variations of this algorithm to different imitation learning algorithms on a set of hard-exploration tasks in the Animal-AI Olympics environment.
To the best of our knowledge, learning a task in a three-dimensional environment with comparable difficulty has never been considered before using only one human demonstration.
****
Efficient Reinforcement Learning in Resource Allocation Problems Through Permutation Invariant Multi-task Learning
Desmond Cai,Shiau Hong Lim,Laura Wynter
One of the main challenges in real-world reinforcement learning is to learn successfully from limited training samples. We show that in certain settings, the available data can be dramatically increased through a form of multi-task learning, by exploiting an invariance property in the tasks. We provide a theoretical performance bound for the gain in sample efficiency under this setting. This motivates a new approach to multi-task learning, which involves the design of an appropriate neural network architecture and a prioritized task-sampling strategy. We demonstrate empirically the effectiveness of the proposed approach on two real-world sequential resource allocation tasks where this invariance property occurs: financial portfolio optimization and meta federated learning.
****
Pea-KD: Parameter-efficient and accurate Knowledge Distillation
IKHYUN CHO,U Kang
How can we efficiently compress a model while maintaining its performance? Knowledge Distillation (KD) is one of the widely known methods for model compression. In essence, KD trains a smaller student model based on a larger teacher model and tries to retain the teacher model's level of performance as much as possible. However, the existing KD methods suffer from the following limitations. First, since the student model is small in absolute size, it inherently lacks model complexity. Second, the absence of an initial guide for the student model makes it difficult for the student to imitate the teacher model to its fullest. Conventional KD methods yield low performance due to these limitations.

In this paper, we propose Pea-KD (Parameter-efficient and accurate Knowledge Distillation), a novel approach to KD. Pea-KD consists of two main parts: Shuffled Parameter Sharing (SPS) and Pretraining with Teacher's Predictions (PTP). Using this combination, we are capable of alleviating the KD's limitations. SPS is a new parameter sharing method that allows greater model complexity for the student model. PTP is a KD-specialized initialization method, which can act as a good initial guide for the student. When combined, this method yields a significant in

crease in student model's performance. Experiments conducted on different datase
ts and tasks show that the proposed approach improves the student model's perfor
mance by 4.4% on average in four GLUE tasks, outperforming existing KD baselines
 by significant margins.

****
A statistical theory of cold posteriors in deep neural networks
Laurence Aitchison
To get Bayesian neural networks to perform comparably to standard neural network
s it is usually necessary to artificially reduce uncertainty using a tempered or
 cold posterior. This is extremely concerning: if the prior is accurate, Bayes i
nference/decision theory is optimal, and any artificial changes to the posterior
 should harm performance. While this suggests that the prior may be at fault, he
re we argue that in fact, BNNs for image classification use the wrong likelihood
. In particular, standard image benchmark datasets such as CIFAR-10 are carefull
y curated. We develop a generative model describing curation which gives a princ
ipled Bayesian account of cold posteriors, because the likelihood under this new
 generative model closely matches the tempered likelihoods used in past work.
****
An Empirical Study of the Expressiveness of Graph Kernels and Graph Neural Netwo
rks
Giannis Nikolentzos,George Panagopoulos,Michalis Vazirgiannis
Graph neural networks and graph kernels have achieved great success in solving m
achine learning problems on graphs. Recently, there has been considerable intere
st in determining the expressive power mainly of graph neural networks and of gr
aph kernels, to a lesser extent. Most studies have focused on the ability of th
ese approaches to distinguish non-isomorphic graphs or to identify specific grap
h properties. However, there is often a need for algorithms whose produced graph
 representations can accurately capture similarity/distance of graphs. This pape
r studies the expressive power of graph neural networks and graph kernels from a
n empirical perspective. Specifically, we compare the graph representations and
similarities produced by these algorithms against those generated by a well-acce
pted, but intractable graph similarity function. We also investigate the impact
of node attributes on the performance of the different models and kernels. Our r
esults reveal interesting findings. For instance, we find that theoretically mor
e powerful models do not necessarily yield higher-quality representations, while
 graph kernels are shown to be very competitive with graph neural networks.
****
Neural Architecture Search without Training
Joseph Mellor,Jack Turner,Amos Storkey,Elliot J. Crowley
The time and effort involved in hand-designing deep neural networks is immense.
This has prompted the development of Neural Architecture Search (NAS) techniques
 to automate this design. However, NAS algorithms tend to be slow and expensive;
 they need to train vast numbers of candidate networks to inform the search proc
ess. This could be remedied if we could infer a network's trained accuracy from
its initial state. In this work, we examine the correlation of linear maps induc
ed by augmented versions of a single image in untrained networks and motivate ho
w this can be used to give a measure which is highly indicative of a network’s t
rained performance. We incorporate this measure into a simple algorithm that all
ows us to search for powerful networks without any training in a matter of secon
ds on a single GPU, and verify its effectiveness on NAS-Bench-101 and NAS-Bench-
201. Finally, we show that our approach can be readily combined with more expens
ive search methods for added value: we modify regularised evolutionary search to
 produce a novel algorithm that outperforms its predecessor.
****
Distributional Reinforcement Learning for Risk-Sensitive Policies
Shiau Hong Lim,Ilyas Malik
We address the problem of learning a risk-sensitive policy based on the CVaR ris
k measure using distributional reinforcement learning. In particular, we show th
at applying the distributional Bellman optimality operator with respect to a ris

k-based action-selection strategy overestimates the dynamic, Markovian CVaR. The resulting policies can however still be overly conservative and one often prefers to learn an optimal policy based on the static, non-Markovian CVaR. To this end, we propose a modification to the existing algorithm and show that it can indeed learn a proper CVaR-optimized policy. Our proposed approach is a simple extension of standard distributional RL algorithms and can therefore take advantage of many of the recent advances in deep RL. On both synthetic and real data, we empirically show that our proposed algorithm is able to produce a family of risk-averse policies that achieves a better tradeoff between risk and the expected return.

****
Generalized Universal Approximation for Certified Networks
Zi Wang,Aws Albarghouthi,Somesh Jha
To certify safety and robustness of neural networks, researchers have successfully applied abstract interpretation, primarily using interval bound propagation. To understand the power of interval bounds, we present the abstract universal approximation (AUA) theorem, a generalization of the recent result by Baader et al. (2020) for ReLU networks to a large class of neural networks. The AUA theorem states that for any continuous function f, there exists a neural network that (1) approximates f (universal approximation) and (2) whose interval bounds are an arbitrarily close approximation of the set semantics of f. The network may be constructed using any activation function from a rich class of functions---sigmoid, tanh, ReLU, ELU, etc.---making our result quite general. The key implication of the AUA theorem is that there always exists certifiably robust neural networks, which can be constructed using a wide range of activation functions.
****
Boost then Convolve: Gradient Boosting Meets Graph Neural Networks
Sergei Ivanov,Liudmila Prokhorenkova
Graph neural networks (GNNs) are powerful models that have been successful in various graph representation learning tasks. Whereas gradient boosted decision trees (GBDT) often outperform other machine learning methods when faced with heterogeneous tabular data. But what approach should be used for graphs with tabular node features? Previous GNN models have mostly focused on networks with homogeneous sparse features and, as we show, are suboptimal in the heterogeneous setting. In this work, we propose a novel architecture that trains GBDT and GNN jointly to get the best of both worlds: the GBDT model deals with heterogeneous features, while GNN accounts for the graph structure. Our model benefits from end-to-end optimization by allowing new trees to fit the gradient updates of GNN. With an extensive experimental comparison to the leading GBDT and GNN models, we demonstrate a significant increase in performance on a variety of graphs with tabular features. The code is available: https://github.com/nd7141/bgnn.
****
Fighting Filterbubbles with Adversarial BERT-Training for News-Recommendation
Lukas Pfahler,Katharina Morik
Recommender engines play a role in the emergence and reinforcement of filter bubbles. When these systems learn that a user prefers content from a particular site, the user will be less likely to be exposed to different sources or opinions and, ultimately, is more likely to develop extremist tendencies.
We trace the roots of this phenomenon to the way the recommender engine represents news articles. The vectorial features modern systems extract from the plain text of news articles are already highly predictive of the associated news outlet. We propose a new training scheme based on adversarial machine learning to tackle this issue . Our experiments show that the features we can extract this way are significantly less predictive of the news outlet and thus offer the possibility to reduce the risk of manifestation of new filter bubbles. We validate our intuitions in a news recommendation task using a recent attention-based recommendation system.
****
Learning not to learn: Nature versus nurture in silico

Robert Tjarko Lange,Henning Sprekeler
Animals are equipped with a rich innate repertoire of sensory, behavioral and mo tor skills, which allows them to interact with the world immediately after birth . At the same time, many behaviors are highly adaptive and can be tailored to sp ecific environments by means of learning. In this work, we use mathematical anal ysis and the framework of meta-learning (or 'learning to learn') to answer when it is beneficial to learn such an adaptive strategy and when to hard-code a heur istic behavior. We find that the interplay of ecological uncertainty, task compl exity and the agents' lifetime has crucial effects on the meta-learned amortized Bayesian inference performed by an agent. There exist two regimes: One in which meta-learning yields a learning algorithm that implements task-dependent inform ation-integration and a second regime in which meta-learning imprints a heuristi c or 'hard-coded' behavior. Further analysis reveals that non-adaptive behaviors are not only optimal for aspects of the environment that are stable across indi viduals, but also in situations where an adaptation to the environment would in fact be highly beneficial, but could not be done quickly enough to be exploited within the remaining lifetime. Hard-coded behaviors should hence not only be tho se that always work, but also those that are too complex to be learned within a reasonable time frame.
****

Genetic Soft Updates for Policy Evolution in Deep Reinforcement Learning
Enrico Marchesini,Davide Corsi,Alessandro Farinelli
The combination of Evolutionary Algorithms (EAs) and Deep Reinforcement Learning (DRL) has been recently proposed to merge the benefits of both solutions. Exist ing mixed approaches, however, have been successfully applied only to actor-crit ic methods and present significant overhead. We address these issues by introduc ing a novel mixed framework that exploits a periodical genetic evaluation to sof t update the weights of a DRL agent. The resulting approach is applicable with a ny DRL method and, in a worst-case scenario, it does not exhibit detrimental beh aviours. Experiments in robotic applications and continuous control benchmarks d emonstrate the versatility of our approach that significantly outperforms prior DRL, EAs, and mixed approaches. Finally, we employ formal verification to confir m the policy improvement, mitigating the inefficient exploration and hyper-param eter sensitivity of DRL.ment, mitigating the inefficient exploration and hyper-p arameter sensitivity of DRL.
****

AutoCleansing: Unbiased Estimation of Deep Learning with Mislabeled Data
Koichi Kuriyama
Mislabeled samples cause prediction errors. This study proposes a solution to th e problem of incorrect labels, called AutoCleansing, to automatically capture th e effect of incorrect labels and mitigate it without removing the mislabeled sam ples. AutoCleansing consists of a base network model and sample-category specifi c constants. Both parameters of the base model and sample-category constants are estimated simultaneously using the training data. Thereafter, predictions for t est data are made using a base model without the constants capturing the mislabe led effects. A theoretical model for AutoCleansing is developed and showing that the gradient of the loss function of the proposed method can be zero at true pa rameters with mislabeled data if the model is correctly constructed. Experiment al results show that AutoCleansing has better performance in test accuracy than previous studies for CIFAR-10, CIFAR-100, SVHN, and ImageNet datasets.
****

Spatially Structured Recurrent Modules
Nasim Rahaman,Anirudh Goyal,Muhammad Waleed Gondal,Manuel Wuthrich,Stefan Bauer, Yash Sharma,Yoshua Bengio,Bernhard Schölkopf
Capturing the structure of a data-generating process by means of appropriate ind uctive biases can help in learning models that generalise well and are robust to changes in the input distribution. While methods that harness spatial and tempo ral structures find broad application, recent work has demonstrated the potentia l of models that leverage sparse and modular structure using an ensemble of spar ingly interacting modules. In this work, we take a step towards dynamic models t

hat are capable of simultaneously exploiting both modular and spatiotemporal str uctures. To this end, we model the dynamical system as a collection of autonomou s but sparsely interacting sub-systems that interact according to a learned topo logy which is informed by the spatial structure of the underlying system. This g ives rise to a class of models that are well suited for capturing the dynamics o f systems that only offer local views into their state, along with corresponding spatial locations of those views. On the tasks of video prediction from cropped frames and multi-agent world modelling from partial observations in the challen ging Starcraft2 domain, we find our models to be more robust to the number of av ailable views and better capable of generalisation to novel tasks without additi onal training than strong baselines that perform equally well or better on the t raining distribution.
****

Generalizing Graph Convolutional Networks via Heat Kernel
Jialin Zhao,Yuxiao Dong,Jie Tang,Ming Ding,Kuansan Wang
Graph convolutional networks (GCNs) have emerged as a powerful framework for min ing and learning with graphs. A recent study shows that GCNs can be simplified a s a linear model by removing nonlinearities and weight matrices across all conse cutive layers, resulting the simple graph convolution (SGC) model. In this paper , we aim to understand GCNs and generalize SGC as a linear model via heat kernel (HKGCN), which acts as a low-pass filter on graphs and enables the aggregation of information from extremely large receptive fields. We theoretically show that HKGCN is in nature a continuous propagation model and GCNs without nonlineariti es (i.e., SGC) are the discrete versions of it. Its low-pass filter and continui ty properties facilitate the fast and smooth convergence of feature propagation. Experiments on million-scale networks show that the linear HKGCN model not only achieves consistently better results than SGC but also can match or even beat a dvanced GCN models, while maintaining SGC’s superiority in efficiency.
****

iPTR: Learning a representation for interactive program translation retrieval
Binger Chen,Ziawasch Abedjan
Program translation contributes to many real world scenarios, such as porting co debases written in an obsolete or deprecated language to a modern one or re-impl ementing existing projects in one's preferred programming language. Existing dat a-driven approaches either require large amounts of training data or neglect sig nificant characteristics of programs. In this paper, we present iPTR for interac tive code translation retrieval from Big Code. iPTR uses a novel code representa tion technique that encodes structural characteristics of a program and a predic tive transformation technique to transform the representation into the target pr ogramming language. The transformed representation is used for code retrieval fr om Big Code. With our succinct representation, the user can easily update and co rrect the returned results to improve the retrieval process. Our experiments sho w that iPTR outperforms supervised baselines in terms of program accuracy.
****

Expressive Power of Invariant and Equivariant Graph Neural Networks
Waiss Azizian,marc lelarge
Various classes of Graph Neural Networks (GNN) have been proposed and shown to b e successful in a wide range of applications with graph structured data. In this paper, we propose a theoretical framework able to compare the expressive power of these GNN architectures. The current universality theorems only apply to intr actable classes of GNNs. Here, we prove the first approximation guarantees for p ractical GNNs, paving the way for a better understanding of their generalization . Our theoretical results are proved for invariant GNNs computing a graph embedd ing (permutation of the nodes of the input graph does not affect the output) and equivariant GNNs computing an embedding of the nodes (permutation of the input permutes the output). We show that Folklore Graph Neural Networks (FGNN), which are tensor based GNNs augmented with matrix multiplication are the most expressi ve architectures proposed so far for a given tensor order. We illustrate our res ults on the Quadratic Assignment Problem (a NP-Hard combinatorial problem) by sh owing that FGNNs are able to learn how to solve the problem, leading to much bet

ter average performances than existing algorithms (based on spectral, SDP or oth
er GNNs architectures). On a practical side, we also implement masked tensors to
 handle batches of graphs of varying sizes.
****

Stochastic Normalized Gradient Descent with Momentum for Large Batch Training
Shen-Yi Zhao,Yin-Peng Xie,Wu-Jun Li
Stochastic gradient descent (SGD) and its variants have been the dominating opti
mization methods in machine learning. Compared with small batch training, SGD wi
th large batch training can better utilize the computational power of current mu
lti-core systems like GPUs and can reduce the number of communication rounds in
 distributed training. Hence, SGD with large batch training has attracted more an
d more attention. However, existing empirical results show that large batch trai
ning typically leads to a drop of generalization accuracy. As a result, large b
atch training has also become a challenging topic. In this paper, we propose a n
ovel method, called stochastic normalized gradient descent with momentum (SNGM),
 for large batch training. We theoretically prove that compared to momentum SGD
(MSGD) which is one of the most widely used variants of SGD, SNGM can adopt a la
rger batch size to converge to the ϵ-stationary point with the same com
putation complexity (total number of gradient computation). Empirical results on
 deep learning also show that SNGM can achieve the state-of-the-art accuracy wit
h a large batch size.
****

On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong B
aselines
Marius Mosbach,Maksym Andriushchenko,Dietrich Klakow
Fine-tuning pre-trained transformer-based language models such as BERT has becom
e a common practice dominating leaderboards across various NLP benchmarks. Despi
te the strong empirical performance of fine-tuned models, fine-tuning is an unst
able process: training the same model with multiple random seeds can result in a
 large variance of the task performance. Previous literature (Devlin et al., 201
9; Lee et al., 2020; Dodge et al., 2020) identified two potential reasons for th
e observed instability: catastrophic forgetting and small size of the fine-tunin
g datasets. In this paper, we show that both hypotheses fail to explain the fine
-tuning instability. We analyze BERT, RoBERTa, and ALBERT, fine-tuned on commonl
y used datasets from the GLUE benchmark, and show that the observed instability
is caused by optimization difficulties that lead to vanishing gradients. Additio
nally, we show that the remaining variance of the downstream task performance ca
n be attributed to differences in generalization where fine-tuned models with th
e same training loss exhibit noticeably different test performance. Based on our
 analysis, we present a simple but strong baseline that makes fine-tuning BERT-b
ased models significantly more stable than the previously proposed approaches. C
ode to reproduce our results is available online: https://github.com/uds-lsv/ber
t-stable-fine-tuning.
****

Mitigating Mode Collapse by Sidestepping Catastrophic Forgetting
Karttikeya Mangalam,Rohin Garg,Jathushan Rajasegaran,Taesung Park
Generative Adversarial Networks (GANs) are a class of generative models used for
 various applications, but they have been known to suffer from the mode collapse
 problem, in which some modes of the target distribution are ignored by the gene
rator. Investigative study using a new data generation procedure indicates that
 the mode collapse of the generator is driven by the discriminator’s inability t
o maintain classification accuracy on previously seen samples, a phenomenon call
ed Catastrophic Forgetting in continual learning. Motivated by this observation,
 we introduce a novel training procedure that dynamically spawns additional dis-
criminators to remember previous modes of generation. On several datasets, we sh
ow that our training scheme can be plugged-in to existing GAN frameworks to miti
gate mode collapse and improve standard metrics for GAN evaluation.
****

Convergence Analysis of Homotopy-SGD for Non-Convex Optimization
Matilde Gargiani,Andrea Zanelli,Moritz Diehl,Quoc Tran-Dinh,Frank Hutter

First-order stochastic methods for solving large-scale non-convex optimization p
roblems are widely used in many big-data applications, e.g. training deep neural
 networks as well as other complex and potentially non-convex machine learning
models. Their inexpensive iterations generally come together with slow global co
nvergence rate (mostly sublinear), leading to the necessity of carrying out a ve
ry high number of iterations before the iterates reach a neighborhood of a minim
izer. In this work, we present a first-order stochastic algorithm based on a com
bination of homotopy methods and SGD, called Homotopy-Stochastic Gradient Descen
t (H-SGD), which finds interesting connections with some proposed heuristics in
the literature, e.g. optimization by Gaussian continuation, training by diffusio
n, mollifying networks. Under some mild and realistic assumptions on the problem
 structure, we conduct a theoretical analysis of the proposed algorithm. Our ana
lysis shows that, with a specifically designed scheme for the homotopy parameter
, H-SGD enjoys a global linear rate of convergence to a neighborhood of a minimi
zer while maintaining fast and inexpensive iterations. Experimental evaluations
confirm the theoretical results and show that H-SGD can outperform standard SGD.
****

Systematic Analysis of Cluster Similarity Indices: How to Validate Validation Me
asures

Martijn Gösgens,Liudmila Prokhorenkova,Aleksei Tikhonov

There are many cluster similarity indices used to evaluate clustering algorithms
, and choosing the best one for a particular task remains an open problem. We de
monstrate that this problem is crucial: there are many disagreements among the i
ndices, these disagreements do affect which algorithms are chosen in application
s, and this can lead to degraded performance in real-world systems. We propose a
 theoretical solution to this problem: we develop a list of desirable properties
 and theoretically verify which indices satisfy them. This allows for making an
informed choice: given a particular application, one can first make a selection
of properties that are desirable for a given application and then identify indic
es satisfying these. We observe that many popular indices have significant drawb
acks. Instead, we advocate using other ones that are not so widely adopted but h
ave beneficial properties.
****

End-to-End Egospheric Spatial Memory

Daniel James Lenton,Stephen James,Ronald Clark,Andrew Davison

Spatial memory, or the ability to remember and recall specific locations and obj
ects, is central to autonomous agents' ability to carry out tasks in real enviro
nments. However, most existing artificial memory modules are not very adept at s
toring spatial information. We propose a parameter-free module, Egospheric Spati
al Memory (ESM), which encodes the memory in an ego-sphere around the agent, ena
bling expressive 3D representations. ESM can be trained end-to-end via either im
itation or reinforcement learning, and improves both training efficiency and fin
al performance against other memory baselines on both drone and manipulator visu
omotor control tasks. The explicit egocentric geometry also enables us to seamle
ssly combine the learned controller with other non-learned modalities, such as l
ocal obstacle avoidance. We further show applications to semantic segmentation o
n the ScanNet dataset, where ESM naturally combines image-level and map-level in
ference modalities. Through our broad set of experiments, we show that ESM provi
des a general computation graph for embodied spatial reasoning, and the module f
orms a bridge between real-time mapping systems and differentiable memory archit
ectures. Implementation at: https://github.com/ivy-dl/memory.
****

Toward Trainability of Quantum Neural Networks

Kaining Zhang,Min-Hsiu Hsieh,Liu Liu,Dacheng Tao

Quantum Neural Networks (QNNs) have been recently proposed as generalizations of
 classical neural networks to achieve the quantum speed-up. Despite the potentia
l to outperform classical models, serious bottlenecks exist for training QNNs; n
amely, QNNs with random structures have poor trainability due to the vanishing g
radient with rate exponential to the input qubit number. The vanishing gradient
could seriously influence the applications of large-size QNNs. In this work, we

provide the first viable solution with theoretical guarantees. Specifically, we prove that QNNs with tree tensor and step controlled architectures have gradients that vanish at most polynomially with the qubit number. Moreover, our result holds irrespective of which encoding methods are employed. We numerically demonstrate QNNs with tree tensor and step controlled structures for the application of binary classification. Simulations show faster convergent rates and better accuracy compared to QNNs with random structures.
****
LEAF: A Learnable Frontend for Audio Classification
Neil Zeghidour,Olivier Teboul,Félix de Chaumont Quitry,Marco Tagliasacchi
Mel-filterbanks are fixed, engineered audio features which emulate human perception and have been used through the history of audio understanding up to today. However, their undeniable qualities are counterbalanced by the fundamental limitations of handmade representations. In this work we show that we can train a single learnable frontend that outperforms mel-filterbanks on a wide range of audio signals, including speech, music, audio events and animal sounds, providing a general-purpose learned frontend for audio classification. To do so, we introduce a new principled, lightweight, fully learnable architecture that can be used as a drop-in replacement of mel-filterbanks. Our system learns all operations of audio features extraction, from filtering to pooling, compression and normalization, and can be integrated into any neural network at a negligible parameter cost. We perform multi-task training on eight diverse audio classification tasks, and show consistent improvements of our model over mel-filterbanks and previous learnable alternatives. Moreover, our system outperforms the current state-of-the-art learnable frontend on Audioset, with orders of magnitude fewer parameters.
****
The Impact of the Mini-batch Size on the Dynamics of SGD: Variance and Beyond
Xin Qian,Diego Klabjan
We study mini-batch stochastic gradient descent (SGD) dynamics under linear regression and deep linear networks by focusing on the variance of the gradients only given the initial weights and mini-batch size, which is the first study of this nature. In the linear regression case, we show that in each iteration the norm of the gradient is a decreasing function of the mini-batch size b and thus the variance of the stochastic gradient estimator is a decreasing function of b. For deep neural networks with L_2 loss we show that the variance of the gradient is a polynomial in $1/b$. The results theoretically back the important intuition that smaller batch sizes yield larger variance of the stochastic gradients and lower loss function values which is a common believe among the researchers. The proof techniques exhibit a relationship between stochastic gradient estimators and initial weights, which is useful for further research on the dynamics of SGD. We empirically provide insights to our results on various datasets and commonly used deep network structures. We further discuss possible extensions of the approaches we build in studying the generalization ability of the deep learning models.
****
Simple Augmentation Goes a Long Way: ADRL for DNN Quantization
Lin Ning,Guoyang Chen,Weifeng Zhang,Xipeng Shen
Mixed precision quantization improves DNN performance by assigning different layers with different bit-width values. Searching for the optimal bit-width for each layer, however, remains a challenge. Deep Reinforcement Learning (DRL) shows some recent promise. It however suffers instability due to function approximation errors, causing large variances in the early training stages, slow convergence, and suboptimal policies in the mixed-precision quantization problem. This paper proposes augmented DRL (ADRL) as a way to alleviate these issues. This new strategy augments the neural networks in DRL with a complementary scheme to boost the performance of learning. The paper examines the effectiveness of ADRL both analytically and empirically, showing that it can produce more accurate quantized models than the state of the art DRL-based quantization while improving the learning speed by 4.5-64 times.
****

Design-Bench: Benchmarks for Data-Driven Offline Model-Based Optimization
Brandon Trabucco,Aviral Kumar,Xinyang Geng,Sergey Levine
Black-box model-based optimization (MBO) problems, where the goal is to find a d esign input that maximizes an unknown objective function, are ubiquitous in a wi de range of domains, such as the design of drugs, aircraft, and robot morphology . Typically, such problems are solved by actively querying the black-box objecti ve on design proposals and using the resulting feedback to improve the proposed designs. However, when the true objective function is expensive or dangerous to evaluate in the real world, we might instead prefer a method that can optimize t his function using only previously collected data, for example from a set of pre viously conducted experiments. This data-driven offline MBO set- ting presents a number of unique challenges, but a number of recent works have demonstrated tha t viable offline MBO methods can be developed even for high- dimensional problem s, using high-capacity deep neural network function approximators. Unfortunately , the lack of standardized evaluation tasks in this emerg- ing new field has mad e tracking progress and comparing recent methods difficult. To address this prob lem, we present Design-Bench, a benchmark suite of offline MBO tasks with a unif ied evaluation protocol and reference implementations of recent methods. Our ben chmark suite includes diverse and realistic tasks derived from real-world proble ms in biology, material science, and robotics that present distinct challenges f or offline MBO methods. Our benchmarks, together with the reference implementati ons, are available at sites.google.com/view/design-bench. We hope that our bench mark can serve as a meaningful metric for the progress of offline MBO methods an d guide future algorithmic development.
****

Multi-View Disentangled Representation
Zongbo Han,Changqing Zhang,Huazhu Fu,Qinghua Hu,Joey Tianyi Zhou
Learning effective representations for data with multiple views is crucial in ma chine learning and pattern recognition. Recently great efforts have focused on l earning unified or latent representations to integrate information from differen t views for specific tasks. These approaches generally assume simple or implicit relationships between different views and as a result are not able to flexibly and explicitly depict the correlations among these views. To address this, we fi rstly propose the definition and conditions for multi-view disentanglement provi ding general instructions for disentangling representations between different vi ews. Furthermore, a novel objective function is derived to explicitly disentangl e the multi-view data into a shared part across different views and a (private) exclusive part within each view. Experiments on a variety of multi-modal dataset s demonstrate that our objective can effectively disentangle information from di fferent views while satisfying the disentangling conditions.
****

Polynomial Graph Convolutional Networks
Luca Pasa,Nicolò Navarin,Alessandro Sperduti
Graph Convolutional Neural Networks (GCNs) exploit convolution operators, based on some neighborhood aggregating scheme, to compute representations of graphs. T he most common convolution operators only exploit local topological information. To consider wider topological receptive fields, the mainstream approach is to n on-linearly stack multiple Graph Convolutional (GC) layers. In this way, however , interactions among GC parameters at different levels pose a bias on the flow o f topological information. In this paper, we propose a different strategy, cons idering a single graph convolution layer that independently exploits neighbourin g nodes at different topological distances, generating decoupled representations for each of them. These representations are then processed by subsequent reado ut layers. We implement this strategy introducing the Polynomial Graph Convoluti on (PGC) layer, that we prove being more expressive than the most common convolu tion operators and their linear stacking. Our contribution is not limited to the definition of a convolution operator with a larger receptive field, but we prov e both theoretically and experimentally that the common way multiple non-linear graph convolutions are stacked limits the neural network expressiveness. Specif ically, we show that a Graph Neural Network architecture with a single PGC layer

achieves state of the art performance on many commonly adopted graph classification benchmarks.
****
The inductive bias of ReLU networks on orthogonally separable data
Mary Phuong,Christoph H Lampert
We study the inductive bias of two-layer ReLU networks trained by gradient flow. We identify a class of easy-to-learn (`orthogonally separable') datasets, and characterise the solution that ReLU networks trained on such datasets converge to. Irrespective of network width, the solution turns out to be a combination of two max-margin classifiers: one corresponding to the positive data subset and one corresponding to the negative data subset.
The proof is based on the recently introduced concept of extremal sectors, for which we prove a number of properties in the context of orthogonal separability. In particular, we prove stationarity of activation patterns from some time T onwards, which enables a reduction of the ReLU network to an ensemble of linear subnetworks.

****
Recycling sub-optimial Hyperparameter Optimization models to generate efficient Ensemble Deep Learning
Pierrick Pochelu,Bruno Conche,Serge G. Petiton
Ensemble Deep Learning improves accuracy over a single model by combining predictions from multiple models. It has established itself to be the core strategy for tackling the most difficult problems, like winning Kaggle challenges. Due to the lack of consensus to design a successful deep learning ensemble, we introduce Hyperband-Dijkstra, a new workflow that automatically explores neural network designs with Hyperband and efficiently combines them with Dijkstra's algorithm. This workflow has the same training cost than standard Hyperband running except sub-optimal solutions are stored and are candidates to be selected in the ensemble selection step (recycling). Next, to predict on new data, the user gives to Dijkstra the maximum number of models wanted in the ensemble to control the tradeoff between accuracy and inference time.
Hyperband is a very efficient algorithm allocating exponentially more resources to the most promising configurations. It is also capable to propose diverse models due to its pure-exploration nature, which allows Dijkstra algorithm with a smart combination of diverse models to achieve a strong variance and bias reduction. The exploding number of possible combinations generated by Hyperband increases the probability that Dijkstra finds an accurate combination which fits the dataset and generalizes on new data.
The two experimentation on CIFAR100 and on our unbalanced microfossils dataset show that our new workflow generates an ensemble far more accurate than any other ensemble of any ResNet models from ResNet18 to ResNet152.
****
Monte-Carlo Planning and Learning with Language Action Value Estimates
Youngsoo Jang,Seokin Seo,Jongmin Lee,Kee-Eung Kim
Interactive Fiction (IF) games provide a useful testbed for language-based reinforcement learning agents, posing significant challenges of natural language understanding, commonsense reasoning, and non-myopic planning in the combinatorial search space. Agents based on standard planning algorithms struggle to play IF games due to the massive search space of language actions. Thus, language-grounded planning is a key ability of such agents, since inferring the consequence of language action based on semantic understanding can drastically improve search. In this paper, we introduce Monte-Carlo planning with Language Action Value Estimates (MC-LAVE) that combines a Monte-Carlo tree search with language-driven exploration. MC-LAVE invests more search effort into semantically promising language actions using locally optimistic language value estimates, yielding a significant reduction in the effective search space of language actions. We then present a reinforcement learning approach via MC-LAVE, which alternates between MC-LAVE planning and supervised learning of the self-generated language actions. In the experiments, we demonstrate that our method achieves new high scores in various I

F games.
****
Learning Energy-Based Models by Diffusion Recovery Likelihood
Ruiqi Gao,Yang Song,Ben Poole,Ying Nian Wu,Diederik P Kingma
While energy-based models (EBMs) exhibit a number of desirable properties, training and sampling on high-dimensional datasets remains challenging. Inspired by recent progress on diffusion probabilistic models, we present a diffusion recovery likelihood method to tractably learn and sample from a sequence of EBMs trained on increasingly noisy versions of a dataset. Each EBM is trained with recovery likelihood, which maximizes the conditional probability of the data at a certain noise level given their noisy versions at a higher noise level. Optimizing recovery likelihood is more tractable than marginal likelihood, as sampling from the conditional distributions is much easier than sampling from the marginal distributions. After training, synthesized images can be generated by the sampling process that initializes from Gaussian white noise distribution and progressively samples the conditional distributions at decreasingly lower noise levels. Our method generates high fidelity samples on various image datasets. On unconditional CIFAR-10 our method achieves FID 9.58 and inception score 8.30, superior to the majority of GANs. Moreover, we demonstrate that unlike previous work on EBMs, our long-run MCMC samples from the conditional distributions do not diverge and still represent realistic images, allowing us to accurately estimate the normalized density of data even for high-dimensional datasets. Our implementation is available at \url{https://github.com/ruiqigao/recovery_likelihood}.

****
Similarity Search for Efficient Active Learning and Search of Rare Concepts
Cody Coleman,Edward Chou,Sean Culatana,Peter Bailis,Alexander C. Berg,Roshan Sumbaly,Matei Zaharia,I. Zeki Yalniz
Many active learning and search approaches are intractable for industrial settings with billions of unlabeled examples. Existing approaches, such as uncertainty sampling or information density, search globally for the optimal examples to label, scaling linearly or even quadratically with the unlabeled data. However, in practice, data is often heavily skewed; only a small fraction of collected data will be relevant for a given learning task. For example, when identifying rare classes, detecting malicious content, or debugging model performance, positive examples can appear in less than 1% of the data. In this work, we exploit this skew in large training datasets to reduce the number of unlabeled examples considered in each selection round by only looking at the nearest neighbors to the labeled examples. Empirically, we observe that learned representations can effectively cluster unseen concepts, making active learning very effective and substantially reducing the number of viable unlabeled examples. We evaluate several selection strategies in this setting on three large-scale computer vision datasets: ImageNet, OpenImages, and a proprietary dataset of 10 billion images from a large internet company. For rare classes, active learning methods need as little as 0.31% of the labeled data to match the average precision of full supervision. By limiting the selection strategies to the immediate neighbors of the labeled data as candidates for labeling, we process as little as 0.1% of the unlabeled data while achieving similar reductions in labeling costs as the traditional global approach. This process of expanding the candidate pool with the nearest neighbors of the labeled set can be done efficiently and reduces the computational complexity of selection by orders of magnitude.
****
On Disentangled Representations Extracted from Pretrained GANs
Valentin Khrulkov,Leyla Mirvakhabova,Ivan Oseledets,Artem Babenko
Constructing disentangled representations is known to be a difficult task, especially in the unsupervised scenario. The dominating paradigm of unsupervised disentanglement is currently to train a generative model that separates different factors of variation in its latent space. This separation is typically enforced by training with specific regularization terms in the model's objective function. These terms, however, introduce additional hyperparameters responsible for the t

rade-off between disentanglement and generation quality. While tuning these hype rparameters is crucial for proper disentanglement, it is often unclear how to tu ne them without external supervision.

This paper investigates an alternative route to disentangled representations. Na mely, we propose to extract such representations from the state-of-the-art GANs trained without disentangling terms in their objectives. This paradigm of post h oc disentanglement employs little or no hyperparameters when learning representa tions, while achieving results on par with existing state-of-the-art, as shown b y comparison in terms of established disentanglement metrics, fairness, and the abstract reasoning task.
All our code and models are publicly available.
****
Real-time Uncertainty Decomposition for Online Learning Control
Jonas Umlauft,Armin Lederer,Thomas Beckers,Sandra Hirche
Safety-critical decisions based on machine learning models require a clear under standing of the involved uncertainties to avoid hazardous or risky situations. W hile aleatoric uncertainty can be explicitly modeled given a parametric descript ion, epistemic uncertainty rather describes the presents or absence of training data. This paper proposes a novel generic method for modeling
epistemic uncertainty and shows its advantages over existing approaches for neur al networks on various data sets. It can be directly combined with aleatoric unc ertainty estimates and
allows for prediction in real-time as the inference is sample-free. We exploit t his property in a model-based quadcopter control setting and demonstrate how the controller benefits from a differentiation between aleatoric and epistemic unce rtainty in online learning of thermal disturbances.
****
Capturing Label Characteristics in VAEs
Tom Joy,Sebastian Schmon,Philip Torr,Siddharth N,Tom Rainforth
We present a principled approach to incorporating labels in variational autoenco ders (VAEs) that captures the rich characteristic information associated with th ose labels. While prior work has typically conflated these by learning latent va riables that directly correspond to label values, we argue this is contrary to t he intended effect of supervision in VAEs—capturing rich label characteristics w ith the latents. For example, we may want to capture the characteristics of a fa ce that make it look young, rather than just the age of the person. To this end, we develop a novel VAE model, the characteristic capturing VAE (CCVAE), which “ reparameterizes” supervision through auxiliary variables and a concomitant varia tional objective. Through judicious structuring of mappings between latent and a uxiliary variables, we show that the CCVAE can effectively learn meaningful repr esentations of the characteristics of interest across a variety of supervision s chemes. In particular, we show that the CCVAE allows for more effective and more general interventions to be performed, such as smooth traversals within the cha racteristics for a given label, diverse conditional generation, and transferring characteristics across datapoints.
****
Learning the Step-size Policy for the Limited-Memory Broyden-Fletcher-Goldfarb-S hanno Algorithm
Lucas N. Egidio,Anders Hansson,Bo Wahlberg
We consider the problem of how to learn a step-size policy for the Limited-Memor y Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm. This is a limited computa tional memory quasi-Newton method widely used for deterministic unconstrained op timization but currently avoided in large-scale problems for requiring step size s to be provided at each iteration. Existing methodologies for the step size sel ection for L-BFGS use heuristic tuning of design parameters and massive re-evalu ations of the objective function and gradient to find appropriate step-lengths. We propose a neural network architecture with local information of the current i terate as the input. The step-length policy is learned from data of similar opti mization problems, avoids additional evaluations of the objective function, and

guarantees that the output step remains inside a pre-defined interval. The corresponding training procedure is formulated as a stochastic optimization problem using the backpropagation through time algorithm. The performance of the proposed method is evaluated on the training of classifiers for the MNIST database for handwritten digits and for CIFAR-10. The results show that the proposed algorithm outperforms heuristically tuned optimizers such as ADAM, RMSprop, L-BFGS with a backtracking line search and L-BFGS with a constant step size. The numerical results also show that a learned policy can be used as a warm-start to train new policies for different problems after a few additional training steps, highlighting its potential use in multiple large-scale optimization problems.
****

CaLFADS: latent factor analysis of dynamical systems in calcium imaging data
Luke Yuri Prince,Shahab Bakhtiari,Colleen J Gillon,Blake Aaron Richards
Dynamic latent variable modelling has been a hugely powerful tool in understanding how spiking activity in populations of neurons can perform computations necessary for adaptive behaviour. The success of such approaches has been enabled by the ability to construct models derived with the characterization of spiking activity as point-processes since spiking dynamics occur on a much faster time-scale than the computational dynamics being inferred. Other experimental techniques, such as calcium imaging, pose a problem for latent variable modelling of computational dynamics, since the time-scales of calcium dynamics and computational dynamics overlap. As such, the success of dynamic latent variable modelling in calcium imaging data rests on being able to disentangle the contribution of these two sources of variation. Here we extend recent advances using variational autoencoders to analyze neural data, by incorporating a ladder architecture that can infer a hierarchy of dynamical systems. Using built-in inductive biases for calcium dynamics, we can capture calcium flux as well as underlying dynamics of neural computation. First, we demonstrate with synthetic calcium data that we can correctly infer an underlying Lorenz attractor at the same time as calcium dynamics. Next, we show that we can infer appropriate rotational dynamics in spiking data from macaque motor cortex after it has been converted into calcium fluorescence data via a calcium dynamics model. Finally, we show that our method applied to real calcium imaging data from primary visual cortex in mice allows us to infer latent factors that carry salient sensory information about unexpected stimuli. These results demonstrate that variational ladder autoencoders are a promising approach for inferring hierarchical dynamics in experimental settings where the measured variable has its own slow dynamics, such as calcium imaging data, thereby providing the neuroscience community with a new analysis tool for a wider array of data modalities.
****

Linear Mode Connectivity in Multitask and Continual Learning
Seyed Iman Mirzadeh,Mehrdad Farajtabar,Dilan Gorur,Razvan Pascanu,Hassan Ghasemzadeh
Continual (sequential) training and multitask (simultaneous) training are often attempting to solve the same overall objective: to find a solution that performs well on all considered tasks. The main difference is in the training regimes, where continual learning can only have access to one task at a time, which for neural networks typically leads to catastrophic forgetting. That is, the solution found for a subsequent task does not perform well on the previous ones anymore.

However, the relationship between the different minima that the two training regimes arrive at is not well understood. What sets them apart? Is there a local structure that could explain the difference in performance achieved by the two different schemes?

Motivated by recent work showing that different minima of the same task are typically connected by very simple curves of low error, we investigate whether multitask and continual solutions are similarly connected. We empirically find that indeed such connectivity can be reliably achieved and, more interestingly, it can be done by a linear path, conditioned on having the same initialization for both. We thoroughly analyze this observation and discuss its significance for the continual learning process.

Furthermore, we exploit this finding to propose an effective algorithm that constrains the sequentially learned minima to behave as the multitask solution. We show that our method outperforms several state of the art continual learning algorithms on various vision benchmarks.
****
Importance and Coherence: Methods for Evaluating Modularity in Neural Networks
Shlomi Hod,Stephen Casper,Daniel Filan,Cody Wild,Andrew Critch,Stuart Russell
As deep neural networks become more advanced and widely-used, it is important to understand their inner workings. Toward this goal, modular interpretations are appealing because they offer flexible levels of abstraction aside from standard architectural building blocks (e.g., neurons, channels, layers). In this paper, we consider the problem of assessing how functionally interpretable a given partitioning of neurons is. 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 input/output features. To measure these proxies, we develop a set of statistical methods based on techniques that have conventionally been used for the interpretation of individual neurons. We apply these methods on partitionings generated by a spectral clustering algorithm which uses a graph representation of the network's neurons and weights. We show that despite our partitioning algorithm using neither activations nor gradients, it reveals clusters with a surprising amount of importance and coherence. Together, these results support the use of modular interpretations, and graph-based partitionings in particular, for interpretability.
****
MC-LSTM: Mass-conserving LSTM
Pieter-Jan Hoedt,Frederik Kratzert,Daniel Klotz,Christina Halmich,Markus Holzleitner,Grey Nearing,Sepp Hochreiter,Günter Klambauer
The success of Convolutional Neural Networks (CNNs) in computer vision is mainly driven by their strong inductive bias, which is strong enough to allow CNNs to solve vision-related tasks with random weights, meaning without learning. Similarly, Long Short-Term Memory (LSTM) has a strong inductive bias towards storing information over time. However, many real-world systems are governed by conservation laws, which lead to the redistribution of particular quantities —e.g. in physical and economical systems. Our novel Mass-Conserving LSTM (MC-LSTM) adheres to these conservation laws by extending the inductive bias of LSTM to model the redistribution of those stored quantities. MC-LSTMs set a new state-of-the-art for neural arithmetic units at learning arithmetic operations, such as addition tasks, which have a strong conservation law, as the sum is constant overtime. Further, MC-LSTM is applied to traffic forecasting, modeling a pendulum, and a large benchmark dataset in hydrology, where it sets a new state-of-the-art for predicting peak flows. In the hydrology example, we show that MC-LSTM states correlate with real world processes and are therefore interpretable.
****
Length-Adaptive Transformer: Train Once with Length Drop, Use Anytime with Search
Gyuwan Kim,Kyunghyun Cho
Although transformers have achieved impressive accuracies in various tasks in natural language processing, they often come with a prohibitive computational cost, that prevents their use in scenarios with limited computational resources for inference. This need for computational efficiency in inference has been addressed by for instance PoWER-BERT (Goyal et al., 2020) which gradually decreases the length of a sequence as it is passed through layers. These approaches however often assume that the target computational complexity is known in advance at the time of training. This implies that a separate model must be trained for each inference scenario with its distinct computational budget. In this paper, we extend PoWER-BERT to address this issue of inefficiency and redundancy. The proposed extension enables us to train a large-scale transformer, called Length-Adaptive Transformer, once and uses it for various inference scenarios without re-training it. To do so, we train a transformer with LengthDrop, a structural variant of dropout, which stochastically determines the length of a sequence at each layer.

We then use a multi-objective evolutionary search to find a length configuration that maximizes the accuracy and minimizes the computational complexity under any given computational budget. Additionally, we significantly extend the applicability of PoWER-BERT beyond sequence-level classification into token-level classification such as span-based question-answering, by introducing the idea of Drop-and-Restore. With Drop-and-Restore, word-vectors are dropped temporarily in intermediate layers and restored at the last layer if necessary. We empirically verify the utility of the proposed approach by demonstrating the superior accuracy-efficiency trade-off under various setups, including SQuAD 1.1, MNLI-m, and SST-2. Upon publication, the code to reproduce our work will be open-sourced.

****
Computational Separation Between Convolutional and Fully-Connected Networks
eran malach,Shai Shalev-Shwartz
Convolutional neural networks (CNN) exhibit unmatched performance in a multitude of computer vision tasks. However, the advantage of using convolutional networks over fully-connected networks is not understood from a theoretical perspective. In this work, we show how convolutional networks can leverage locality in the data, and thus achieve a computational advantage over fully-connected networks. Specifically, we show a class of problems that can be efficiently solved using convolutional networks trained with gradient-descent, but at the same time is hard to learn using a polynomial-size fully-connected network.
****
Model-based Asynchronous Hyperparameter and Neural Architecture Search
Aaron Klein,Louis Chi-Chun Tiao,Thibaut Lienart,Cedric Archambeau,Matthias Seeger
We introduce a model-based asynchronous multi-fidelity method for hyperparameter and neural architecture search that combines the strengths of asynchronous Successive Halving and Gaussian process-based Bayesian optimization. At the heart of our method is a probabilistic model that can simultaneously reason across hyperparameters and resource levels, and supports decision-making in the presence of pending evaluations. We demonstrate the effectiveness of our method on a wide range of challenging benchmarks, for tabular data, image classification and language modelling, and report substantial speed-ups over current state-of-the-art methods. Our new methods, along with asynchronous baselines, are implemented in a distributed framework which will be open sourced along with this publication.
****
ItNet: iterative neural networks for fast and efficient anytime prediction
Thomas Pfeil
Deep neural networks have usually to be compressed and accelerated for their usage in low-power, e.g. mobile, devices. Common requirements are high accuracy, high throughput, low latency, and a small memory footprint. A good trade-off between accuracy and latency has been shown by networks comprising multiple intermediate outputs. In this study, we introduce a multi-output network that has a tiny memory footprint in terms of its computational graph, which allows its execution on novel, massively-parallel hardware accelerators designed for extremely high throughput. To this end, the graph is designed to contain loops by iteratively executing a single network building block. These so-called iterative neural networks enable stateof-the-art results for semantic segmentation on the CamVid and Cityscapes datasets that are especially demanding in terms of computational resources. In ablation studies, the improvement of network training by intermediate network outputs as well as the trade-off between weight sharing over iterations and the network size are investigated.
****
The Negative Pretraining Effect in Sequential Deep Learning and Three Ways to Fix It
Julian G. Zilly,Franziska Eckert,Bhairav Mehta,Andrea Censi,Emilio Frazzoli
Negative pretraining is a prominent sequential learning effect of neural networks where a pretrained model obtains a worse generalization performance than a model that is trained from scratch when either are trained on a target task. We con

ceptualize the ingredients of this problem setting and examine the negative pret raining effect experimentally by providing three interventions to remove and fix it. First, acting on the learning process, altering the learning rate after pre training can yield even better results than training directly on the target task . Second, on the learning task-level, we intervene by increasing the discretizat ion of data distribution changes from start to target task instead of “jumping” to a target task. Finally at the model-level, resetting network biases to larger values likewise removes negative pretraining effects, albeit to a smaller degre e. With these intervention experiments, we aim to provide new evidence to help u nderstand the subtle influences that neural network training and pretraining can have on final generalization performance on a target task in the context of neg ative pretraining.
****

Benchmarking Unsupervised Object Representations for Video Sequences
Marissa A. Weis,Kashyap Chitta,Yash Sharma,Wieland Brendel,Matthias Bethge,Andre as Geiger,Alexander S Ecker
Perceiving the world in terms of objects and tracking them through time is a cru cial prerequisite for reasoning and scene understanding. Recently, several metho ds have been proposed for unsupervised learning of object-centric representation s. However, since these models have been evaluated with respect to different dow nstream tasks, it remains unclear how they compare in terms of basic perceptual abilities such as detection, figure-ground segmentation and tracking of individu al objects. To close this gap, we design a benchmark with three datasets of vary ing complexity and seven additional test sets which feature challenging tracking scenarios relevant for natural videos. Using this benchmark, we compare the per ceptual abilities of four unsupervised object-centric learning approaches: ViMON , a video-extension of MONet, based on a recurrent spatial attention mechanism, OP3, which exploits clustering via spatial mixture models, as well as TBA and SC ALOR, which use an explicit factorization via spatial transformers. Our results suggest that architectures with unconstrained latent representations and full-im age object masks such as ViMON and OP3 are able to learn more powerful represent ations in terms of object detection, segmentation and tracking than the explicit ly parameterized spatial transformer based architecture of TBA and SCALOR. We al so observe that none of the methods are able to gracefully handle the most chall enging tracking scenarios despite their synthetic nature, suggesting that our be nchmark may provide fruitful guidance towards learning more robust object-centri c video representations.
****

Rethinking Embedding Coupling in Pre-trained Language Models
Hyung Won Chung,Thibault Fevry,Henry Tsai,Melvin Johnson,Sebastian Ruder
We re-evaluate the standard practice of sharing weights between input and output embeddings in state-of-the-art pre-trained language models. We show that decoup led embeddings provide increased modeling flexibility, allowing us to significan tly improve the efficiency of parameter allocation in the input embedding of mul tilingual models. By reallocating the input embedding parameters in the Transfor mer layers, we achieve dramatically better performance on standard natural langu age understanding tasks with the same number of parameters during fine-tuning. W e also show that allocating additional capacity to the output embedding provides benefits to the model that persist through the fine-tuning stage even though th e output embedding is discarded after pre-training. Our analysis shows that larg er output embeddings prevent the model's last layers from overspecializing to th e pre-training task and encourage Transformer representations to be more general and more transferable to other tasks and languages. Harnessing these findings, we are able to train models that achieve strong performance on the XTREME benchm ark without increasing the number of parameters at the fine-tuning stage.
****

Recursive Neighborhood Pooling for Graph Representation Learning
Behrooz Tahmasebi,Stefanie Jegelka
While massage passing based Graph Neural Networks (GNNs) have become increasingl y popular architectures for learning with graphs, recent works have revealed imp

ortant shortcomings in their expressive power. In response, several higher-order GNNs have been proposed, which substantially increase the expressive power, but at a large computational cost.
Motivated by this gap, we introduce and analyze a new recursive pooling technique of local neighborhoods that allows different tradeoffs of computational cost and expressive power. First, we show that this model can count subgraphs of size k, and thereby overcomes a known limitation of low-order GNNs. Second, we prove that, in several cases, RNP-GNNs can greatly reduce computational complexity compared to the existing higher-order k-GNN and Local Relational Pooling (LRP) networks.

****
Addressing Distribution Shift in Online Reinforcement Learning with Offline Datasets
Seunghyun Lee,Younggyo Seo,Kimin Lee,Pieter Abbeel,Jinwoo Shin
Recent progress in offline reinforcement learning (RL) has made it possible to train strong RL agents from previously-collected, static datasets. However, depending on the quality of the trained agents and the application being considered, it is often desirable to improve such offline RL agents with further online interaction. As it turns out, fine-tuning offline RL agents is a non-trivial challenge, due to distribution shift – the agent encounters out-of-distribution samples during online interaction, which may cause bootstrapping error in Q-learning and instability during fine-tuning. In order to address the issue, we present a simple yet effective framework, which incorporates a balanced replay scheme and an ensemble distillation scheme. First, we propose to keep separate offline and online replay buffers, and carefully balance the number of samples from each buffer during updates. By utilizing samples from a wider distribution, i.e., both online and offline samples, we stabilize the Q-learning. Next, we present an ensemble distillation scheme, where we train an ensemble of independent actor-critic agents, then distill the policies into a single policy. In turn, we improve the policy using the Q-ensemble during fine-tuning, which allows the policy updates to be more robust to error in each individual Q-function. We demonstrate the superiority of our method on MuJoCo datasets from the recently proposed D4RL benchmark suite.

****
Stable Weight Decay Regularization
Zeke Xie,Issei Sato,Masashi Sugiyama
Weight decay is a popular regularization technique for training of deep neural networks. Modern deep learning libraries mainly use L_{2} regularization as the default implementation of weight decay. \citet{loshchilov2018decoupled} demonstrated that L_{2} regularization is not identical to weight decay for adaptive gradient methods, such as Adaptive Momentum Estimation (Adam), and proposed Adam with Decoupled Weight Decay (AdamW). However, we found that the popular implementations of weight decay, including L_{2} regularization and decoupled weight decay, in modern deep learning libraries usually damage performance. First, the L_{2} regularization is unstable weight decay for all optimizers that use Momentum, such as stochastic gradient descent (SGD). Second, decoupled weight decay is highly unstable for all adaptive gradient methods. We further propose the Stable Weight Decay (SWD) method to fix the unstable weight decay problem from a dynamical perspective. The proposed SWD method makes significant improvements over L_{2} regularization and decoupled weight decay in our experiments. Simply fixing weight decay in Adam by SWD, with no extra hyperparameter, can outperform complex Adam variants, which have more hyperparameters.
****
Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity
JangHyun Kim,Wonho Choo,Hosan Jeong,Hyun Oh Song
While deep neural networks show great performance on fitting to the training distribution, improving the networks' generalization performance to the test distribution and robustness to the sensitivity to input perturbations still remain as

a challenge. Although a number of mixup based augmentation strategies have been proposed to partially address them, it remains unclear as to how to best utilize the supervisory signal within each input data for mixup from the optimization perspective. We propose a new perspective on batch mixup and formulate the optimal construction of a batch of mixup data maximizing the data saliency measure of each individual mixup data and encouraging the supermodular diversity among the constructed mixup data. This leads to a novel discrete optimization problem minimizing the difference between submodular functions. We also propose an efficient modular approximation based iterative submodular minimization algorithm for efficient mixup computation per each minibatch suitable for minibatch based neural network training. Our experiments show the proposed method achieves the state of the art generalization, calibration, and weakly supervised localization results compared to other mixup methods. The source code is available at https://github.com/snu-mllab/Co-Mixup.
****

Physics-aware, probabilistic model order reduction with guaranteed stability
Sebastian Kaltenbach,Phaedon Stelios Koutsourelakis
Given (small amounts of) time-series' data from a high-dimensional, fine-grained, multiscale dynamical system, we propose a generative framework for learning an effective, lower-dimensional, coarse-grained dynamical model that is predictive of the fine-grained system's long-term evolution but also of its behavior under different initial conditions.
We target fine-grained models as they arise in physical applications (e.g. molecular dynamics, agent-based models), the dynamics of which are strongly non-stationary but their transition to equilibrium is governed by unknown slow processes which are largely inaccessible by brute-force simulations.
Approaches based on domain knowledge heavily rely on physical insight in identifying temporally slow features and fail to enforce the long-term stability of the learned dynamics. On the other hand, purely statistical frameworks lack interpretability and rely on large amounts of expensive simulation data (long and multiple trajectories) as they cannot infuse domain knowledge.
The generative framework proposed achieves the aforementioned desiderata by employing a flexible prior on the complex plane for the latent, slow processes, and an intermediate layer of physics-motivated latent variables that reduces reliance on data and imbues inductive bias. In contrast to existing schemes, it does not require the a priori definition of projection operators from the fine-grained description and addresses simultaneously the tasks of dimensionality reduction and model estimation.
We demonstrate its efficacy and accuracy in multiscale physical systems of particle dynamics where probabilistic, long-term predictions of phenomena not contained in the training data are produced.
****

Improving the accuracy of neural networks in analog computing-in-memory systems by a generalized quantization method
Lingjun Dai,Qingtian Zhang,Huaqiang Wu
Crossbar-enabled analog computing-in-memory (CACIM) systems can significantly improve the computation speed and energy efficiency of deep neural networks (DNNs). However, the transition of DNN from the digital systems to CACIM systems usually reduces its accuracy. The major issue is that the weights of DNN are stored and calculated directly on analog quantities in CACIM systems. The variation and programming overhead of the analog weight limit the precision.
Therefore, a suitable quantization algorithm is important when deploying a DNN into CACIM systems to obtain less accuracy loss. The analog weight has its unique advantages when doing quantization. Because there is no encoding and decoding process, the set of quanta will not affect the computing process. Therefore, a generalized quantization method that does not constrain the range of quanta and can obtain less quantization error will be effective in CACIM systems. For the first time, we introduced a generalized quantization method into CACIM systems and showed superior performance on a series of computer vision tasks, such as image classification, object detection, and semantic segmentation. Using the generaliz

ed quantization method, the DNN with 8-level analog weights can outperform the 3 2-bit networks. With fewer levels, the generalized quantization method can obtai n less accuracy loss than other uniform quantization methods.
****

Learning Binary Trees via Sparse Relaxation
Valentina Zantedeschi,Matt Kusner,Vlad Niculae
One of the most classical problems in machine learning is how to learn binary tr ees that split data into meaningful partitions. From classification/regression v ia decision trees to hierarchical clustering, binary trees are useful because th ey (a) are often easy to visualize; (b) make computationally-efficient predictio ns; and (c) allow for flexible partitioning. Because of this there has been exte nsive research on how to learn such trees. Optimization generally falls into one of three categories: 1. greedy node-by-node optimization; 2. probabilistic rela xations for differentiability; 3. mixed-integer programming (MIP). Each of these have downsides: greedy can myopically choose poor splits, probabilistic relaxat ions do not have principled ways to prune trees, MIP methods can be slow on larg e problems and may not generalize. In this work we derive a novel sparse relaxat ion for binary tree learning. By sparsely relaxing a new MIP, our approach is ab le to learn tree splits and tree pruning using state-of-the-art gradient-based a pproaches. We demonstrate how our approach is easily visualizable, is efficient, and is competitive with current work in classification/regression and hierarchi cal clustering.
****

Disentangling 3D Prototypical Networks for Few-Shot Concept Learning
Mihir Prabhudesai,Shamit Lal,Darshan Patil,Hsiao-Yu Tung,Adam W Harley,Katerina Fragkiadaki
We present neural architectures that disentangle RGB-D images into objects’ shap es and styles and a map of the background scene, and explore their applications for few-shot 3D object detection and few-shot concept classification. Our networ ks incorporate architectural biases that reflect the image formation process, 3D geometry of the world scene, and shape-style interplay. They are trained end-t o-end self-supervised by predicting views in static scenes, alongside a small nu mber of 3D object boxes. Objects and scenes are represented in terms of 3D featu re grids in the bottleneck of the network. We show the proposed 3D neural repres entations are compositional: they can generate novel 3D scene feature maps by mi xing object shapes and styles, resizing and adding the resulting object 3D featu re maps over background scene feature maps. We show object detectors trained on hallucinated 3D neural scenes generalize better to novel environments. We show c lassifiers for object categories, color, materials, and spatial relationships tr ained over the disentangled 3D feature sub-spaces generalize better with dramat ically fewer exemplars over the current state-of-the-art, and enable a visual qu estion answering system that uses them as its modules to generalize one-shot to novel objects in the scene.
****

Conditional Generative Modeling for De Novo Hierarchical Multi-Label Functional Protein Design
Tim Kucera,Karsten Michael Borgwardt,Matteo Togninalli,Laetitia Papaxanthos
The availability of vast protein sequence information and rich functional annota tions thereof has a large potential for protein design applications in biomedici ne and synthetic biology. To this date, there exists no method for the general-p urpose design of proteins without any prior knowledge about the protein of inter est, such as costly and rare structure information or seed sequence fragments. H owever, the Gene Ontology (GO) database provides information about the hierarchi cal organisation of protein functions, and thus could inform generative models a bout the underlying complex sequence-function relationships, replacing the need for structural data. We therefore propose to use conditional generative adversar ial networks (cGANs) on the task of fast de novo hierarchical multi-label protei n design. We generate protein sequences exhibiting properties of a large set of molecular functions extracted from the GO database, using a single model and wit hout any prior information. We shed light on efficient conditioning mechanisms a

nd adapted network architectures thanks to a thorough hyperparameter selection p
rocess and analysis. We further provide statistically- and biologically-driven e
valuation measures for generative models in the context of protein design to ass
ess the quality of the generated sequences and facilitate progress in the field.
 We show that our proposed model, ProteoGAN, outperforms several baselines when
designing proteins given a functional label and generates well-formed sequences.
****

Sequence Metric Learning as Synchronization of Recurrent Neural Networks
Paul Compagnon,Grégoire Lefebvre,Stefan Duffner,Christophe Garcia
Sequence metric learning is becoming a widely adopted approach for various appli
cations dealing with sequential multi-variate data such as activity recognition
or natural language processing and is most of the time tackled with sequence ali
gnment approaches or representation learning.
In this paper, we propose to study this subject from the point of view of dynami
cal system theory by drawing the analogy between synchronized trajectories produ
ced by dynamical systems and the distance between similar sequences processed by
 a siamese recurrent neural network.
Indeed, a siamese recurrent network comprises two identical sub-networks, two id
entical dynamical systems which can theoretically achieve complete synchronizati
on if a coupling is introduced between them.
We therefore propose a new neural network model that implements this coupling wi
th a new gate integrated into the classical Gated Recurrent Unit architecture. T
his model is thus able to simultaneously learn a similarity metric and the synch
ronization of unaligned multi-variate sequences in a weakly supervised way.
Our experiments show that introducing such a coupling improves the performance o
f the siamese Gated Recurrent Unit architecture on an activity recognition datas
et.
****

Time Series Counterfactual Inference with Hidden Confounders
Guangyu Li,Jiahao Chen,Samuel A Assefa,Yan Liu
We present augmented counterfactual ordinary differential equations (ACODEs), a
new approach to counterfactual inference on time series data with a focus on hea
lthcare applications. ACODEs model interventions in continuous time with differe
ntial equations, augmented by auxiliary confounding variables to reduce inferenc
e bias. Experiments on tumor growth simulation and sepsis patient treatment resp
onse show that ACODEs outperform other methods like counterfactual Gaussian proc
esses, recurrent marginal structural networks, and time series deconfounders in
the accuracy of counterfactual inference. The learned auxiliary variables also r
eveal new insights into causal interventions and hidden confounders.
****

Multi-Head Attention: Collaborate Instead of Concatenate
Jean-Baptiste Cordonnier,Andreas Loukas,Martin Jaggi
Attention layers are widely used in natural language processing (NLP) and are be
ginning to influence computer vision architectures. However, they suffer from ov
er-parameterization. For instance, it was shown that the majority of attention h
eads could be pruned without impacting accuracy. This work aims to enhance curre
nt understanding on how multiple heads interact. Motivated by the observation th
at trained attention heads share common key/query projections, we propose a coll
aborative multi-head attention layer that enables heads to learn shared projecti
ons. Our scheme decreases the number of parameters in an attention layer and can
 be used as a drop-in replacement in any transformer architecture.For instance,
by allowing heads to collaborate on a neural machine translation task, we can re
duce the key dimension by 4× without any loss in performance. We also show that
it is possible to re-parametrize a pre-trained multi-head attention layer into o
ur collaborative attention layer. Even without retraining, collaborative multi-h
ead attention manages to reduce the size of the key and query projections by hal
f without sacrificing accuracy. Our code is public.
****

Secure Byzantine-Robust Machine Learning
Lie He,Sai Praneeth Karimireddy,Martin Jaggi

Increasingly machine learning systems are being deployed to edge servers and dev
ices (e.g. mobile phones) and trained in a collaborative manner. Such distribute
d/federated/decentralized training raises a number of concerns about the robustn
ess, privacy, and security of the procedure. While extensive work has been done
in tackling with robustness, privacy, or security individually, their combinatio
n has rarely been studied. In this paper, we propose a secure multi-server proto
col that offers both input privacy and Byzantine-robustness. In addition, this p
rotocol is communication-efficient, fault-tolerant, and enjoys local differentia
l privacy.
****

LiftPool: Bidirectional ConvNet Pooling
Jiaojiao Zhao,Cees G. M. Snoek
Pooling is a critical operation in convolutional neural networks for increasing
receptive fields and improving robustness to input variations. Most existing poo
ling operations downsample the feature maps, which is a lossy process. Moreov
er, they are not invertible: upsampling a downscaled feature map can not recover
the lost information in the downsampling. By adopting the philosophy of the cl
assical Lifting Scheme from signal processing, we propose LiftPool for bidirecti
onal pooling layers, including LiftDownPool and LiftUpPool. LiftDownPool decomp
oses a feature map into various downsized sub-bands, each of which contains inf
ormation with different frequencies. As the pooling function in LiftDownPool is
perfectly invertible, by performing LiftDownPool backward, a corresponding up-po
oling layer LiftUpPool is able to generate a refined upsampled feature map using
the detail subbands, which is useful for image-to-image translation challenges.
Experiments show the proposed methods achieve better results on image classifi
cation and semantic segmentation, using various backbones. Moreover, LiftDownPo
ol offers better robustness to input corruptions and perturbations.
****

Latent Convergent Cross Mapping
Edward De Brouwer,Adam Arany,Jaak Simm,Yves Moreau
Discovering causal structures of temporal processes is a major tool of scientifi
c inquiry because it helps us better understand and explain the mechanisms drivi
ng a phenomenon of interest, thereby facilitating analysis, reasoning, and synth
esis for such systems.
However, accurately inferring causal structures within a phenomenon based on obs
ervational data only is still an open problem. Indeed, this type of data usually
consists in short time series with missing or noisy values for which causal inf
erence is increasingly difficult. In this work, we propose a method to uncover c
ausal relations in chaotic dynamical systems from short, noisy and sporadic time
series (that is, incomplete observations at infrequent and irregular intervals)
where the classical convergent cross mapping (CCM) fails. Our method works by l
earning a Neural ODE latent process modeling the state-space dynamics of the tim
e series and by checking the existence of a continuous map between the resulting
processes. We provide theoretical analysis and show empirically that Latent-CCM
can reliably uncover the true causal pattern, unlike traditional methods.
****

You Only Need Adversarial Supervision for Semantic Image Synthesis
Edgar Schönfeld,Vadim Sushko,Dan Zhang,Juergen Gall,Bernt Schiele,Anna Khoreva
Despite their recent successes, GAN models for semantic image synthesis still su
ffer from poor image quality when trained with only adversarial supervision. His
torically, additionally employing the VGG-based perceptual loss has helped to ov
ercome this issue, significantly improving the synthesis quality, but at the sam
e time limiting the progress of GAN models for semantic image synthesis. In this
work, we propose a novel, simplified GAN model, which needs only adversarial su
pervision to achieve high quality results. We re-design the discriminator as a s
emantic segmentation network, directly using the given semantic label maps as th
e ground truth for training. By providing stronger supervision to the discrimina
tor as well as to the generator through spatially- and semantically-aware discri
minator feedback, we are able to synthesize images of higher fidelity with bette
r alignment to their input label maps, making the use of the perceptual loss sup

erfluous. Moreover, we enable high-quality multi-modal image synthesis through g lobal and local sampling of a 3D noise tensor injected into the generator, which allows complete or partial image change. We show that images synthesized by our model are more diverse and follow the color and texture distributions of real i mages more closely. We achieve an average improvement of 6 FID and 5 mIoU po ints over the state of the art across different datasets using only adversarial supervision.
****
MARS: Markov Molecular Sampling for Multi-objective Drug Discovery
Yutong Xie,Chence Shi,Hao Zhou,Yuwei Yang,Weinan Zhang,Yong Yu,Lei Li
Searching for novel molecules with desired chemical properties is crucial in dru g discovery. Existing work focuses on developing neural models to generate eithe r molecular sequences or chemical graphs. However, it remains a big challenge to find novel and diverse compounds satisfying several properties. In this paper, we propose MARS, a method for multi-objective drug molecule discovery. MARS is b ased on the idea of generating the chemical candidates by iteratively editing fr agments of molecular graphs. To search for high-quality candidates, it employs M arkov chain Monte Carlo sampling (MCMC) on molecules with an annealing scheme an d an adaptive proposal. To further improve sample efficiency, MARS uses a graph neural network (GNN) to represent and select candidate edits, where the GNN is t rained on-the-fly with samples from MCMC. Experiments show that MARS achieves st ate-of-the-art performance in various multi-objective settings where molecular b io-activity, drug-likeness, and synthesizability are considered. Remarkably, in the most challenging setting where all four objectives are simultaneously optimi zed, our approach outperforms previous methods significantly in comprehensive ev aluations. The code is available at https://github.com/yutxie/mars.
****
Unsupervised Active Pre-Training for Reinforcement Learning
Hao Liu,Pieter Abbeel
We introduce a new unsupervised pre-training method for reinforcement learning c alled $\textbf{APT}$, which stands for $\textbf{A}\text{ctive}\textbf{P}\text{re -}\textbf{T}\text{raining}$. APT learns a representation and a policy initializa tion by actively searching for novel states in reward-free environments. We use the contrastive learning framework for learning the representation from collecte d transitions. The key novel idea is to collect data during pre-training by maxi mizing a particle based entropy computed in the learned latent representation sp ace. By doing particle based entropy maximization, we alleviate the need for cha llenging density modeling and are thus able to scale our approach to image obser vations. APT successfully learns meaningful representations as well as policy in itializations without using any reward. We empirically evaluate APT on the Atari game suite and DMControl suite by exposing task-specific reward to agent after a long unsupervised pre-training phase. On Atari games, APT achieves human-level performance on 12 games and obtains highly competitive performance compared t o canonical fully supervised RL algorithms. On DMControl suite, APT beats all ba selines in terms of asymptotic performance and data efficiency and dramatically improves performance on tasks that are extremely difficult for training from scr atch. Importantly, the pre-trained models can be fine-tuned to solve different t asks as long as the environment does not change. Finally, we also pre-train mult i-environment encoders on data from multiple environments and show generalizatio n to a broad set of RL tasks.
****
The Emergence of Individuality in Multi-Agent Reinforcement Learning
Jiechuan Jiang,Zongqing Lu
Individuality is essential in human society, which induces the division of labor and thus improves the efficiency and productivity. Similarly, it should also be a key to multi-agent cooperation. Inspired by that individuality is of being an individual separate from others, we propose a simple yet efficient method for t he emergence of individuality (EOI) in multi-agent reinforcement learning (MARL) . EOI learns a probabilistic classifier that predicts a probability distribution over agents given their observation and gives each agent an intrinsic reward of

being correctly predicted by the classifier. The intrinsic reward encourages th
e agents to visit their own familiar observations, and learning the classifier b
y such observations makes the intrinsic reward signals stronger and in turn make
s the agents more identifiable. To further enhance the intrinsic reward and prom
ote the emergence of individuality, two regularizers are proposed to increase th
e discriminability of the classifier. We implement EOI on top of popular MARL al
gorithms. Empirically, we show that EOI outperforms existing methods in a variet
y of multi-agent cooperative scenarios.
****
Bounded Myopic Adversaries for Deep Reinforcement Learning Agents
Ezgi Korkmaz,Henrik Sandberg,Gyorgy Dan
Adversarial attacks against deep neural networks have been widely studied. Adver
sarial examples for deep reinforcement learning (DeepRL) have significant securi
ty implications, due to the deployment of these algorithms in many application d
omains. In this work we formalize an optimal myopic adversary for deep reinforce
ment learning agents. Our adversary attempts to find a bounded perturbation of t
he state which minimizes the value of the action taken by the agent. We show wit
h experiments in various games in the Atari environment that our attack formulat
ion achieves significantly larger impact as compared to the current state-of-th
e-art. Furthermore, this enables us to lower the bounds by several orders of mag
nitude on the perturbation needed to efficiently achieve significant impacts on
DeepRL agents.
****
On Self-Supervised Image Representations for GAN Evaluation
Stanislav Morozov,Andrey Voynov,Artem Babenko
The embeddings from CNNs pretrained on Imagenet classification are de-facto stan
dard image representations for assessing GANs via FID, Precision and Recall meas
ures. Despite broad previous criticism of their usage for non-Imagenet domains,
these embeddings are still the top choice in most of the GAN literature.

In this paper, we advocate the usage of the state-of-the-art self-supervised rep
resentations to evaluate GANs on the established non-Imagenet benchmarks. These
representations, typically obtained via contrastive learning, are shown to provi
de better transfer to new tasks and domains, therefore, can serve as more univer
sal embeddings of natural images. With extensive comparison of the recent GANs o
n the common datasets, we show that self-supervised representations produce a mo
re reasonable ranking of models in terms of FID/Precision/Recall, while the rank
ing with classification-pretrained embeddings often can be misleading.
****
A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Expon
entially Favors Flat Minima
Zeke Xie,Issei Sato,Masashi Sugiyama
Stochastic Gradient Descent (SGD) and its variants are mainstream methods for tr
aining deep networks in practice. SGD is known to find a flat minimum that often
generalizes well. However, it is mathematically unclear how deep learning can s
elect a flat minimum among so many minima. To answer the question quantitatively
, we develop a density diffusion theory to reveal how minima selection quantitat
ively depends on the minima sharpness and the hyperparameters. To the best of ou
r knowledge, we are the first to theoretically and empirically prove that, benef
ited from the Hessian-dependent covariance of stochastic gradient noise, SGD fav
ors flat minima exponentially more than sharp minima, while Gradient Descent (GD
) with injected white noise favors flat minima only polynomially more than sharp
minima. We also reveal that either a small learning rate or large-batch trainin
g requires exponentially many iterations to escape from minima in terms of the r
atio of the batch size and learning rate. Thus, large-batch training cannot sear
ch flat minima efficiently in a realistic computational time.
****
Robust Learning of Fixed-Structure Bayesian Networks in Nearly-Linear Time
Yu Cheng,Honghao Lin
We study the problem of learning Bayesian networks where an ϵ-fraction

of the samples are adversarially corrupted. We focus on the fully-observable ca se where the underlying graph structure is known. In this work, we present the first nearly-linear time algorithm for this problem with a dimension-independent error guarantee. Previous robust algorithms with comparable error guarantees a re slower by at least a factor of (d/ϵ), where d is the number of var iables in the Bayesian network and ϵ is the fraction of corrupted sampl es.

Our algorithm and analysis are considerably simpler than those in previous work. We achieve this by establishing a direct connection between robust learning of Bayesian networks and robust mean estimation. As a subroutine in our algorithm , we develop a robust mean estimation algorithm whose runtime is nearly-linear i n the number of nonzeros in the input samples, which may be of independent inter est.
****
Contrastive Explanations for Reinforcement Learning via Embedded Self Prediction s
Zhengxian Lin,Kin-Ho Lam,Alan Fern
We investigate a deep reinforcement learning (RL) architecture that supports exp laining why a learned agent prefers one action over another. The key idea is to learn action-values that are directly represented via human-understandable prope rties of expected futures. This is realized via the embedded self-prediction (ES P) model, which learns said properties in terms of human provided features. Acti on preferences can then be explained by contrasting the future properties predic ted for each action. To address cases where there are a large number of features , we develop a novel method for computing minimal sufficient explanations from a n ESP. Our case studies in three domains, including a complex strategy game, sho w that ESP models can be effectively learned and support insightful explanations .
****
Explore with Dynamic Map: Graph Structured Reinforcement Learning
Jiarui Jin,Sijin Zhou,Weinan Zhang,Rasool Fakoor,David Wipf,Tong He,Yong Yu,Zhen g Zhang,Alex Smola
In reinforcement learning, a map with states and transitions built based on hist orical trajectories is often helpful in exploration and exploitation. Even so, l earning and planning on such a map within a sparse environment remains a challen ge. As a step towards this goal, we propose Graph Structured Reinforcement Learn ing (GSRL), which utilizes historical trajectories to slowly adjust exploration directions and learn related experiences while rapidly updating the value functi on estimation. GSRL constructs a dynamic graph on top of state transitions in th e replay buffer based on historical trajectories, and develops an attention stra tegy on the map to select an appropriate goal direction, which decomposes the ta sk of reaching a distant goal state into a sequence of easier tasks. We also lev erage graph structure to sample related trajectories for efficient value learnin g. Results demonstrate that GSRL can outperform the state-of-the-art algorithms in terms of sample efficiency on benchmarks with sparse reward functions.
****
Hybrid Discriminative-Generative Training via Contrastive Learning
Hao Liu,Pieter Abbeel
Contrastive learning and supervised learning have both seen significant progress and success. However, thus far they have largely been treated as two separate o bjectives, brought together only by having a shared neural network. In this pape r we show that through the perspective of hybrid discriminative-generative train ing of energy-based models we can make a direct connection between contrastive l earning and supervised learning. Beyond presenting this unified view, we show ou r specific choice of approximation of the energy-based loss significantly improv es energy-based models and contrastive learning based methods in confidence-cali bration, out-of-distribution detection, adversarial robustness, generative model ing, and image classification tasks. In addition to significantly improved perfo rmance, our method also gets rid of SGLD training and does not suffer from train

ing instability. Our evaluations also demonstrate that our method performs bette r than or on par with state-of-the-art hand-tailored methods in each task.
****
Small Input Noise is Enough to Defend Against Query-based Black-box Attacks
Junyoung Byun,Hyojun Go,Changick Kim
While deep neural networks show unprecedented performance in various tasks, the vulnerability to adversarial examples hinders their deployment in safety-critica l systems. Many studies have shown that attacks are also possible even in a bla ck-box setting where an adversary cannot access the target model's internal info rmation. Most black-box attacks are based on queries, each of which obtains the target model's output for an input, and many recent studies focus on reducing th e number of required queries. In this paper, we pay attention to an implicit ass umption of these attacks that the target model's output exactly corresponds to t he query input. If some randomness is introduced into the model to break this as sumption, query-based attacks may have tremendous difficulty in both gradient es timation and local search, which are the core of their attack process. From this motivation, we observe even a small additive input noise can neutralize most qu ery-based attacks and name this simple yet effective approach Small Noise Defens e (SND). We analyze how SND can defend against query-based black-box attacks and demonstrate its effectiveness against eight different state-of-the-art attacks with CIFAR-10 and ImageNet datasets. Even with strong defense ability, SND almos t maintains the original clean accuracy and computational speed. SND is readily applicable to pre-trained models by adding only one line of code at the inferenc e stage, so we hope that it will be used as a baseline of defense against query-based black-box attacks in the future.
****
RetCL: A Selection-based Approach for Retrosynthesis via Contrastive Learning
Hankook Lee,Sungsoo Ahn,Seung-Woo Seo,You Young Song,Eunho Yang,Sung Ju Hwang,Ji nwoo Shin
Retrosynthesis, of which the goal is to find a set of reactants for synthesizing a target product, is an emerging research area of deep learning. While the exis ting approaches have shown promising results, they currently lack the ability to consider availability (e.g., stability or purchasability) of the reactants or g eneralize to unseen reaction templates (i.e., chemical reaction rules).
In this paper, we propose a new approach that mitigates the issues by reformulat ing retrosynthesis into a selection problem of reactants from a candidate set of commercially available molecules. To this end, we design an efficient reactant selection framework, named RetCL (retrosynthesis via contrastive learning), for enumerating all of the candidate molecules based on selection scores computed by graph neural networks. For learning the score functions, we also propose a nove l contrastive training scheme with hard negative mining. Extensive experiments d emonstrate the benefits of the proposed selection-based approach. For example, w hen all 671k reactants in the USPTO database are given as candidates, our RetCL achieves top-1 exact match accuracy of 71.3% for the USPTO-50k benchmark, while a recent transformer-based approach achieves 59.6%. We also demonstrate that Ret CL generalizes well to unseen templates in various settings in contrast to templ ate-based approaches. The code will be released.
****
Learning a Max-Margin Classifier for Cross-Domain Sentiment Analysis
Mohammad Rostami,Aram Galstyan
Sentiment analysis is a costly yet necessary task for enterprises to study the opinions of their costumers to improve their products and services and to determ ine optimal marketing strategies. Due to existence of a wide range of domains across different products and services, cross-domain sentiment analysis methods have received significant attention in recent years. These methods mitigate the domain gap between different applications by training cross-domain generalizable classifiers which help to relax the need for individual data annotation per eac h domain. Most existing methods focus on learning domain-agnostic representation s that are invariant with respect to both the source and the target domains. As a result, a classifier that is trained using annotated data in a source domain,

would generalize well in a related target domain. In this work, we introduce a new domain adaptation method which induces large margins between different classes in an embedding space based on the notion of prototypical distribution. This embedding space is trained to be domain-agnostic by matching the data distributions across the domains. Large margins in the source domain help to reduce the effect of ``domain shift'' on the performance of a trained classifier in the target domain. Theoreticaland empirical analysis are provided to demonstrate that the method is effective.
****
Activation-level uncertainty in deep neural networks
Pablo Morales-Alvarez,Daniel Hernández-Lobato,Rafael Molina,José Miguel Hernández-Lobato
Current approaches for uncertainty estimation in deep learning often produce too confident results. Bayesian Neural Networks (BNNs) model uncertainty in the space of weights, which is usually high-dimensional and limits the quality of variational approximations. The more recent functional BNNs (fBNNs) address this only partially because, although the prior is specified in the space of functions, the posterior approximation is still defined in terms of stochastic weights. In this work we propose to move uncertainty from the weights (which are deterministic) to the activation function. Specifically, the activations are modelled with simple 1D Gaussian Processes (GP), for which a triangular kernel inspired by the ReLu non-linearity is explored. Our experiments show that activation-level stochasticity provides more reliable uncertainty estimates than BNN and fBNN, whereas it performs competitively in standard prediction tasks. We also study the connection with deep GPs, both theoretically and empirically. More precisely, we show that activation-level uncertainty requires fewer inducing points and is better suited for deep architectures.
****
Consensus Clustering with Unsupervised Representation Learning
Jayanth Reddy Regatti,Aniket Anand Deshmukh,Eren Manavoglu,Urun Dogan
Recent advances in deep clustering and unsupervised representation learning are based on the idea that different views of an input image (generated through data augmentation techniques) must either be closer in the representation space, or have a similar cluster assignment. In this work, we leverage this idea together with ensemble learning to perform clustering and representation learning. Ensemble learning is widely used in the supervised learning setting but has not yet been practical in deep clustering. Previous works on ensemble learning for clustering neither work on the feature space nor learn features. We propose a novel ensemble learning algorithm dubbed Consensus Clustering with Unsupervised Representation Learning (ConCURL) which learns representations by creating a consensus on multiple clustering outputs. Specifically, we generate a cluster ensemble using random transformations on the embedding space, and define a consensus loss function that measures the disagreement among the constituents of the ensemble. Thus, diverse ensembles minimize this loss function in a synergistic way, which leads to better representations that work with all cluster ensemble constituents. Our proposed method ConCURL is easy to implement and integrate into any representation learning or deep clustering block. ConCURL outperforms all state of the art methods on various computer vision datasets. Specifically, we beat the closest state of the art method by 5.9 percent on the ImageNet-10 dataset, and by 18 percent on the ImageNet-Dogs dataset in terms of clustering accuracy. We further shed some light on the under-studied overfitting issue in clustering and show that our method does not overfit as much as existing methods, and thereby generalizes better for new data samples.
****
Reconnaissance for reinforcement learning with safety constraints
Shin-ichi Maeda,Hayato Watahiki,Yi Ouyang,Shintarou Okada,Masanori Koyama
Practical reinforcement learning problems are often formulated as constrained Markov decision process (CMDP) problems, in which the agent has to maximize the expected return while satisfying a set of prescribed safety constraints. In this study, we consider a situation in which the agent has access to the generative mo

del which provides us with a next state sample for any given state-action pair, and propose a model to solve a CMDP problem by decomposing the CMDP into a pair of MDPs; \textit{reconnaissance} MDP (R-MDP) and \textit{planning} MDP (P-MDP). In R-MDP, we train threat function, the Q-function analogue of danger that can d etermine whether a given state-action pair is safe or not. In P-MDP, we train a reward-seeking policy while using a fixed threat function to determine the safen ess of each action. With the help of generative model, we can efficiently train the threat function by preferentially sampling rare dangerous events. Once the t hreat function for a baseline policy is computed, we can solve other CMDP proble ms with different reward and different danger-constraint without the need to re-train the model. We also present an efficient approximation method for the threa t function that can greatly reduce the difficulty of solving R-MDP. We will demo nstrate the efficacy of our method over classical approaches in benchmark datase t and complex collision-free navigation tasks.
****

SkipW: Resource Adaptable RNN with Strict Upper Computational Limit
Tsiry Mayet,Anne Lambert,Pascal Leguyadec,Francoise Le Bolzer,François Schnitzle r
We introduce Skip-Window, a method to allow recurrent neural networks (RNNs) to trade off accuracy for computational cost during the analysis of a sequence. Sim ilarly to existing approaches, Skip-Window extends existing RNN cells by adding a mechanism to encourage the model to process fewer inputs. Unlike existing appr oaches, Skip-Window is able to respect a strict computational budget, making thi s model more suitable for limited hardware. We evaluate this approach on two dat asets: a human activity recognition task and adding task. Our results show that Skip-Window is able to exceed the accuracy of existing approaches for a lower co mputational cost while strictly limiting said cost.
****

Wasserstein-2 Generative Networks
Alexander Korotin,Vage Egiazarian,Arip Asadulaev,Alexander Safin,Evgeny Burnaev
We propose a novel end-to-end non-minimax algorithm for training optimal transpo rt mappings for the quadratic cost (Wasserstein-2 distance). The algorithm uses input convex neural networks and a cycle-consistency regularization to approxima te Wasserstein-2 distance. In contrast to popular entropic and quadratic regular izers, cycle-consistency does not introduce bias and scales well to high dimensi ons. From the theoretical side, we estimate the properties of the generative map ping fitted by our algorithm. From the practical side, we evaluate our algorithm on a wide range of tasks: image-to-image color transfer, latent space optimal t ransport, image-to-image style transfer, and domain adaptation.
****

QTRAN++: Improved Value Transformation for Cooperative Multi-Agent Reinforcement Learning
Kyunghwan Son,Sungsoo Ahn,Roben D. Delos Reyes,Jinwoo Shin,Yung Yi
QTRAN is a multi-agent reinforcement learning (MARL) algorithm capable of learni ng the largest class of joint-action value functions up to date. However, despit e its strong theoretical guarantee, it has shown poor empirical performance in c omplex environments, such as Starcraft Multi-Agent Challenge (SMAC). In this pap er, we identify the performance bottleneck of QTRAN and propose a substantially improved version, coined QTRAN++. Our gains come from (i) stabilizing the traini ng objective of QTRAN, (ii) removing the strict role separation between the acti on-value estimators of QTRAN, and (iii) introducing a multi-head mixing network for value transformation. Through extensive evaluation, we confirm that our diag nosis is correct, and QTRAN++ successfully bridges the gap between empirical per formance and theoretical guarantee. In particular, QTRAN++ newly achieves state-of-the-art performance in the SMAC environment. The code will be released.
****

Non-iterative Parallel Text Generation via Glancing Transformer
Lihua Qian,Hao Zhou,Yu Bao,Mingxuan Wang,Lin Qiu,Weinan Zhang,Yong Yu,Lei Li
Although non-autoregressive models with one-iteration generation achieve remarka ble inference speed-up, they still fall behind their autoregressive counterparts

in prediction accuracy. The non-autoregressive models with the best accuracy currently rely on multiple decoding iterations, which largely sacrifice the inference speed of non-autoregressive models. Inspired by the way of learning word dependencies in autoregressive and iterative-decoding models, we propose Glancing Transformer (GLAT) with a glancing language model (GLM), which learns to capture the word dependency gradually. Experiments on three benchmarks demonstrate that our approach can significantly improve the accuracy of non-autoregressive models without multiple decoding iterations. In particular, GLAT achieves state-of-the-art results among non-iterative models and even outperforms top iterative counterparts in some specific benchmarks.

****

Offline Meta-Reinforcement Learning with Advantage Weighting

Eric Mitchell,Rafael Rafailov,Xue Bin Peng,Sergey Levine,Chelsea Finn

This paper introduces the offline meta-reinforcement learning (offline meta-RL) problem setting and proposes an algorithm that performs well in this setting. Offline meta-RL is analogous to the widely successful supervised learning strategy of pre-training a model on a large batch of fixed, pre-collected data (possibly from various tasks) and fine-tuning the model to a new task with relatively little data. That is, in offline meta-RL, we meta-train on fixed, pre-collected data from several tasks and adapt to a new task with a very small amount (less than 5 trajectories) of data from the new task. By nature of being offline, algorithms for offline meta-RL can utilize the largest possible pool of training data available and eliminate potentially unsafe or costly data collection during meta-training. This setting inherits the challenges of offline RL, but it differs significantly because offline RL does not generally consider a) transfer to new tasks or b) limited data from the test task, both of which we face in offline meta-RL. Targeting the offline meta-RL setting, we propose Meta-Actor Critic with Advantage Weighting (MACAW). MACAW is an optimization-based meta-learning algorithm that uses simple, supervised regression objectives for both the inner and outer loop of meta-training. On offline variants of common meta-RL benchmarks, we empirically find that this approach enables fully offline meta-reinforcement learning and achieves notable gains over prior methods.

****

Group Equivariant Stand-Alone Self-Attention For Vision

David W. Romero,Jean-Baptiste Cordonnier

We provide a general self-attention formulation to impose group equivariance to arbitrary symmetry groups. This is achieved by defining positional encodings that are invariant to the action of the group considered. Since the group acts on the positional encoding directly, group equivariant self-attention networks (GSA-Nets) are steerable by nature. Our experiments on vision benchmarks demonstrate consistent improvements of GSA-Nets over non-equivariant self-attention networks.

****

A Simple and Effective Baseline for Out-of-Distribution Detection using Abstention

Sunil Thulasidasan,Sushil Thapa,Sayera Dhaubhadel,Gopinath Chennupati,Tanmoy Bhattacharya,Jeff Bilmes

Refraining from confidently predicting when faced with categories of inputs different from those seen during training is an important requirement for the safe deployment of deep learning systems. While simple to state, this has been a particularly challenging problem in deep learning, where models often end up making overconfident predictions in such situations. In this work we present a simple, but highly effective approach to deal with out-of-distribution detection that uses the principle of abstention: when encountering a sample from an unseen class, the desired behavior is to abstain from predicting. Our approach uses a network with an extra abstention class and is trained on a dataset that is augmented with an uncurated set that consists of a large number of out-of-distribution (OoD) samples that are assigned the label of the abstention class; the model is then trained to learn an effective discriminator between in and out-of-distribution samples.

We compare this relatively simple approach against a wide variety of more complex methods that have been proposed both for out-of-distribution detection as well as uncertainty modeling in deep learning, and empirically demonstrate its effectiveness on a wide variety of of benchmarks and deep architectures for image recognition and text classification, often outperforming existing approaches by significant margins. Given the simplicity and effectiveness of this method, we propose that this approach be used as a new additional baseline for future work in this domain.
****

LIME: Learning Inductive Bias for Primitives of Mathematical Reasoning
Yuhuai Wu,Markus Norman Rabe,Wenda Li,Jimmy Ba,Roger Baker Grosse,Christian Szegedy
While designing inductive bias in neural architectures has been widely studied, we hypothesize that transformer networks are flexible enough to learn inductive bias from suitable generic tasks. Here, we replace architecture engineering by encoding inductive bias in the form of datasets. Inspired by Peirce's view that deduction, induction, and abduction form an irreducible set of reasoning primitives, we design three synthetic tasks that are intended to require the model to have these three abilities. We specifically design these synthetic tasks in a way that they are devoid of mathematical knowledge to ensure that only the fundamental reasoning biases can be learned from these tasks. This defines a new pre-training methodology called ``"LIME" (Learning Inductive bias for Mathematical rEasoning). Models trained with LIME significantly outperform vanilla transformers on three very different large mathematical reasoning benchmarks. Unlike dominating the computation cost as traditional pre-training approaches, LIME requires only a small fraction of the computation cost of the typical downstream task.
****

Lipschitz-Bounded Equilibrium Networks
Max Revay,Ruigang Wang,Ian Manchester
This paper introduces new parameterizations of equilibrium neural networks, i.e. networks defined by implicit equations. This model class includes standard multilayer
and residual networks as special cases. The new parameterization admits
a Lipschitz bound during training via unconstrained optimization, i.e. no projections
or barrier functions are required. Lipschitz bounds are a common proxy for
robustness and appear in many generalization bounds. Furthermore, compared to
previous works we show well-posedness (existence of solutions) under less restrictive
conditions on the network weights and more natural assumptions on the
activation functions: that they are monotone and slope restricted. These results
are proved by establishing novel connections with convex optimization, operator
splitting on non-Euclidean spaces, and contracting neural ODEs. In image classification
experiments we show that the Lipschitz bounds are very accurate and
improve robustness to adversarial attacks.
****

Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization
Alexander Korotin,Lingxiao Li,Justin Solomon,Evgeny Burnaev
Wasserstein barycenters provide a geometric notion of the weighted average of probability measures based on optimal transport. In this paper, we present a scalable algorithm to compute Wasserstein-2 barycenters given sample access to the input measures, which are not restricted to being discrete. While past approaches rely on entropic or quadratic regularization, we employ input convex neural networks and cycle-consistency regularization to avoid introducing bias. As a result, our approach does not resort to minimax optimization. We provide theoretical analysis on error bounds as well as empirical evidence of the effectiveness of the proposed approach in low-dimensional qualitative scenarios and high-dimensional quantitative experiments.

****
PODS: Policy Optimization via Differentiable Simulation
Miguel Angel Zamora Mora,Momchil Peychev,Sehoon Ha,Martin Vechev,Stelian Coros
Current reinforcement learning (RL) methods use simulation models as simple blac k-box oracles. In this paper, with the goal of improving the performance exhibit ed by RL algorithms, we explore a systematic way of leveraging the additional in formation provided by an emerging class of differentiable simulators. Building o n concepts established by Deterministic Policy Gradients (DPG) methods, the neur al network policies learned with our approach represent deterministic actions. I n a departure from standard methodologies, however, learning these policy does n ot hinge on approximations of the value function that must be learned concurrent ly in an actor-critic fashion. Instead, we exploit differentiable simulators to directly compute the analytic gradient of a policy's value function with respect to the actions it outputs. This, in turn, allows us to efficiently perform loca lly optimal policy improvement iterations. Compared against other state-of-the-a rt RL methods, we show that with minimal hyper-parameter tuning our approach con sistently leads to better asymptotic behavior across a set of payload manipulati on tasks that demand high precision.
****
Regioned Episodic Reinforcement Learning
Jiarui Jin,Cong Chen,Ming Zhou,Weinan Zhang,Rasool Fakoor,David Wipf,Yong Yu,Jun Wang,Alex Smola
Goal-oriented reinforcement learning algorithms are often good at exploration, n ot exploitation, while episodic algorithms excel at exploitation, not exploratio n. As a result, neither of these approaches alone can lead to a sample efficient algorithm in complex environments with high dimensional state space and delayed rewards. Motivated by these observations and shortcomings, in this paper, we in troduce Regioned Episodic Reinforcement Learning (RERL) that combines the episod ic and goal-oriented learning strengths and leads to a more sample efficient and ef- fective algorithm. RERL achieves this by decomposing the space into several sub-space regions and constructing regions that lead to more effective explorat ion and high values trajectories. Extensive experiments on various benchmark tas ks show that RERL outperforms existing methods in terms of sample efficiency and final rewards.
****
Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies
Dominik Schmidt,Georgia Koppe,Zahra Monfared,Max Beutelspacher,Daniel Durstewitz
A main theoretical interest in biology and physics is to identify the nonlinear dynamical system (DS) that generated observed time series. Recurrent Neural Netw orks (RNN) are, in principle, powerful enough to approximate any underlying DS, but in their vanilla form suffer from the exploding vs. vanishing gradients prob lem. Previous attempts to alleviate this problem resulted either in more complic ated, mathematically less tractable RNN architectures, or strongly limited the d ynamical expressiveness of the RNN.
Here we address this issue by suggesting a simple regularization scheme for vani lla RNN with ReLU activation which enables them to solve long-range dependency p roblems and express slow time scales, while retaining a simple mathematical stru cture which makes their DS properties partly analytically accessible. We prove t wo theorems that establish a tight connection between the regularized RNN dynami cs and their gradients, illustrate on DS benchmarks that our regularization appr oach strongly eases the reconstruction of DS which harbor widely differing time scales, and show that our method is also en par with other long-range architectu res like LSTMs on several tasks.
****
Semantic-Guided Representation Enhancement for Self-supervised Monocular Trained Depth Estimation
Rui Li,Qing Mao,Pei Wang,Xiantuo He,Yu Zhu,Jinqiu Sun,Yanning Zhang
Self-supervised depth estimation has shown its great effectiveness in producing high quality depth maps given only image sequences as input. However, its perfor

mance usually drops when estimating on border areas or objects with thin structu
res due to the limited depth representation ability. In this paper, we address t
his problem by proposing a semantic-guided depth representation enhancement meth
od, which promotes both local and global depth feature representations by levera
ging rich contextual information. In stead of a single depth network as used in
conventional paradigms, we propose an extra semantic segmentation branch to offe
r extra contextual features for depth estimation. Based on this framework, we en
hance the local feature representation by sampling and feeding the point-based f
eatures that locate on the semantic edges to an individual Semantic-guided Edge
Enhancement module (SEEM), which is specifically designed for promoting depth es
timation on the challenging semantic borders. Then, we improve the global featur
e representation by proposing a semantic-guided multi-level attention mechanism,
 which enhances the semantic and depth features by exploring pixel-wise correlat
ions in the multi-level depth decoding scheme. Extensive experiments validate th
e distinct superiority of our method in capturing highly accurate depth on the c
hallenging image areas such as semantic category borders and thin objects. Both
quantitative and qualitative experiments on KITTI show that our method outperfor
ms the state-of-the-art methods.
****

Memory Augmented Design of Graph Neural Networks
Tao Xiong,Liang Zhu,Ruofan Wu,Yuan Qi
The expressive power of graph neural networks (GNN) has drawn much interest rece
ntly. Most existent work focused on measuring the expressiveness of GNN through
the task of distinguishing between graphs. In this paper, we inspect the represe
ntation limits of locally unordered messaging passing (LUMP) GNN architecture th
rough the lens of \emph{node classification}. For GNNs based on permutation inva
riant local aggregators, we characterize graph-theoretic conditions under which
such GNNs fail to discriminate simple instances, regardless of underlying archit
ecture or network depth. To overcome this limitation, we propose a novel framewo
rk to augment GNNs with global graph information called \emph{memory augmentatio
n}. Specifically, we allow every node in the original graph to interact with a g
roup of memory nodes. For each node, information from all the other nodes in the
 graph can be gleaned through the relay of the memory nodes. For proper backbone
 architectures like GAT and GCN, memory augmented GNNs are theoretically shown t
o be more expressive than LUMP GNNs. Empirical evaluations demonstrate the signi
ficant improvement of memory augmentation. In particular, memory augmented GAT a
nd GCN are shown to either outperform or closely match state-of-the-art performa
nce across various benchmark datasets.
****

RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs
Meng Qu,Junkun Chen,Louis-Pascal Xhonneux,Yoshua Bengio,Jian Tang
This paper studies learning logic rules for reasoning on knowledge graphs. Logic
 rules provide interpretable explanations when used for prediction as well as be
ing able to generalize to other tasks, and hence are critical to learn. Existing
 methods either suffer from the problem of searching in a large search space (e.
g., neural logic programming) or ineffective optimization due to sparse rewards
(e.g., techniques based on reinforcement learning). To address these limitations
, this paper proposes a probabilistic model called RNNLogic. RNNLogic treats log
ic rules as a latent variable, and simultaneously trains a rule generator as wel
l as a reasoning predictor with logic rules. We develop an EM-based algorithm fo
r optimization. In each iteration, the reasoning predictor is updated to explore
 some generated logic rules for reasoning. Then in the E-step, we select a set o
f high-quality rules from all generated rules with both the rule generator and r
easoning predictor via posterior inference; and in the M-step, the rule generato
r is updated with the rules selected in the E-step. Experiments on four datasets
 prove the effectiveness of RNNLogic.
****

ARMCMC: Online Model Parameters full probability Estimation in Bayesian Paradigm
Pedram Agand,Mo Chen,Hamid D. Taghirad
Although the Bayesian paradigm provides a rigorous framework to estimate the ful

l probability distribution over unknown parameters, its online implementation can be challenging due to heavy computational costs. This paper proposes Adaptive Recursive Markov Chain Monte Carlo (ARMCMC) which estimates full probability density of model parameters while alleviating shortcomings of conventional online approaches. These shortcomings include: being solely able to account for Gaussian noise, being applicable to systems with linear in the parameters (LIP) constraint, or having requirements on persistence excitation (PE). In ARMCMC, we propose a variable jump distribution, which depends on a temporal forgetting factor.
This allows one to adjust the trade-off between exploitation and exploration, depending on whether there is an abrupt change to the parameter being estimated. We prove that ARMCMC requires fewer samples to achieve the same precision and reliability compared to conventional MCMC approaches. We demonstrate our approach on two challenging benchmark: the estimation of parameters in a soft bending actuator and the Hunt-Crossley dynamic model. Our method shows at-least 70% improvement in parameter point estimation accuracy and approximately 55% reduction in tracking error of the value of interest compared to recursive least squares and conventional MCMC.

****

Learning to Solve Multi-Robot Task Allocation with a Covariant-Attention based Neural Architecture

Steve Paul,Payam Ghassemi,Souma Chowdhury

This paper presents a new graph neural network architecture over which reinforcement learning can be performed to yield online policies for an important class of multi-robot task allocation (MRTA) problems, one that involves tasks with deadlines, and robots with ferry range and payload constraints and multi-tour capability. While drawing motivation from recent graph learning methods that learn to solve combinatorial optimization problems of the mTSP/VRP type, this paper seeks to provide better convergence and generalizability specifically for MRTA problems. The proposed neural architecture, called Covariant Attention-based Model or CAM, includes three main components: 1) an encoder: a covariant compositional node-based embedding is used to represent each task as a learnable feature vector in manner that preserves the local structure of the task graph while being invariant to the ordering of graph nodes; 2) context: a vector representation of the mission time and state of the concerned robot and its peers; and 2) a decoder: builds upon the attention mechanism to facilitate a sequential output. In order to train the CAM model, a policy-gradient method based on REINFORCE is used. While the new architecture can solve the broad class of MRTA problems stated above, to demonstrate real-world applicability we use a multi-unmanned aerial vehicle or multi-UAV-based flood response problem for evaluation purposes. For comparison, the well-known attention-based approach (designed to solve mTSP/VRP problems) is extended and applied to the MRTA problem, as a baseline. The results show that the proposed CAM method is not only superior to the baseline AM method in terms of the cost function (over training and unseen test scenarios), but also provide significantly faster convergence and yields learnt policies that can be executed within 2.4ms/robot, thereby allowing real-time application.

****

Redesigning the Classification Layer by Randomizing the Class Representation Vectors

Gabi Shalev,Gal Lev Shalev,Yossi Keshet

Neural image classification models typically consist of two components. The first is an image encoder, which is responsible for encoding a given raw image into a representative vector. The second is the classification component, which is often implemented by projecting the representative vector onto target class vectors. The target class vectors, along with the rest of the model parameters, are estimated so as to minimize the loss function.

In this paper, we analyze how simple design choices for the classification layer affect the learning dynamics. We show that the standard cross-entropy training implicitly captures visual similarities between different classes, which might deteriorate accuracy or even prevents some models from converging. We propose to

draw the class vectors randomly and set them as fixed during training, thus invalidating the visual similarities encoded in these vectors. We analyze the effects of keeping the class vectors fixed and show that it can increase the inter-class separability, intra-class compactness, and the overall model accuracy, while maintaining the robustness to image corruptions and the generalization of the learned concepts.

****

A Multi-Modal and Multitask Benchmark in the Clinical Domain

Yong Huang,Edgar Mariano Marroquin,Volodymyr Kuleshov

Healthcare represents one of the most promising application areas for machine learning algorithms, including modern methods based on deep learning. Modern deep learning algorithms perform best on large datasets and on unstructured modalities such as text or image data; advances in deep learning have often been driven by the availability of such large datasets. Here, we introduce Multi-Modal Multitask MIMIC-III (M3) — a dataset and benchmark for evaluating machine learning algorithms in the healthcare domain. This dataset contains multi-modal patient data collected from intensive care units — including physiological time series, clinical notes, ECG waveforms, and tabular inputs — and defines six clinical tasks — including predicting mortality, decompensation, readmission, and other outcomes — which serve as benchmarks for comparing algorithms. We introduce new multi-modal and multitask models for this dataset, and show that they outperform previous state-of-the-art results that only rely on a subset of all tasks and modalities. This highlights the potential of multitask and multi-modal learning to improve the performance of algorithms in the healthcare domain. More generally, we envision M3 as a general resource that will help accelerate research in applying machine learning to healthcare.

****

Selective Classification Can Magnify Disparities Across Groups

Erik Jones,Shiori Sagawa,Pang Wei Koh,Ananya Kumar,Percy Liang

Selective classification, in which models can abstain on uncertain predictions, is a natural approach to improving accuracy in settings where errors are costly but abstentions are manageable. In this paper, we find that while selective classification can improve average accuracies, it can simultaneously magnify existing accuracy disparities between various groups within a population, especially in the presence of spurious correlations. We observe this behavior consistently across five vision and NLP datasets. Surprisingly, increasing abstentions can even decrease accuracies on some groups. To better understand this phenomenon, we study the margin distribution, which captures the model’s confidences over all predictions. For symmetric margin distributions, we prove that whether selective classification monotonically improves or worsens accuracy is fully determined by the accuracy at full coverage (i.e., without any abstentions) and whether the distribution satisfies a property we call left-log-concavity. Our analysis also shows that selective classification tends to magnify full-coverage accuracy disparities. Motivated by our analysis, we train distributionally-robust models that achieve similar full-coverage accuracies across groups and show that selective classification uniformly improves each group on these models. Altogether, our results suggest that selective classification should be used with care and underscore the importance of training models to perform equally well across groups at full coverage.

****

A Provably Convergent and Practical Algorithm for Min-Max Optimization with Applications to GANs

Oren Mangoubi,Sushant Sachdeva,Nisheeth K Vishnoi

We present a first-order algorithm for nonconvex-nonconcave min-max optimization problems such as those that arise in training GANs. Our algorithm provably converges in $\mathrm{poly}(d,L, b)$ steps for any loss function $f:\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ which is b-bounded with ${L}$-Lipschitz gradient. To achieve convergence, we 1) give a novel approximation to the global strategy of the max-player based on first-order algorithms such as gradient ascent, and 2) empower the min-player to look ahead and simulate the max-player

’s response for arbitrarily many steps, but restrict the min-player to move acco rding to updates sampled from a stochastic gradient oracle. Our algorithm, when used to train GANs on synthetic and real-world datasets, does not cycle, results in GANs that seem to avoid mode collapse, and achieves a training time per iter ation and memory requirement similar to gradient descent-ascent.

****
What are effective labels for augmented data? Improving robustness with AutoLabe l
Yao Qin,Xuezhi Wang,Balaji Lakshminarayanan,Ed Chi,Alex Beutel
A wide breadth of research has devised data augmentation approaches that can imp rove both accuracy and generalization performance for neural networks. However, augmented data can end up being far from the clean data and what is the appropri ate label is less clear. Despite this, most existing work simply reuses the orig inal label from the clean data, and the choice of label accompanying the augment ed data is relatively less explored. In this paper, we propose AutoLabel to auto matically learn the labels for augmented data, based on the distance between the clean distribution and augmented distribution. AutoLabel is built on label smoo thing and is guided by the calibration-performance over a hold-out validation se t. We show that AutoLabel is a generic framework that can be easily applied to e xisting data augmentation methods, including AugMix, mixup, and adversarial trai ning. Experiments on CIFAR-10, CIFAR-100 and ImageNet show that AutoLabel can im prove models' accuracy and calibration performance, especially under distributio nal shift. Additionally, we demonstrate that AutoLabel can help adversarial trai ning by bridging the gap between clean accuracy and adversarial robustness.
****
Individuality in the hive - Learning to embed lifetime social behaviour of honey bees
Benjamin Wild,David Dormagen,Michael L Smith,Tim Landgraf
Honey bees are a popular model for complex social systems, in which global behav ior emerges from the actions and interactions of thousands of individuals. While the average life of a bee is organized as a sequence of tasks roughly determine d by age, there is substantial variation at the individual level. For example, y oung bees can become foragers early in life, depending on the colony’s needs. Us ing a unique dataset containing lifetime trajectories of all individuals over mu ltiple generations in two honey bee colonies, we propose a new temporal matrix f actorization model that jointly learns the average developmental path and struct ured variations of individuals in the social network over their entire lives. Ou r method yields inherently interpretable embeddings that are biologically plausi ble and consistent over time, which allow one to compare individuals regardless of when, or in which colony, they lived. Our method provides a quantitative fram ework for understanding behavioral heterogeneity in complex social systems appli cable in fields such as behavioral biology, social sciences, neuroscience, and i nformation science.
****
Matrix Shuffle-Exchange Networks for Hard 2D Tasks
Em■ls Ozoli■š,Karlis Freivalds,Agris Šostaks
Convolutional neural networks have become the main tools for processing two-dime nsional data. They work well for images, yet convolutions have a limited recepti ve field that prevents its applications to more complex 2D tasks. We propose a n ew neural model, called Matrix Shuffle-Exchange network, that can efficiently ex ploit long-range dependencies in 2D data and has comparable speed to a convoluti onal neural network. It is derived from Neural Shuffle-Exchange network and has $\mathcal{O}(\log{n})$ layers and $\mathcal{O}(n^2 \log{n})$ total time and s pace complexity for processing a $n \times n$ data matrix. We show that the Matr ix Shuffle-Exchange network is well-suited for algorithmic and logical reasoning tasks on matrices and dense graphs, exceeding convolutional and graph neural ne twork baselines. Its distinct advantage is the capability of retaining full long -range dependency modelling when generalizing to larger instances -- much larger than could be processed with models equipped with a dense attention mechanism.

****
FedMix: Approximation of Mixup under Mean Augmented Federated Learning
Tehrim Yoon,Sumin Shin,Sung Ju Hwang,Eunho Yang
Federated learning (FL) allows edge devices to collectively learn a model without directly sharing data within each device, thus preserving privacy and eliminating the need to store data globally. While there are promising results under the assumption of independent and identically distributed (iid) local data, current state-of-the-art algorithms suffer a performance degradation as the heterogeneity of local data across clients increases. To resolve this issue, we propose a simple framework, \emph{Mean Augmented Federated Learning (MAFL)}, where clients send and receive \emph{averaged} local data, subject to the privacy requirements of target applications. Under our framework, we propose a new augmentation algorithm, named \emph{FedMix}, which is inspired by a phenomenal yet simple data augmentation method, Mixup, but does not require local raw data to be directly shared among devices. Our method shows greatly improved performance in the standard benchmark datasets of FL, under highly non-iid federated settings, compared to conventional algorithms.
****
Improving Calibration through the Relationship with Adversarial Robustness
Yao Qin,Xuezhi Wang,Alex Beutel,Ed Chi
Neural networks lack adversarial robustness -- they are vulnerable to adversarial examples that through small perturbations to inputs cause incorrect predictions. Further, trust is undermined when models give miscalibrated uncertainty estimates, i.e. the predicted probability is not a good indicator of how much we should trust our model. In this paper, we study the connection between adversarial robustness and calibration on four classification networks and datasets. We find that the inputs for which the model is sensitive to small perturbations (are easily attacked) are more likely to have poorly calibrated predictions. Based on this insight, we examine if calibration can be improved by addressing those adversarially unrobust inputs. To this end, we propose Adversarial Robustness based Adaptive Label Smoothing (AR-AdaLS) that integrates the correlations of adversarial robustness and uncertainty into training by adaptively softening labels for an example based on how easily it can be attacked by an adversary. We find that our method, taking the adversarial robustness of the in-distribution data into consideration, leads to better calibration over the model even under distributional shifts. In addition, AR-AdaLS can also be applied to an ensemble model to further improve model's calibration.
****
Memformer: The Memory-Augmented Transformer
Qingyang Wu,Zhenzhong Lan,Jing Gu,Zhou Yu
Transformer models have obtained remarkable accomplishments in various NLP tasks. However, these models have efficiency issues on long sequences, as the complexity of their self-attention module scales quadratically with the sequence length. To remedy the limitation, we present Memformer, a novel language model that utilizes a single unified memory to encode and retrieve past information. It includes a new optimization scheme, Memory Replay Back-Propagation, which promotes long-range back-propagation through time with a significantly reduced memory requirement. Memformer achieves $\mathcal{O}(n)$ time complexity and $\mathcal{O}(1)$ space complexity in processing long sequences, meaning that the model can handle an infinite length sequence during inference. Our model is also compatible with other self-supervised tasks to further improve the performance on language modeling. Experimental results show that Memformer outperforms the previous long-range sequence models on WikiText-103, including Transformer-XL and Compressive Transformer.
****
In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness
Sang Michael Xie,Ananya Kumar,Robbie Jones,Fereshte Khani,Tengyu Ma,Percy Liang
Consider a prediction setting with few in-distribution labeled examples and many unlabeled examples both in- and out-of-distribution (OOD). The goal is to learn

a model which performs well both in-distribution and OOD. In these settings, auxiliary information is often cheaply available for every input. How should we best leverage this auxiliary information for the prediction task? Empirically across three image and time-series datasets, and theoretically in a multi-task linear regression setting, we show that (i) using auxiliary information as input features improves in-distribution error but can hurt OOD error; but (ii) using auxiliary information as outputs of auxiliary pre-training tasks improves OOD error. To get the best of both worlds, we introduce In-N-Out, which first trains a model with auxiliary inputs and uses it to pseudolabel all the in-distribution inputs, then pre-trains a model on OOD auxiliary outputs and fine-tunes this model with the pseudolabels (self-training). We show both theoretically and empirically that In-N-Out outperforms auxiliary inputs or outputs alone on both in-distribution and OOD error.

****

On Relating "Why?" and "Why Not?" Explanations

Alexey Ignatiev,Nina Narodytska,Nicholas Asher,Joao Marques-Silva

Explanations of Machine Learning (ML) models often address a ‘Why?’ question. Such explanations can be related with selecting feature-value pairs which are sufficient for the prediction. Recent work has investigated explanations that address
a ‘Why Not?’ question, i.e. finding a change of feature values that guarantee a change of prediction. Given their goals, these two forms of explaining predictions of ML models appear to be mostly unrelated. However, this paper demonstrates otherwise, and establishes a rigorous formal relationship between ‘Why?’ and ‘Why Not?’ explanations. Concretely, the paper proves that, for any given instance, ‘Why?’ explanations are minimal hitting sets of ‘Why Not?’ explanations and vice-versa. Furthermore, the paper devises novel algorithms for extracting and enumerating both forms of explanations.

****

Learning Contextualized Knowledge Structures for Commonsense Reasoning

Jun Yan,Mrigank Raman,Tianyu Zhang,Ryan Rossi,Handong Zhao,Sungchul Kim,Nedim Lipka,Xiang Ren

Recently, neural-symbolic architectures have achieved success on commonsense reasoning through effectively encoding relational structures retrieved from external knowledge graphs (KGs) and obtained state-of-the-art results in tasks such as (commonsense) question answering and natural language inference. However, current neural-symbolic reasoning methods rely on quality and contextualized knowledge structures (i.e., fact triples) that can be retrieved at the pre-processing stage and overlook challenges such as dealing with incompleteness of a KG (low coverage), limited expressiveness of its relations, and irrelevant retrieved facts in the reasoning context.

In this paper, we present a novel neural-symbolic approach, named Hybrid Graph Network (HGN), which jointly generates feature representations for new triples (as complement to the existing edges in the KG), determines relevance of the triples to the reasoning context, and learns graph model parameters for encoding the relational information. Our method learns a compact graph structure (comprising both retrieved and generated edges) through filtering edges that are unhelpful to the reasoning process. We show marked improvements on three commonsense reasoning benchmarks and demonstrate the superiority of the learned graph structures with user studies.

****

Sample-Efficient Automated Deep Reinforcement Learning

Jörg K.H. Franke,Gregor Koehler,André Biedenkapp,Frank Hutter

Despite significant progress in challenging problems across various domains, applying state-of-the-art deep reinforcement learning (RL) algorithms remains challenging due to their sensitivity to the choice of hyperparameters. This sensitivity can partly be attributed to the non-stationarity of the RL problem, potentially requiring different hyperparameter settings at various stages of the learning process. Additionally, in the RL setting, hyperparameter optimization (HPO) req

uires a large number of environment interactions, hindering the transfer of the successes in RL to real-world applications. In this work, we tackle the issues of sample-efficient and dynamic HPO in RL. We propose a population-based automated RL (AutoRL) framework to meta-optimize arbitrary off-policy RL algorithms. In this framework, we optimize the hyperparameters and also the neural architecture while simultaneously training the agent. By sharing the collected experience across the population, we substantially increase the sample efficiency of the meta-optimization. We demonstrate the capabilities of our sample-efficient AutoRL approach in a case study with the popular TD3 algorithm in the MuJoCo benchmark suite, where we reduce the number of environment interactions needed for meta-optimization by up to an order of magnitude compared to population-based training.
****

Quantum Deformed Neural Networks
Roberto Bondesan,Max Welling
We develop a new quantum neural network layer designed to run efficiently on a quantum computer but that can be simulated on a classical computer when restricted in the way it entangles input states. We first ask how a classical neural network architecture, both fully connected or convolutional, can be executed on a quantum computer using quantum phase estimation. We then deform the classical layer into a quantum design which entangles activations and weights into quantum superpositions. While the full model would need the exponential speedups delivered by a quantum computer, a restricted class of designs represent interesting new classical network layers that still use quantum features. We show that these quantum deformed neural networks can be trained and executed on normal data such as images, and even classically deliver modest improvements over standard architectures.
****

A Temporal Kernel Approach for Deep Learning with Continuous-time Information
Da Xu,Chuanwei Ruan,Evren Korpeoglu,Sushant Kumar,Kannan Achan
Sequential deep learning models such as RNN, causal CNN and attention mechanism do not readily consume continuous-time information. Discretizing the temporal data, as we show, causes inconsistency even for simple continuous-time processes. Current approaches often handle time in a heuristic manner to be consistent with the existing deep learning architectures and implementations. In this paper, we provide a principled way to characterize continuous-time systems using deep learning tools. Notably, the proposed approach applies to all the major deep learning architectures and requires little modifications to the implementation. The critical insight is to represent the continuous-time system by composing neural networks with a temporal kernel, where we gain our intuition from the recent advancements in understanding deep learning with Gaussian process and neural tangent kernel. To represent the temporal kernel, we introduce the random feature approach and convert the kernel learning problem to spectral density estimation under reparameterization. We further prove the convergence and consistency results even when the temporal kernel is non-stationary, and the spectral density is misspecified. The simulations and real-data experiments demonstrate the empirical effectiveness of our temporal kernel approach in a broad range of settings.
****

Explainability for fair machine learning
Tom Begley,Tobias Schwedes,Christopher Frye,Ilya Feige
As the decisions made or influenced by machine learning models increasingly impact our lives, it is crucial to detect, understand, and mitigate unfairness. But even simply determining what ``unfairness'' should mean in a given context is non-trivial: there are many competing definitions, and choosing between them often requires a deep understanding of the underlying task. It is thus tempting to use model explainability to gain insights into model fairness, however existing explainability tools do not reliably indicate whether a model is indeed fair. In this work we present a new approach to explaining fairness in machine learning, based on the Shapley value paradigm. Our fairness explanations attribute a model's overall unfairness to individual input features, even in cases where the model does not operate on sensitive attributes directly. Moreover, motivated by the l

inearity of Shapley explainability, we propose a meta algorithm for applying exi sting training-time fairness interventions, wherein one trains a perturbation to the original model, rather than a new model entirely. By explaining the origina l model, the perturbation, and the fair-corrected model, we gain insight into th e accuracy-fairness trade-off that is being made by the intervention. We further show that this meta algorithm enjoys both flexibility and stability benefits wi th no loss in performance.
****
Convex Regularization behind Neural Reconstruction
Arda Sahiner,Morteza Mardani,Batu Ozturkler,Mert Pilanci,John M. Pauly
Neural networks have shown tremendous potential for reconstructing high-resoluti on images in inverse problems. The non-convex and opaque nature of neural networ ks, however, hinders their utility in sensitive applications such as medical ima ging. To cope with this challenge, this paper advocates a convex duality framewo rk that makes a two-layer fully-convolutional ReLU denoising network amenable to convex optimization. The convex dual network not only offers the optimum traini ng with convex solvers, but also facilitates interpreting training and predictio n. In particular, it implies training neural networks with weight decay regulari zation induces path sparsity while the prediction is piecewise linear filtering. A range of experiments with MNIST and fastMRI datasets confirm the efficacy of the dual network optimization problem.
****
A Representational Model of Grid Cells' Path Integration Based on Matrix Lie Alg ebras
Ruiqi Gao,Jianwen Xie,Xue-Xin Wei,Song-Chun Zhu,Ying Nian Wu
The grid cells in the mammalian medial entorhinal cortex exhibit striking hexago n firing patterns when the agent navigates in the open field. It is hypothesized that the grid cells are involved in path integration so that the agent is aware of its self-position by accumulating its self-motion. Assuming the grid cells f orm a vector representation of self-position, we elucidate a minimally simple re current model for grid cells' path integration based on two coupled matrix Lie a lgebras that underlie two coupled rotation systems that mirror the agent's self- motion: (1) When the agent moves along a certain direction, the vector is rotate d by a generator matrix. (2) When the agent changes direction, the generator mat rix is rotated by another generator matrix. Our experiments show that our model learns hexagonal grid response patterns that resemble the firing patterns observ ed from the grid cells in the brain. Furthermore, the learned model is capable o f near exact path integration, and it is also capable of error correction. Our m odel is novel and simple, with explicit geometric and algebraic structures.
****
Vector-output ReLU Neural Network Problems are Copositive Programs: Convex Analy sis of Two Layer Networks and Polynomial-time Algorithms
Arda Sahiner,Tolga Ergen,John M. Pauly,Mert Pilanci
We describe the convex semi-infinite dual of the two-layer vector-output ReLU ne ural network training problem. This semi-infinite dual admits a finite dimension al representation, but its support is over a convex set which is difficult to ch aracterize. In particular, we demonstrate that the non-convex neural network tra ining problem is equivalent to a finite-dimensional convex copositive program. O ur work is the first to identify this strong connection between the global optim a of neural networks and those of copositive programs. We thus demonstrate how n eural networks implicitly attempt to solve copositive programs via semi-nonnegat ive matrix factorization, and draw key insights from this formulation. We descri be the first algorithms for provably finding the global minimum of the vector ou tput neural network training problem, which are polynomial in the number of samp les for a fixed data rank, yet exponential in the dimension. However, in the cas e of convolutional architectures, the computational complexity is exponential in only the filter size and polynomial in all other parameters. We describe the ci rcumstances in which we can find the global optimum of this neural network train ing problem exactly with soft-thresholded SVD, and provide a copositive relaxati on which is guaranteed to be exact for certain classes of problems, and which co

rresponds with the solution of Stochastic Gradient Descent in practice.
****
Multi-Level Generative Models for Partial Label Learning with Non-random Label Noise
Yan Yan,Yuhong Guo
Partial label (PL) learning tackles the problem where each training instance is associated with a set of candidate labels that include both the true label and irrelevant noise labels. In this paper, we propose a novel multi-level generative model for partial label learning (MGPLL), which tackles the PL problem by learning both a label level adversarial generator and a feature level adversarial generator under a bi-directional mapping framework between the label vectors and the data samples. MGPLL uses a conditional noise label generation network to model the non-random noise labels and perform label denoising, and uses a multi-class predictor to map the training instances to the denoised label vectors, while a conditional data feature generator is used to form an inverse mapping from the denoised label vectors to data samples. Both the noise label generator and the data feature generator are learned in an adversarial manner to match the observed candidate labels and data features respectively. We conduct extensive experiments on both synthesized and real-world partial label datasets. The proposed approach demonstrates the state-of-the- art performance for partial label learning.
****
Beyond COVID-19 Diagnosis: Prognosis with Hierarchical Graph Representation Learning
CHEN LIU,Jinze Cui,Dailin Gan,Guosheng Yin
Coronavirus disease 2019 (COVID-19), the pandemic that is spreading fast globally, has caused over 34 million confirmed cases. Apart from the reverse transcription polymerase chain reaction (RT-PCR), the chest computed tomography (CT) is viewed as a standard and effective tool for disease diagnosis and progression monitoring. We propose a diagnosis and prognosis model based on graph convolutional networks (GCNs). The chest CT scan of a patient, typically involving hundreds of sectional images in sequential order, is formulated as a densely connected weighted graph. A novel distance aware pooling is proposed to abstract the node information hierarchically, which is robust and efficient for such densely connected graphs. Our method, combining GCNs and distance aware pooling, can integrate the information from all slices in the chest CT scans for optimal decision making, which leads to the state-of-the-art accuracy in the COVID-19 diagnosis and prognosis. With less than 1% number of total parameters in the baseline 3D ResNet model, our method achieves 94.8% accuracy for diagnosis. It has a 2.4% improvement compared with the baseline model on the same dataset. In addition, we can localize the most informative slices with disease lesions for COVID-19 within a large sequence of chest CT images. The proposed model can produce visual explanations for the diagnosis and prognosis, making the decision more transparent and explainable, while RT-PCR only leads to the test result with no prognosis information. The prognosis analysis can help hospitals or clinical centers designate medical resources more efficiently and better support clinicians to determine the proper clinical treatment.
****
H-divergence: A Decision-Theoretic Probability Discrepancy Measure
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. Based on ideas from decision theory, we investigate a new class of discrepancies that are based on the optimal decision loss. Two probability distributions are different if the optimal decision loss is higher on the mixture distribution than on each individual distribution. We show that this generalizes popular notions of discrepancy measurements such as the Jensen Shannon divergence and the maximum mean discrepancy. We apply our approach to two-sample tests, which evaluates whether two sets of samples come from the same distribution. On various benchmark and real datasets, we demonstrate that tests based on our generalized notion of discrepancy is able to achieve su

perior test power. We also apply our approach to sample quality evaluation as an alternative to the FID score, and to understanding the effects of climate change on different social and economic activities.
****

AttackDist: Characterizing Zero-day Adversarial Samples by Counter Attack
Simin Chen,Zihe Song,Lei Ma,Cong Liu,Wei Yang
Deep Neural Networks (DNNs) have been shown vulnerable to adversarial attacks, which could produce adversarial samples that easily fool the state-of-the-art DNNs. The harmfulness of adversarial attacks calls for the defense mechanisms under fire. However, the relationship between adversarial attacks and defenses is like spear and shield. Whenever a defense method is proposed, a new attack would be followed to bypass the defense immediately. Devising a definitive defense against new attacks~(zero-day attacks) is proven to be challenging. We tackle this challenge by characterizing the intrinsic properties of adversarial samples, via measuring the norm of the perturbation after a counterattack. Our method is based on the idea that, from an optimization perspective, adversarial samples would be closer to the decision boundary; thus the perturbation to counterattack adversarial samples would be significantly smaller than normal cases. Motivated by this, we propose AttackDist, an attack-agnostic property to characterize adversarial samples. We first theoretically clarify under which condition AttackDist can provide a certified detecting performance, then show that a potential application of AttackDist is distinguishing zero-day adversarial examples without knowing the mechanisms of new attacks. As a proof-of-concept, we evaluate AttackDist on two widely used benchmarks. The evaluation results show that AttackDist can outperform the state-of-the-art detection measures by large margins in detecting zero-day adversarial attacks.
****

Learning Better Structured Representations Using Low-rank Adaptive Label Smoothing
Asish Ghoshal,Xilun Chen,Sonal Gupta,Luke Zettlemoyer,Yashar Mehdad
Training with soft targets instead of hard targets has been shown to improve performance and calibration of deep neural networks. Label smoothing is a popular way of computing soft targets, where one-hot encoding of a class is smoothed with a uniform distribution. Owing to its simplicity, label smoothing has found wide-spread use for training deep neural networks on a wide variety of tasks, ranging from image and text classification to machine translation and semantic parsing. Complementing recent empirical justification for label smoothing, we obtain PAC-Bayesian generalization bounds for label smoothing and show that the generalization error depends on the choice of the noise (smoothing) distribution. Then we propose low-rank adaptive label smoothing (LORAS): a simple yet novel method for training with learned soft targets that generalizes label smoothing and adapts to the latent structure of the label space in structured prediction tasks. Specifically, we evaluate our method on semantic parsing tasks and show that training with appropriately smoothed soft targets can significantly improve accuracy and model calibration, especially in low-resource settings. Used in conjunction with pre-trained sequence-to-sequence models, our method achieves state of the art performance on four semantic parsing data sets. LORAS can be used with any model, improves performance and implicit model calibration without increasing the number of model parameters, and can be scaled to problems with large label spaces containing tens of thousands of labels.
****

Adaptive Gradient Methods Can Be Provably Faster than SGD with Random Shuffling
Xunpeng Huang,Vicky Jiaqi Zhang,Hao Zhou,Lei Li
Adaptive gradient methods have been shown to outperform SGD in many tasks of training neural networks. However, the acceleration effect is yet to be explained in the non-convex setting since the best convergence rate of adaptive gradient methods is worse than that of SGD in literature. In this paper, we prove that adaptive gradient methods exhibit an $\small\tilde{O}(T^{-1/2})$-convergence rate for finding first-order stationary points under the strong growth condition, which improves previous best convergence results of adaptive gradient methods and ran

dom shuffling SGD by factors of $\small O(T^{-1/4})$ and $\small O(T^{-1/6})$, r
espectively. In particular, we study two variants of AdaGrad with random shuffli
ng for finite sum minimization. Our analysis suggests that the combination of ra
ndom shuffling and adaptive learning rates gives rise to better convergence.
****

Towards certifying ℓ_∞ robustness using Neural networks with $\ell_\in
fty$-dist Neurons
Bohang Zhang,Zhou Lu,Tianle Cai,Di He,Liwei Wang
It is well-known that standard neural networks, even with a high classification
accuracy, are vulnerable to small ℓ_∞ perturbations. Many attempts hav
e been tried to learn a network that can resist such adversarial attacks. Howeve
r, most previous works either can only provide empirical verification of the def
ense to a particular attack method or can only develop a theoretical guarantee o
f the model robustness in limited scenarios. In this paper, we develop a theoret
ically principled neural network that inherently resists ℓ_∞ perturbat
ions. In particular, we design a novel neuron that uses ℓ_∞ distance a
s its basic operation, which we call ℓ_∞-dist neuron. We show that the
 ℓ_∞-dist neuron is naturally a 1-Lipschitz function with respect to t
he ℓ_∞ norm, and the neural networks constructed with ℓ_∞-di
st neuron (ℓ_{∞}-dist Nets) enjoy the same property. This directly pro
vides a theoretical guarantee of the certified robustness based on the margin of
 the prediction outputs. We further prove that the ℓ_{∞}-dist Nets hav
e enough expressiveness power to approximate any 1-Lipschitz function, and can g
eneralize well as the robust test error can be upper-bounded by the performance
of a large margin classifier on the training data. Preliminary experiments show
that even without the help of adversarial training, the learned networks with hi
gh classification accuracy are already provably robust.
****

Understanding the role of importance weighting for deep learning
Da Xu,Yuting Ye,Chuanwei Ruan
The recent paper by Byrd & Lipton (2019), based on empirical observations, raise
s a major concern on the impact of importance weighting for the over-parameteriz
ed deep learning models. They observe that as long as the model can separate the
 training data, the impact of importance weighting diminishes as the training pr
oceeds. Nevertheless, there lacks a rigorous characterization of this phenomenon
. In this paper, we provide formal characterizations and theoretical justificati
ons on the role of importance weighting with respect to the implicit bias of gra
dient descent and margin-based learning theory. We reveal both the optimization
dynamics and generalization performance under deep learning models. Our work not
 only explains the various novel phenomenons observed for importance weighting i
n deep learning, but also extends to the studies where the weights are being opt
imized as part of the model, which applies to a number of topics under active re
search.
****

Powers of layers for image-to-image translation
Hugo Touvron,Matthijs Douze,Matthieu Cord,Herve Jegou
We propose a simple architecture to address unpaired image-to-image translation
 tasks: style or class transfer, denoising, deblurring, deblocking, etc.
We start from an image autoencoder architecture with fixed weights.
For each task we learn a residual block operating in the latent space, which is
iteratively called until the target domain is reached.
A specific training schedule is required to alleviate the exponentiation effect
of the iterations.
At test time, it offers several advantages: the number of weight parameters is l
imited and the compositional design allows one to modulate the strength of the t
ransformation with the number of iterations.
This is useful, for instance, when the type or amount of noise to suppress is no
t known in advance.
Experimentally, we show that the performance of our model is comparable or bette
r than CycleGAN and Nice-GAN with fewer parameters.

****
Training GANs with Stronger Augmentations via Contrastive Discriminator
Jongheon Jeong,Jinwoo Shin
Recent works in Generative Adversarial Networks (GANs) are actively revisiting various data augmentation techniques as an effective way to prevent discriminator overfitting. It is still unclear, however, that which augmentations could actually improve GANs, and in particular, how to apply a wider range of augmentations in training. In this paper, we propose a novel way to address these questions by incorporating a recent contrastive representation learning scheme into the GAN discriminator, coined ContraD. This "fusion" enables the discriminators to work with much stronger augmentations without increasing their training instability, thereby preventing the discriminator overfitting issue in GANs more effectively. Even better, we observe that the contrastive learning itself also benefits from our GAN training, i.e., by maintaining discriminative features between real and fake samples, suggesting a strong coherence between the two worlds: good contrastive representations are also good for GAN discriminators, and vice versa. Our experimental results show that GANs with ContraD consistently improve FID and IS compared to other recent techniques incorporating data augmentations, still maintaining highly discriminative features in the discriminator in terms of the linear evaluation. Finally, as a byproduct, we also show that our GANs trained in an unsupervised manner (without labels) can induce many conditional generative models via a simple latent sampling, leveraging the learned features of ContraD. Code is available at https://github.com/jh-jeong/ContraD.
****
A Unified Bayesian Framework for Discriminative and Generative Continual Learning
Abhishek Kumar,Sunabha Chatterjee,Piyush Rai
Continual Learning is a learning paradigm where learning systems are trained on a sequence of tasks. The goal here is to perform well on the current task without suffering from a performance drop on the previous tasks. Two notable directions among the recent advances in continual learning with neural networks are (1) variational Bayes based regularization by learning priors from previous tasks, and, (2) learning the structure of deep networks to adapt to new tasks. So far, these two approaches have been orthogonal. We present a novel Bayesian framework for continual learning based on learning the structure of deep neural networks, addressing the shortcomings of both these approaches. The proposed framework learns the deep structure for each task by learning which weights to be used, and supports inter-task transfer through the overlapping of different sparse subsets of weights learned by different tasks. An appealing aspect of our proposed continual learning framework is that it is applicable to both discriminative (supervised) and generative (unsupervised) settings. Experimental results on supervised and unsupervised benchmarks shows that our model performs comparably or better than recent advances in continual learning.
****
Neural Partial Differential Equations with Functional Convolution
Ziqian Wu,Xingzhe He,Michael Zhang,Yijun Li,Cheng Yang,Rui Liu,Shiying Xiong,Bo Zhu
We present a lightweighted neural PDE representation to discover the hidden structure and predict the solution of different nonlinear PDEs. Our key idea is to leverage the prior of ``"translational similarity" of numerical PDE differential operators to drastically reduce the scale of learning model and training data. We implemented three central network components, including a neural functional convolution operator, a Picard forward iterative procedure, and an adjoint backward gradient calculator. Our novel paradigm fully leverages the multifaceted priors that stem from the sparse and smooth nature of the physical PDE solution manifold and the various mature numerical techniques such as adjoint solver, linearization, and iterative procedure to accelerate the computation. We demonstrate the efficacy of our method by robustly discovering the model and accurately predicting the solutions of various types of PDEs with small-scale networks and training sets. We highlight that all the PDE examples we showed were trained with up to

8 data samples and within 325 network parameters.
****

Transformers are Deep Infinite-Dimensional Non-Mercer Binary Kernel Machines
Matthew A Wright,Joseph E. Gonzalez
Despite their ubiquity in core AI fields like natural language processing, the mechanics of deep attention-based neural networks like the ``Transformer'' model are not fully understood. In this article, we present a new perspective towards understanding how Transformers work. In particular, we show that the ``"dot-product attention" that is the core of the Transformer's operation can be characterized as a kernel learning method on a pair of Banach spaces. In particular, the Transformer's kernel is characterized as having an infinite feature dimension. Along the way we generalize the standard kernel learning problem to what we term a "binary" kernel learning problem, where data come from two input domains and a response is defined for every cross-domain pair. We prove a new representer theorem for these binary kernel machines with non-Mercer (indefinite, asymmetric) kernels (implying that the functions learned are elements of reproducing kernel Banach spaces rather than Hilbert spaces), and also prove a new universal approximation theorem showing that the Transformer calculation can learn any binary non-Mercer reproducing kernel Banach space pair. We experiment with new kernels in Transformers, and obtain results that suggest the infinite dimensionality of the standard Transformer kernel is partially responsible for its performance. This paper's results provide a new theoretical understanding of a very important but poorly understood model in modern machine learning.
****

Dynamic Feature Selection for Efficient and Interpretable Human Activity Recognition
Randy Ardywibowo,Shahin Boluki,Zhangyang Wang,Bobak J Mortazavi,Shuai Huang,Xiaoning Qian
In many machine learning tasks, input features with varying degrees of predictive capability are usually acquired at some cost. For example, in human activity recognition (HAR) and mobile health (mHealth) applications, monitoring performance should be achieved with a low cost to gather different sensory features, as maintaining sensors incur monetary, computation, and energy cost. We propose an adaptive feature selection method that dynamically selects features for prediction at any given time point. We formulate this problem as an ℓ_0 minimization problem across time, and cast the combinatorial optimization problem into a stochastic optimization formulation. We then utilize a differentiable relaxation to make the problem amenable to gradient-based optimization. Our evaluations on four activity recognition datasets show that our method achieves a favorable trade-off between performance and the number of features used. Moreover, the dynamically selected features of our approach are shown to be interpretable and associated with the actual activity types.
****

Deep k-NN Label Smoothing Improves Reproducibility of Neural Network Predictions
Dara Bahri,Heinrich Jiang
Training modern neural networks is an inherently noisy process that can lead to high \emph{prediction churn}-- disagreements between re-trainings of the same model due to factors such as randomization in the parameter initialization and mini-batches-- even when the trained models all attain high accuracies. Such prediction churn can be very undesirable in practice. In this paper, we present several baselines for reducing churn and show that utilizing the k-NN predictions to smooth the labels results in a new and principled method that often outperforms the baselines on churn while improving accuracy on a variety of benchmark classification tasks and model architectures.
****

Variational Structured Attention Networks for Dense Pixel-Wise Prediction
Guanglei Yang,Paolo Rota,Xavier Alameda-Pineda,Dan Xu,Mingli Ding,Elisa Ricci
State-of-the-art performances in dense pixel-wise prediction tasks are obtained with specifically designed convolutional networks. These models often benefit fr

om attention mechanisms that allow better learning of deep representations. Rece nt works showed the importance of estimating both spatial- and channel-wise atte ntion tensors. In this paper, we propose a unified approach to jointly estimate spatial attention maps and channel attention vectors so as to structure the resu lting attention tensor. Moreover, we integrate the estimation of the attention w ithin a probabilistic framework, leading to VarIational STructured Attention net works(VISTA). We implement the inference rules within the neural network, thus a llowing for joint learning of the probabilistic and the CNN front-end parameters . Importantly, as demonstrated by our extensive empirical evaluation on six larg e-scale datasets VISTA outperforms the state-of-the-art in multiple continuous a nd discrete pixel-level prediction tasks, thus confirming the benefit of structu ring the attention tensor and of inferring it within a probabilistic formulation .
****
Private Image Reconstruction from System Side Channels Using Generative Models
Yuanyuan Yuan,Shuai Wang,Junping Zhang
System side channels denote effects imposed on the underlying system and hardwar e when running a program, such as its accessed CPU cache lines. Side channel ana lysis (SCA) allows attackers to infer program secrets based on observed side cha nnel signals. Given the ever-growing adoption of machine learning as a service (MLaaS), image analysis software on cloud platforms has been exploited by reconst ructing private user images from system side channels. Nevertheless, to date, SC A is still highly challenging, requiring technical knowledge of victim software' s internal operations. For existing SCA attacks, comprehending such internal ope rations requires heavyweight program analysis or manual efforts.

This research proposes an attack framework to reconstruct private user images pr ocessed by media software via system side channels. The framework forms an effec tive workflow by incorporating convolutional networks, variational autoencoders, and generative adversarial networks. Our evaluation of two popular side channel s shows that the reconstructed images consistently match user inputs, making pri vacy leakage attacks more practical. We also show surprising results that even o ne-bit data read/write pattern side channels, which are deemed minimally informa tive, can be used to reconstruct quality images using our framework.
****
RMSprop converges with proper hyper-parameter
Naichen Shi,Dawei Li,Mingyi Hong,Ruoyu Sun
Despite the existence of divergence examples, RMSprop remains
one of the most popular algorithms in machine learning. Towards closing the gap between theory and practice, we prove that RMSprop converges with proper choice of hyper-parameters under certain conditions. More specifically, we prove that w hen the hyper-parameter β_2 is close enough to 1, RMSprop and its random shuffling version converge to a bounded region in general, and to critical poin ts in the interpolation regime. It is worth mentioning that our results do not d epend on ``bounded gradient" assumption, which is often the key assumption uti lized by existing theoretical work for Adam-type adaptive gradient method. Remov ing this assumption allows us to establish a phase transition from divergence to non-divergence for RMSprop.

Finally, based on our theory, we conjecture that in practice there is a critical threshold $\sf{\beta_2^*}$, such that RMSprop generates reasonably good results only if $1>\beta_2\ge \sf{\beta_2^*}$. We provide empirical evidence for such a phase transition in our numerical experiments.
****
Estimating Treatment Effects via Orthogonal Regularization
Tobias Hatt,Stefan Feuerriegel
Decision-making often requires accurate estimation of causal effects from observ ational data. This is challenging as outcomes of alternative decisions are not o bserved and have to be estimated. Previous methods estimate outcomes based on un confoundedness but neglect any constraints that unconfoundedness imposes on the

outcomes. In this paper, we propose a novel regularization framework in which we formalize unconfoundedness as an orthogonality constraint. We provide theoretical guarantees that this yields an asymptotically normal estimator for the average causal effect. Compared to other estimators, its asymptotic variance is strictly smaller. Based on our regularization framework, we develop deep orthogonal networks for unconfounded treatments (DONUT) which learn outcomes that are orthogonal to the treatment assignment. Using a variety of benchmark datasets for causal inference, we demonstrate that DONUT outperforms the state-of-the-art substantially.
****
On Nondeterminism and Instability in Neural Network Optimization
Cecilia Summers,Michael J. Dinneen
Optimization nondeterminism causes uncertainty when improving neural networks, with small changes in performance difficult to discern from run-to-run variability. While uncertainty can be reduced by training multiple copies of a model with different random seeds, doing so is time-consuming, costly, and makes reproducibility challenging. Despite this, little attention has been paid towards establishing an understanding of this problem. In this work, we establish an experimental protocol for understanding the effect of optimization nondeterminism on model diversity, which allows us to study the independent effects of a variety of sources of nondeterminism. Surprisingly, we find that each source of nondeterminism all have similar effects on multiple measures of model diversity. To explain this intriguing fact, we examine and identify the instability of model training, when taken as an end-to-end procedure, as the key determinant. We show that even one-bit changes in initial model parameters result in models that converge to vastly different values. Last, we demonstrate that recent methods in accelerated model ensembling hold promise for reducing the effects of instability on run-to-run variability.
****
Novelty Detection via Robust Variational Autoencoding
Chieh-Hsin Lai,Dongmian Zou,Gilad Lerman
We propose a new method for novelty detection that can tolerate high corruption of the training points, whereas previous works assumed either no or very low corruption. Our method trains a robust variational autoencoder (VAE), which aims to generate a model for the uncorrupted training points. To gain robustness to high corruption, we incorporate the following four changes to the common VAE: 1. Extracting crucial features of the latent code by a carefully designed dimension reduction component for distributions; 2. Modeling the latent distribution as a mixture of Gaussian low-rank inliers and full-rank outliers, where the testing only uses the inlier model; 3. Applying the Wasserstein-1 metric for regularization, instead of the Kullback-Leibler (KL) divergence; and 4. Using a least absolute deviation error for reconstruction. We establish both robustness to outliers and suitability to low-rank modeling of the Wasserstein metric as opposed to the KL divergence. We illustrate state-of-the-art results on standard benchmarks for novelty detection.
****
A Unified View on Graph Neural Networks as Graph Signal Denoising
Yao Ma,Xiaorui Liu,Tong Zhao,Yozen Liu,Jiliang Tang,Neil Shah
Graph Neural Networks (GNNs) have risen to prominence in learning representations for graph structured data. A single GNN layer typically consists of a feature transformation and a feature aggregation operation. The former normally uses feed-forward networks to transform features, while the latter aggregates the transformed features over the graph. Numerous recent works have proposed GNN models with different designs in the aggregation operation. In this work, we establish mathematically that the aggregation processes in a group of representative GNN models including GCN, GAT, PPNP, and APPNP can be regarded as (approximately) solving a graph denoising problem with a smoothness assumption. Such a unified view across GNNs not only provides a new perspective to understand a variety of aggregation operations but also enables us to develop a unified graph neural network framework UGNN. To demonstrate its promising potential, we instantiate a novel GN

N model, ADA-UGNN, derived from UGNN, to handle graphs with adaptive smoothness across nodes. Comprehensive experiments show the effectiveness of ADA-UGNN.
****

Physics-aware Spatiotemporal Modules with Auxiliary Tasks for Meta-Learning
Sungyong Seo,Chuizheng Meng,Sirisha Rambhatla,Yan Liu
Modeling the dynamics of real-world physical systems is critical for spatiotemporal prediction tasks, but challenging when data is limited. The scarcity of real-world data and the difficulty in reproducing the data distribution hinder directly applying meta-learning techniques. Although the knowledge of governing partial differential equations (PDE) of the data can be helpful for the fast adaptation to few observations, it is mostly infeasible to exactly find the equation for observations in real-world physical systems. In this work, we propose a framework, physics-aware meta-learning with auxiliary tasks whose spatial modules incorporate PDE-independent knowledge and temporal modules utilize the generalized features from the spatial modules to be adapted to the limited data, respectively. The framework is inspired by a local conservation law expressed mathematically as a continuity equation and does not require the exact form of governing equation to model the spatiotemporal observations. The proposed method mitigates the need for a large number of real-world tasks for meta-learning by leveraging spatial information in simulated data to meta-initialize the spatial modules. We apply the proposed framework to both synthetic and real-world spatiotemporal prediction tasks and demonstrate its superior performance with limited observations.
****

Deep Q-Learning with Low Switching Cost
Shusheng Xu,Simon Shaolei Du,Yi Wu
We initiate the study on deep reinforcement learning problems that require low switching cost, i.e., small number of policy switches during training. Such a requirement is ubiquitous in many applications, such as medical domains, recommendation systems, education, robotics, dialogue agents, etc, where the deployed policy that actually interacts with the environment cannot change frequently. Our paper investigates different policy switching criteria based on deep Q-networks and further proposes an adaptive approach based on the feature distance between the deployed Q-network and the underlying learning Q-network. Through extensive experiments on a medical treatment environment and a collection of the Atari games, we find our feature-switching criterion substantially decreases the switching cost while maintains a similar sample efficiency to the case without the low-switching-cost constraint. We also complement this empirical finding with a theoretical justification from a representation learning perspective.
****

Constraint-Driven Explanations of Black-Box ML Models
Aditya Aniruddha Shrotri,Nina Narodytska,Alexey Ignatiev,Joao Marques-Silva,Kuldeep S. Meel,Moshe Vardi
Modern machine learning techniques have enjoyed widespread success, but are plagued by lack of transparency in their decision making, which has led to the emergence of the field of explainable AI. One popular approach called LIME, seeks to explain an opaque model's behavior, by training a surrogate interpretable model to be locally faithful on perturbed instances.
Despite being model-agnostic and easy-to-use, it is known that LIME's explanations can be unstable and are susceptible to adversarial attacks as a result of Out-Of-Distribution (OOD) sampling. Quality of explanations is also calculated heuristically, and lacks a strong theoretical foundation. In spite of numerous attempts to remedy some of these issues, making the LIME framework more trustworthy and reliable remains an open problem.

In this work, we demonstrate that the OOD sampling problem stems from rigidity of the perturbation procedure. To resolve this issue, we propose a theoretically sound framework based on uniform sampling of user-defined subspaces. Through logical constraints, we afford the end-user the flexibility to delineate the precise subspace of the input domain to be explained. This not only helps mitigate the problem of OOD sampling, but also allow experts to drill down and uncover bug

s deep inside the model. For testing the quality of generated explanations, we d
evelop an efficient estimation algorithm that is able to certifiably measure the
 true value of metrics such as fidelity up to any desired degree of accuracy, wh
ich can help in building trust in the generated explanations. Our framework call
ed CLIME can be applied to any ML model, and extensive experiments demonstrate i
ts versatility on real-world problems.

****
Learning to Make Decisions via Submodular Regularization
Ayya Alieva,Aiden Aceves,Jialin Song,Stephen Mayo,Yisong Yue,Yuxin Chen
Many sequential decision making tasks can be viewed as combinatorial optimizatio
n problems over a large number of actions. When the cost of evaluating an action
 is high, even a greedy algorithm, which iteratively picks the best action given
 the history, is prohibitive to run. In this paper, we aim to learn a greedy heu
ristic for sequentially selecting actions as a surrogate for invoking the expens
ive oracle when evaluating an action. In particular, we focus on a class of comb
inatorial problems that can be solved via submodular maximization (either direct
ly on the objective function or via submodular surrogates). We introduce a data-
driven optimization framework based on the submodular-norm loss, a novel loss fu
nction that encourages the resulting objective to exhibit diminishing returns. O
ur framework outputs a surrogate objective that is efficient to train, approxima
tely submodular, and can be made permutation-invariant. The latter two propertie
s allow us to prove strong approximation guarantees for the learned greedy heuri
stic. Furthermore, we show that our model can be easily integrated with modern d
eep imitation learning pipelines for sequential prediction tasks. We demonstrate
 the performance of our algorithm on a variety of batched and sequential optimiz
ation tasks, including set cover, active learning, and Bayesian optimization for
 protein engineering.
****
Intelligent Matrix Exponentiation
Thomas Fischbacher,Iulia Maria Comsa,Krzysztof Potempa,Moritz Firsching,Luca Ver
sari,Jyrki Alakuijala
We present a novel machine learning architecture that uses a single high-dimensi
onal nonlinearity consisting of the exponential of a single input-dependent matr
ix. The mathematical simplicity of this architecture allows a detailed analysis
of its behaviour, providing robustness guarantees via Lipschitz bounds. Despite
its simplicity, a single matrix exponential layer already provides universal app
roximation properties and can learn and extrapolate fundamental functions of the
 input, such as periodic structure or geometric invariants. This architecture ou
tperforms other general-purpose architectures on benchmark problems, including C
IFAR-10, using fewer parameters.
****
The large learning rate phase of deep learning
Aitor Lewkowycz,Yasaman Bahri,Ethan Dyer,Jascha Sohl-Dickstein,Guy Gur-Ari
The choice of initial learning rate can have a profound effect on the performanc
e of deep networks. We present empirical evidence that networks exhibit sharply
distinct behaviors at small and large learning rates. In the small learning rate
 phase, training can be understood using the existing theory of infinitely wide
neural networks. At large learning rates, we find that networks exhibit qualitat
ively distinct phenomena that cannot be explained by existing theory: The loss g
rows during the early part of training, and optimization eventually converges to
 a flatter minimum. Furthermore, we find that the optimal performance is often f
ound in the large learning rate phase. To better understand this behavior we ana
lyze the dynamics of a two-layer linear network and prove that it exhibits these
 different phases. We find good agreement between our analysis and the training
dynamics observed in realistic deep learning settings.

****
SEQUENCE-LEVEL FEATURES: HOW GRU AND LSTM CELLS CAPTURE N-GRAMS

Xiaobing Sun,Wei Lu
Modern recurrent neural networks (RNN) such as Gated Recurrent Units (GRU) and L
ong Short-term Memory (LSTM) have demonstrated impressive results on tasks invol
ving sequential data in practice. Despite continuous efforts on interpreting the
ir behaviors, the exact mechanism underlying their successes in capturing sequen
ce-level information have not been thoroughly understood. In this work, we prese
nt a study on understanding the essential features captured by GRU/LSTM cells by
 mathematically expanding and unrolling the hidden states. Based on the expanded
 and unrolled hidden states, we find there was a type of sequence-level represen
tations brought in by the gating mechanism, which enables the cells to encode se
quence-level features along with token-level features. Specifically, we show tha
t the cells would consist of such sequence-level features similar to those of N-
grams. Based on such a finding, we also found that replacing the hidden states o
f the standard cells with N-gram representations does not necessarily degrade pe
rformance on the sentiment analysis and language modeling tasks, indicating such
 features may play a significant role for GRU/LSTM cells.
****

UserBERT: Self-supervised User Representation Learning
Tianyu Li,Ali Cevahir,Derek Cho,Hao Gong,DuyKhuong Nguyen,Bjorn Stenger
This paper extends the BERT model to user data for pretraining user representati
ons in a self-supervised way. By viewing actions (e.g., purchases and clicks) in
 behavior sequences (i.e., usage history) in an analogous way to words in senten
ces, we propose methods for the tokenization, the generation of input representa
tion vectors and a novel pretext task to enable the pretraining model to learn f
rom its own input, omitting the burden of collecting additional data. Further, o
ur model adopts a unified structure to simultaneously learn from long-term and s
hort-term user behavior as well as user profiles. Extensive experiments demonstr
ate that the learned representations result in significant improvements when tr
ansferred to three different real-world tasks, particularly in comparison with t
ask-specific modeling and representations obtained from multi-task learning.
****

The Recurrent Neural Tangent Kernel
Sina Alemohammad,Zichao Wang,Randall Balestriero,Richard Baraniuk
The study of deep neural networks (DNNs) in the infinite-width limit, via the so
-called neural tangent kernel (NTK) approach, has provided new insights into the
 dynamics of learning, generalization, and the impact of initialization. One key
 DNN architecture remains to be kernelized, namely, the recurrent neural network
 (RNN). In this paper we introduce and study the Recurrent Neural Tangent Kerne
l (RNTK), which provides new insights into the behavior of overparametrized RNN
s. A key property of the RNTK should greatly benefit practitioners is its abilit
y to compare inputs of different length. To this end, we characterize how the RN
TK weights different time steps to form its output under different initializatio
n parameters and nonlinearity choices. A synthetic and 56 real-world data experi
ments demonstrate that the RNTK offers significant performance gains over other
kernels, including standard NTKs, across a wide array of data sets.
****

Learning to Use Future Information in Simultaneous Translation
Xueqing Wu,Yingce Xia,Lijun Wu,Shufang Xie,Weiqing Liu,Tao Qin,Tie-Yan Liu
Simultaneous neural machine translation (briefly, NMT) has attracted much attent
ion recently. In contrast to standard NMT, where the NMT system can access the f
ull input sentence, simultaneous NMT is a prefix-to-prefix problem, where the sy
stem can only utilize the prefix of the input sentence and thus more uncertainty
 and difficulty are introduced to decoding. Wait-k inference is a simple yet eff
ective strategy for simultaneous NMT, where the decoder generates the output seq
uence k words behind the input words. For wait-k inference, we observe that wa
it-m training with $m>k$ in simultaneous NMT (i.e., using more future informatio
n for training than inference) generally outperforms wait-k training. Based on t
his observation, we propose a method that automatically learns how much future i
nformation to use in training for simultaneous NMT. Specifically, we introduce a
 controller to adaptively select wait-m training strategies according to the net

work status of the translation model and current training sentence pairs, and th e controller is jointly trained with the translation model through bi-level opti mization. Experiments on four datasets show that our method brings 1 to 3 BLEU p oint improvement over baselines under the same latency. Our code is available at https://github.com/P2F-research/simulNMT .

****

Understanding the Effect of Bias in Deep Anomaly Detection

Ziyu Ye,Yuxin Chen,Haitao Zheng

Anomaly detection presents a unique challenge in machine learning, due to the sc arcity of labeled anomaly data. Recent work attempts to mitigate such problems b y augmenting training of deep anomaly detection models with additional labeled a nomaly samples. However, the labeled data often does not align with the target d istribution and introduces harmful bias to the trained model. In this paper, we aim to understand the effect of a biased anomaly set on anomaly detection. We fo rmally state the anomaly detection problem as a supervised learning task, and fo cus on the anomaly detector’s recall at a given false positive rate as the main performance metric. Given two different anomaly score functions, we formally def ine their difference in performance as the relative scoring bias of the anomaly detectors. Along this line, our work provides two key contributions. We establis h the first finite sample rates for estimating the relative scoring bias for dee p anomaly detection, and empirically validate our theoretical results on both sy nthetic and real-world datasets. We also provide extensive empirical study on ho w a biased training anomaly set affects the anomaly score function and therefore the detection performance on different anomaly classes. Our study demonstrates scenarios in which the biased anomaly set can be useful or problematic, and prov ides a solid benchmark for future research.

****

Jumpy Recurrent Neural Networks

Samuel James Greydanus,Stefan Lee,Alan Fern

Recurrent neural networks (RNNs) can learn complex, long-range structure in time series data simply by predicting one point at a time. Because of this ability, they have enjoyed widespread adoption in commercial and academic contexts. Yet R NNs have a fundamental limitation: they represent time as a series of discrete, uniform time steps. As a result, they force a tradeoff between temporal resoluti on and the computational expense of predicting far into the future. To resolve t his tension, we propose a Jumpy RNN model which does not predict state transitio ns over uniform intervals of time. Instead, it predicts a sequence of linear dyn amics functions in latent space and intervals of time over which their predictio ns can be expected to be accurate. This structure enables our model to jump over long time intervals while retaining the ability to produce fine-grained or cont inuous-time predictions when necessary. In simple physics simulations, our model can skip over long spans of predictable motion and focus on key events such as collisions between two balls. On a set of physics tasks including coordinate and pixel observations of a small-scale billiards environment, our model matches th e performance of a baseline RNN while using a fifth of the compute. On a real-wo rld weather forecasting dataset, it makes more accurate predictions while using fewer sampling steps. When used for model-based planning, our method matches a b aseline RNN while using half the compute.

****

Architecture Agnostic Neural Networks

Sabera J Talukder,Guruprasad Raghavan,Yisong Yue

In this paper, we explore an alternate method for synthesizing neural network ar chitectures, inspired by the brain's stochastic synaptic pruning. During a perso n’s lifetime, numerous distinct neuronal architectures are responsible for perfo rming the same tasks. This indicates that biological neural networks are, to som e degree, architecture agnostic. However, artificial networks rely on their fine -tuned weights and hand-crafted architectures for their remarkable performance. This contrast begs the question: Can we build artificial architecture agnostic n eural networks? To ground this study we utilize sparse, binary neural networks t hat parallel the brain’s circuits. Within this sparse, binary paradigm we sample

many binary architectures to create families of architecture agnostic neural networks not trained via backpropagation. These high-performing network families share the same sparsity, distribution of binary weights, and succeed in both static and dynamic tasks. In summation, we create an architecture manifold search procedure to discover families of architecture agnostic neural networks.
****
Low Complexity Approximate Bayesian Logistic Regression for Sparse Online Learning
Gil I. Shamir,Wojciech Szpankowski
Theoretical results show that Bayesian methods can achieve lower bounds on regret for online logistic regression. In practice, however, such techniques may not be feasible especially for very large feature sets. Various approximations that, for huge sparse feature sets, diminish the theoretical advantages, must be used. Often, they apply stochastic gradient methods with hyper-parameters that must be tuned on some surrogate loss, defeating theoretical advantages of Bayesian methods. The surrogate loss, defined to approximate the mixture, requires techniques as Monte Carlo sampling, increasing computations per example. We propose low complexity analytical approximations for sparse online logistic and probit regressions. Unlike variational inference and other methods, our methods use analytical closed forms, substantially lowering computations. Unlike dense solutions,
as Gaussian Mixtures, our methods allow for sparse problems with huge feature sets without increasing complexity. With the analytical closed forms, there is also no need for applying stochastic gradient methods on surrogate losses, and for tuning and balancing learning and regularization hyper-parameters. Empirical results top the performance of the more computationally involved methods. Like such methods, our methods still reveal per feature and per example uncertainty measures.

****
Predicting the impact of dataset composition on model performance
Tatsunori Hashimoto
Real-world machine learning systems are often are trained using a mix of data sources with varying cost and quality. Understanding how the size and composition of a training dataset affect model performance is critical for advancing our understanding of generalization, as well as designing more effective data collection policies. We show that there is a simple, accurate way to predict the loss incurred by a model based on data size and composition. Our work expands recent observations of log-linear generalization error and uses this to cast model performance prediction as a learning problem. Using the theory of optimal experimental design, we derive a simple rational function approximation to generalization error that can be fitted using a few model training runs. Our approach achieves nearly exact ($r^2>.93$) predictions of model performance under substantial extrapolation in two different standard supervised learning tasks and is accurate ($r^2 > .83$) on more challenging machine translation and question answering tasks where baselines achieve worse-than-random performance.
****
SBEVNet: End-to-End Deep Stereo Layout Estimation
Divam Gupta,Wei Pu,Trenton Tabor,Jeff Schneider
Accurate layout estimation is crucial for planning and navigation, for robotics applications such as self driving. In this paper, we introduce stereo bird's eye view network SBEVNet, a novel supervised end-to-end framework for estimation of bird's eye view layout from a pair of stereo images. Although our network reuses the building blocks from the state-of-the-art deep learning networks for disparity estimation, we show that accurate depth estimation is neither sufficient nor necessary. Instead, the learning of a good internal bird's eye view feature representation is essential for layout estimation. Specifically, we first generate a disparity feature volume using the features of the stereo images and then project it to the bird's eye view coordinates. This gives us coarse grained scene structural information. We also apply inverse perspective mapping (IPM) to map th

e input images and their features to the bird's eye view. This gives us fine gra
ined texture information. The concatenated IPM features with the projected featu
re volume creates a rich bird's eye view representation which is capable of spat
ial reasoning. We use this representation to estimate the BEV semantic map. Addi
tionally, we show that using the IPM features as a supervisory signal for stereo
 features can give an improvement in performance. We demonstrate our approach on
 two datasets: KITTI dataset and synthetically generated dataset using the CARLA
 simulator. For both of the datasets, we establish state-of-the-art performance
beyond other baselines.
****

Why Are Convolutional Nets More Sample-Efficient than Fully-Connected Nets?
Zhiyuan Li,Yi Zhang,Sanjeev Arora
Convolutional neural networks often dominate fully-connected counterparts in gen
eralization performance, especially on image classification tasks. This is often
 explained in terms of \textquotedblleft better inductive bias.\textquotedblrigh
t\ However, this has not been made mathematically rigorous, and the hurdle is t
hat the sufficiently wide fully-connected net can always simulate the convolutio
nal net. Thus the training algorithm plays a role. The current work describes a
natural task on which a provable sample complexity gap can be shown, for standar
d training algorithms. We construct a single natural distribution on $\mathbb{R}
^d\times\{\pm 1\}$ on which any orthogonal-invariant algorithm (i.e. fully-conne
cted networks trained with most gradient-based methods from gaussian initializat
ion) requires $\Omega(d^2)$ samples to generalize while $O(1)$ samples suffice f
or convolutional architectures. Furthermore, we demonstrate a single target func
tion, learning which on all possible distributions leads to an $O(1)$ vs $\Omega
(d^2/\varepsilon)$ gap. The proof relies on the fact that SGD on fully-connected
 network is orthogonal equivariant. Similar results are achieved for ℓ_2 re
gression and adaptive training algorithms, e.g. Adam and AdaGrad, which are only
 permutation equivariant.
****

Deep Q Learning from Dynamic Demonstration with Behavioral Cloning
Xiaoshuang Li,Junchen Jin,Xiao Wang,Fei-Yue Wang
    Although Deep Reinforcement Learning (DRL) has proven its capability to lear
n optimal policies by directly interacting with simulation environments, how to
combine DRL with supervised learning and leverage additional knowledge to assist
 the DRL agent effectively still remains difficult. This study proposes a novel
approach integrating deep Q learning from dynamic demonstrations with a behavior
al cloning model (DQfDD-BC), which includes a supervised learning technique of i
nstructing a DRL model to enhance its performance. Specifically, the DQfDD-BC mo
del leverages historical demonstrations to pre-train a supervised BC model and c
onsistently update it by learning the dynamically updated demonstrations. Then t
he DQfDD-BC model manages the sample complexity by exploiting both the historica
l and generated demonstrations. An expert loss function is designed to compare a
ctions generated by the DRL model with those obtained from the BC model to provi
de advantageous guidance for policy improvements. Experimental results in severa
l OpenAI Gym environments show that the proposed approach adapts to different pe
rformance levels of demonstrations, and meanwhile, accelerates the learning proc
esses. As illustrated in an ablation study, the dynamic demonstration and expert
 loss mechanisms with the utilization of a BC model contribute to improving the
learning convergence performance compared with the origin DQfD model.
****

Linear Representation Meta-Reinforcement Learning for Instant Adaptation
Matt Peng,Banghua Zhu,Jiantao Jiao
This paper introduces Fast Linearized Adaptive Policy (FLAP), a new meta-reinfor
cement learning (meta-RL) method that is able to extrapolate well to out-of-dist
ribution tasks without the need to reuse data from training, and adapt almost in
stantaneously with the need of only a few samples during testing. FLAP builds up
on the idea of learning a shared linear representation of the policy so that whe
n adapting to a new task, it suffices to predict a set of linear weights. A sepa
rate adapter network is trained simultaneously with the policy such that during

adaptation, we can directly use the adapter network to predict these linear weig hts instead of updating a meta-policy via gradient descent such as in prior Meta -RL algorithms like MAML to obtain the new policy. The application of the separ ate feed-forward network not only speeds up the adaptation run-time significantl y, but also generalizes extremely well to very different tasks that prior Meta-R L methods fail to generalize to. Experiments on standard continuous-control meta -RL benchmarks show FLAP presenting significantly stronger performance on out-of -distribution tasks with up to double the average return and up to 8X faster ada ptation run-time speeds when compared to prior methods.
****

Evaluation of Similarity-based Explanations
Kazuaki Hanawa,Sho Yokoi,Satoshi Hara,Kentaro Inui
Explaining the predictions made by complex machine learning models helps users t o understand and accept the predicted outputs with confidence. One promising way is to use similarity-based explanation that provides similar instances as evide nce to support model predictions. Several relevance metrics are used for this pu rpose. In this study, we investigated relevance metrics that can provide reasona ble explanations to users. Specifically, we adopted three tests to evaluate whet her the relevance metrics satisfy the minimal requirements for similarity-based explanation. Our experiments revealed that the cosine similarity of the gradient s of the loss performs best, which would be a recommended choice in practice. In addition, we showed that some metrics perform poorly in our tests and analyzed the reasons of their failure. We expect our insights to help practitioners in se lecting appropriate relevance metrics and also aid further researches for design ing better relevance metrics for explanations.
****

Adaptive Procedural Task Generation for Hard-Exploration Problems
Kuan Fang,Yuke Zhu,Silvio Savarese,L. Fei-Fei
We introduce Adaptive Procedural Task Generation (APT-Gen), an approach to progr essively generate a sequence of tasks as curricula to facilitate reinforcement l earning in hard-exploration problems. At the heart of our approach, a task gener ator learns to create tasks from a parameterized task space via a black-box proc edural generation module. To enable curriculum learning in the absence of a dire ct indicator of learning progress, we propose to train the task generator by bal ancing the agent's performance in the generated tasks and the similarity to the target tasks. Through adversarial training, the task similarity is adaptively es timated by a task discriminator defined on the agent's experiences, allowing the generated tasks to approximate target tasks of unknown parameterization or outs ide of the predefined task space. Our experiments on the grid world and robotic manipulation task domains show that APT-Gen achieves substantially better perfor mance than various existing baselines by generating suitable tasks of rich varia tions.
****

The impacts of known and unknown demonstrator irrationality on reward inference
Lawrence Chan,Andrew Critch,Anca Dragan
Algorithms inferring rewards from human behavior typically assume that people ar e (approximately) rational. In reality, people exhibit a wide array of irrationa lities. Motivated by understanding the benefits of modeling these irrationalitie s, we analyze the effects that demonstrator irrationality has on reward inferenc e. We propose operationalizing several forms of irrationality in the language of MDPs, by altering the Bellman optimality equation, and use this framework to st udy how these alterations affect inference.

We find that incorrectly assuming noisy-rationality for an irrational demonstrat or can lead to remarkably poor reward inference accuracy, even in situations whe re inference with the correct model leads to good inference. This suggests a nee d to either model irrationalities or find reward inference algorithms that are m ore robust to misspecification of the demonstrator model. Surprisingly, we find that if we give the learner access to the correct model of the demonstrator's ir rationality, these irrationalities can actually help reward inference. In other

words, if we could choose between a world where humans were perfectly rational a
nd the current world where humans have systematic biases, the current world migh
t counter-intuitively be preferable for reward inference. We reproduce this effe
ct in several domains. While this finding is mainly conceptual, it is perhaps ac
tionable as well: we might ask human demonstrators for myopic demonstrations ins
tead of optimal ones, as they are more informative for the learner and might be
easier for a human to generate.
****
Implicit bias of gradient descent for mean squared error regression with wide ne
ural networks
Hui Jin,Guido Montufar
We investigate gradient descent training of wide neural networks and the corresp
onding implicit bias in function space. For 1D regression, we show that the solu
tion of training a width-n shallow ReLU network is within $n^{- 1/2}$ of the f
unction which fits the training data and whose difference from initialization ha
s smallest 2-norm of the second derivative weighted by $1/\zeta$. The curvature
penalty function $1/\zeta$ is expressed in terms of the probability distribution
 that is utilized to initialize the network parameters, and we compute it explic
itly for various common initialization procedures. For instance, asymmetric init
ialization with a uniform distribution yields a constant curvature penalty, and
thence the solution function is the natural cubic spline interpolation of the tr
aining data. While similar results have been obtained in previous works, our ana
lysis clarifies important details and allows us to obtain significant generaliza
tions. In particular, the result generalizes to multivariate regression and diff
erent activation functions. Moreover, we show that the training trajectories are
 captured by trajectories of spatially adaptive smoothing splines with decreasin
g regularization strength.
****
Linear Last-iterate Convergence in Constrained Saddle-point Optimization
Chen-Yu Wei,Chung-Wei Lee,Mengxiao Zhang,Haipeng Luo
Optimistic Gradient Descent Ascent (OGDA) and Optimistic Multiplicative Weights
Update (OMWU) for saddle-point optimization have received growing attention due
to their favorable last-iterate convergence. However, their behaviors for simple
 bilinear games over the probability simplex are still not fully understood ---
previous analysis lacks explicit convergence rates, only applies to an exponenti
ally small learning rate, or requires additional assumptions such as the uniquen
ess of the optimal solution.

In this work, we significantly expand the understanding of last-iterate converge
nce for OGDA and OMWU in the constrained setting. Specifically, for OMWU in bili
near games over the simplex, we show that when the equilibrium is unique, linear
 last-iterate convergence is achievable with a constant learning rate, which imp
roves the result of (Daskalakis & Panageas, 2019) under the same assumption. We
then significantly extend the results to more general objectives and feasible se
ts for the projected OGDA algorithm, by introducing a sufficient condition under
 which OGDA exhibits concrete last-iterate convergence rates with a constant lea
rning rate. We show that bilinear games over any polytope satisfy this condition
 and OGDA converges exponentially fast even without the unique equilibrium assum
ption. Our condition also holds for strongly-convex-strongly-concave functions,
recovering the result of (Hsieh et al., 2019). Finally, we provide experimental
results to further support our theory.
****
Memory Representation in Transformer
Mikhail Burtsev,Yurii Kuratov,Anton Peganov,Grigory V. Sapunov
Transformer-based models have achieved state-of-the-art results in many natural
language processing tasks. The self-attention architecture allows transformer to
 combine information from all elements of a sequence into context-aware represen
tations. However, information about the context is stored mostly in the same ele
ment-wise representations. This might limit the processing of properties related
 to the sequence as a whole more difficult. Adding trainable memory to selective

ly store local as well as global representations of a sequence is a promising di rection to improve the Transformer model. Memory-augmented neural networks (MANN s) extend traditional neural architectures with general-purpose memory for repre sentations. MANNs have demonstrated the capability to learn simple algorithms li ke Copy or Reverse and can be successfully trained via backpropagation on divers e tasks from question answering to language modeling outperforming RNNs and LSTM s of comparable complexity. In this work, we propose and study few extensions of the Transformer baseline (1) by adding memory tokens to store non-local represe ntations, (2) creating memory bottleneck for the global information, (3) control ling memory update with dedicated layer. We evaluate these memory augmented Tran sformers and demonstrate that presence of memory positively correlates with the model performance for machine translation and language modelling tasks. Augmenta tion of pre-trained masked language model with memory tokens shows mixed results for tasks from GLUE benchmark. Visualization of attention patterns over the mem ory suggest that it improves the model's ability to process a global context.
****

On Graph Neural Networks versus Graph-Augmented MLPs
Lei Chen,Zhengdao Chen,Joan Bruna
From the perspectives of expressive power and learning, this work compares multi -layer Graph Neural Networks (GNNs) with a simplified alternative that we call G raph-Augmented Multi-Layer Perceptrons (GA-MLPs), which first augments node feat ures with certain multi-hop operators on the graph and then applies learnable no de-wise functions. From the perspective of graph isomorphism testing, we show bo th theoretically and numerically that GA-MLPs with suitable operators can distin guish almost all non-isomorphic graphs, just like the Weisfeiler-Lehman (WL) tes t and GNNs. However, by viewing them as node-level functions and examining the e quivalence classes they induce on rooted graphs, we prove a separation in expres sive power between GA-MLPs and GNNs that grows exponentially in depth. In partic ular, unlike GNNs, GA-MLPs are unable to count the number of attributed walks. W e also demonstrate via community detection experiments that GA-MLPs can be limit ed by their choice of operator family, whereas GNNs have higher flexibility in l earning.
****

Improving Hierarchical Adversarial Robustness of Deep Neural Networks
Avery Ma,Aladin Virmaux,Kevin Scaman,Juwei Lu
Do all adversarial examples have the same consequences? An autonomous driving sy stem misclassifying a pedestrian as a car may induce a far more dangerous --and even potentially lethal-- behavior than, for instance, a car as a bus. In order to better tackle this important problematic, we introduce the concept of hierarc hical adversarial robustness. Given a dataset whose classes can be grouped into coarse-level labels, we define hierarchical adversarial examples as the ones lea ding to a misclassification at the coarse level. To improve the resistance of ne ural networks to hierarchical attacks, we introduce a hierarchical adversarially robust (HAR) network design that decomposes a single classification task into o ne coarse and multiple fine classification tasks, before being specifically trai ned by adversarial defense techniques. As an alternative to an end-to-end learni ng approach, we show that HAR significantly improves the robustness of the netwo rk against ℓ_{∞} and ℓ_{2}bounded hierarchical attacks on CIFAR-100.
****

CorDial: Coarse-to-fine Abstractive Dialogue Summarization with Controllable Gra nularity
Chien-Sheng Wu,Linqing Liu,Wenhao Liu,Pontus Stenetorp,Caiming Xiong
Dialogue summarization is challenging due to its multi-speaker standpoints, casu al spoken language, and limited labeled data. In this paper, we propose CorDial, aiming to improve the abstractive dialogue summarization quality and at the sam e time enable granularity controllability. We propose 1) a coarse-to-fine genera tion strategy that generates a summary draft followed by a final summary in an a utoregressive way. The summary draft, which provides weakly-supervised signals, is composed of pseudo-labeled interrogative pronoun categories and noisy key phr

ases extracted with a constituency parser. 2) A simple strategy to control the g
ranularity of the final summary. CorDial can predict and control the number of s
ummary sentences for a given dialogue by predicting and highlighting different t
ext spans from the source text. Our model achieves state-of-the-art performance
on the largest dialogue summarization corpus SAMSum. We conduct comprehensive er
ror analysis and show competitive human evaluation results to annotated summarie
s.
****
Meta-Reinforcement Learning Robust to Distributional Shift via Model Identificat
ion and Experience Relabeling
Russell Mendonca,Xinyang Geng,Chelsea Finn,Sergey Levine
Reinforcement learning algorithms can acquire policies for complex tasks autonom
ously. However, the number of samples required to learn a diverse set of skills
can be prohibitively large. While meta-reinforcement learning methods have enabl
ed agents to leverage prior experience to adapt quickly to new tasks, their perf
ormance depends crucially on how close the new task is to the previously experie
nced tasks. Current approaches are either not able to extrapolate well, or can d
o so at the expense of requiring extremely large amounts of data for on-policy m
eta-training. In this work, we present model identification and experience relab
eling (MIER), a meta-reinforcement learning algorithm that is both efficient and
 extrapolates well when faced with out-of-distribution tasks at test time. Our m
ethod is based on a simple insight: we recognize that dynamics models can be ada
pted efficiently and consistently with off-policy data, more easily than policie
s and value functions. These dynamics models can then be used to continue traini
ng policies and value functions for out-of-distribution tasks without using meta
-reinforcement learning at all, by generating synthetic experience for the new t
ask.
****
Deep Positive Unlabeled Learning with a Sequential Bias
Walter Gerych,Thomas Hartvigsen,Luke Buquicchio,Kavin Chandrasekaran,Hamid Manso
or,Abdulaziz alajaji
For many domains, from video stream analytics to human activity recognition, onl
y weakly-labeled datasets are available.
Worse yet, the given labels are often assigned sequentially, resulting in sequen
tial bias. Current Positive Unlabeled (PU) classifiers, a state-of-the-art famil
y of robust semi-supervised methods, are ineffective under sequential bias. In t
his work, we propose DeepSPU, the first method to address this sequential bias
problem. DeepSPU tackles the two interdependent subproblems of learning both the
 latent labeling process and the true class likelihoods within one architecture.
 We achieve this by developing a novel iterative learning strategy aided by theo
retically-justified cost terms to avoid collapsing into a naive classifier. Our
experimental studies demonstrate that DeepSPU outperforms state-of-the-art metho
ds by over 10% on diverse real-world datasets.
****
An Efficient Protocol for Distributed Column Subset Selection in the Entrywise $
\ell_p$ Norm
Shuli Jiang,Dongyu Li,Irene Mengze Li,Arvind V. Mahankali,David Woodruff
We give a distributed protocol with nearly-optimal communication and number of r
ounds for Column Subset Selection with respect to the entrywise {ℓ_1} norm
(k-CSS$_1$), and more generally, for the ℓ_p-norm with $1 \leq p < 2$. We
 study matrix factorization in ℓ_1-norm loss, rather than the more standard
 Frobenius norm loss, because the ℓ_1 norm is more robust to noise, which i
s observed to lead to improved performance in a wide range of computer vision an
d robotics problems.
In the distributed setting, we consider s servers in the standard coordinator
model of communication, where the columns of the input matrix $A \in \mathbb{R}^
{d \times n}$ ($n \gg d$) are distributed across the s servers. We give a prot
ocol in this model with $\widetilde{O}(sdk)$ communication, 1 round, and polyn
omial running time, and which achieves a multiplicative $k^{\frac{1}{p} - \frac{
1}{2}}\poly(\log nd)$-approximation to the best possible column subset. A key in

gredient in our proof is the reduction to the $\ell_{p,2}$-norm, which correspon ds to the p-norm of the vector of Euclidean norms of each of the columns of $A $. This enables us to use strong coreset constructions for Euclidean norms, whic h previously had not been used in this context. This naturally also allows us to implement our algorithm in the popular streaming model of computation. We furth er propose a greedy algorithm for selecting columns, which can be used by the co ordinator, and show the first provable guarantees for a greedy algorithm for the $\ell_{1,2}$ norm. Finally, we implement our protocol and give significant prac tical advantages on real-world data analysis tasks.
****
Policy Optimization in Zero-Sum Markov Games: Fictitious Self-Play Provably Atta ins Nash Equilibria
Boyi Liu,Zhuoran Yang,Zhaoran Wang
Fictitious Self-Play (FSP) has achieved significant empirical success in solving extensive-form games.
However, from a theoretical perspective, it remains unknown whether FSP is guara nteed to converge to Nash equilibria in Markov games.
As an initial attempt, we propose an FSP algorithm for two-player zero-sum Marko v games, dubbed as smooth FSP, where both agents adopt an entropy-regularized po licy optimization method against each other.
Smooth FSP builds upon a connection between smooth fictitious play and the polic y optimization framework. Specifically, in each iteration, each player infers th e policy of the opponent implicitly via policy evaluation and improves its curre nt policy by taking the smoothed best-response via a proximal policy optimizatio n (PPO) step.
Moreover, to tame the non-stationarity caused by the opponent, we propose to inc orporate entropy regularization in PPO for algorithmic stability.
When both players adopt smooth FSP simultaneously, i.e., with self-play, we prov e that the sequence of joint policies converges to a neighborhood of a Nash equi librium at a sublinear $\tilde{O}(1/T)$ rate, where T is the number of iterati ons. To our best knowledge, we establish the first finite-time convergence guara ntee for FSP-type algorithms in zero-sum Markov games.
****
Localized Meta-Learning: A PAC-Bayes Analysis for Meta-Learning Beyond Global Pr ior
Chenghao Liu,Tao Lu,Doyen Sahoo,Yuan Fang,Kun Zhang,Steven Hoi
Meta-learning methods learn the meta-knowledge among various training tasks and aim to promote the learning of new tasks under the task similarity assumption. S uch meta-knowledge is often represented as a fixed distribution; this, however, may be too restrictive to capture various specific task information because the discriminative patterns in the data may change dramatically across tasks. In thi s work, we aim to equip the meta learner with the ability to model and produce t ask-specific meta-knowledge and, accordingly, present a localized meta-learning framework based on the PAC-Bayes theory. In particular, we propose a Local Coord inate Coding (LCC) based prior predictor that allows the meta learner to generat e local meta-knowledge for specific tasks adaptively. We further develop a pract ical algorithm with deep neural network based on the bound. Empirical results on real-world datasets demonstrate the efficacy of the proposed method.
****
Reinforcement Learning with Bayesian Classifiers: Efficient Skill Learning from Outcome Examples
Kevin Li,Abhishek Gupta,Vitchyr H. Pong,Ashwin Reddy,Aurick Zhou,Justin Yu,Serge y Levine
Exploration in reinforcement learning is, in general, a challenging problem. In this work, we study a more tractable class of reinforcement learning problems de fined by data that provides examples of successful outcome states. In this case, the reward function can be obtained automatically by training a classifier to c lassify states as successful or not. We argue that, with appropriate representat ion and regularization, such a classifier can guide a reinforcement learning alg orithm to an effective solution. However, as we will show, this requires the cla

ssifier to make uncertainty-aware predictions that are very difficult with stand ard deep networks. To address this, we propose a novel mechanism for obtaining c alibrated uncertainty based on an amortized technique for computing the normaliz ed maximum likelihood distribution. We show that the resulting algorithm has a n umber of intriguing connections to both count-based exploration methods and prio r algorithms for learning reward functions from data, while being able to guide algorithms towards the specified goal more effectively. We show how using amorti zed normalized maximum likelihood for reward inference is able to provide effect ive reward guidance for solving a number of challenging navigation and robotic m anipulation tasks which prove difficult for other algorithms.
****

Acoustic Neighbor Embeddings
Woojay Jeon
This paper proposes a novel acoustic word embedding called Acoustic Neighbor Emb eddings where speech or text of arbitrary length are mapped to a vector space of fixed, reduced dimensions by adapting stochastic neighbor embedding (SNE) to se quential inputs. The Euclidean distance between coordinates in the embedding spa ce reflects the phonetic confusability between their corresponding sequences. Tw o encoder neural networks are trained: an acoustic encoder that accepts speech s ignals in the form of frame-wise subword posterior probabilities obtained from a n acoustic model and a text encoder that accepts text in the form of subword tra nscriptions. Compared to a triplet loss criterion, the proposed method is shown to have more effective gradients for neural network training. Experimentally, it also gives more accurate results with low-dimensional embeddings when the two e ncoder networks are used in tandem in a word (name) recognition task, and when t he text encoder network is used standalone in an approximate phonetic matching t ask. In particular, in an isolated name recognition task depending solely on Euc lidean nearest-neighbor search between the proposed embedding vectors, the recog nition accuracy is identical to that of conventional finite state transducer(FST)-based decoding using test data with up to 1 million names in the vocabulary an d 40 dimensions in the embeddings.
****

Solving Compositional Reinforcement Learning Problems via Task Reduction
Yunfei Li,Yilin Wu,Huazhe Xu,Xiaolong Wang,Yi Wu
We propose a novel learning paradigm, Self-Imitation via Reduction (SIR), for so lving compositional reinforcement learning problems. SIR is based on two core id eas: task reduction and self-imitation. Task reduction tackles a hard-to-solve t ask by actively reducing it to an easier task whose solution is known by the RL agent. Once the original hard task is successfully solved by task reduction, the agent naturally obtains a self-generated solution trajectory to imitate. By con tinuously collecting and imitating such demonstrations, the agent is able to pro gressively expand the solved subspace in the entire task space. Experiment resul ts show that SIR can significantly accelerate and improve learning on a variety of challenging sparse-reward continuous-control problems with compositional stru ctures. Code and videos are available at https://sites.google.com/view/sir-compo sitional.
****

SOAR: Second-Order Adversarial Regularization
Avery Ma,Fartash Faghri,Nicolas Papernot,Amir-massoud Farahmand
Adversarial training is a common approach to improving the robustness of deep ne ural networks against adversarial examples. In this work, we propose a novel reg ularization approach as an alternative. To derive the regularizer, we formulate the adversarial robustness problem under the robust optimization framework and a pproximate the loss function using a second-order Taylor series expansion. Our p roposed second-order adversarial regularizer (SOAR) is an upper bound based on t he Taylor approximation of the inner-max in the robust optimization objective. W e empirically show that the proposed method improves the robustness of networks against the ℓ_∞ and ℓ_2 bounded perturbations on CIFAR-10 and SVH N.
****

Learning Chess Blindfolded
Shubham Toshniwal,Sam Wiseman,Karen Livescu,Kevin Gimpel
Transformer language models have made tremendous strides in natural language understanding. However, the complexity of natural language makes it challenging to ascertain how accurately these models are tracking the world state underlying the text. Motivated by this issue, we consider the task of language modeling for the game of chess. Unlike natural language, chess notations describe a simple, constrained, and deterministic domain. Moreover, we observe that chess notation itself allows for directly probing the world state, without requiring any additional probing-related machinery. Additionally, we have access to a vast number of chess games coupled with the exact state at every move, allowing us to measure the impact of various ways of including grounding during language model training. Overall, we find that with enough training data, transformer language models can learn to track pieces and predict legal moves when trained solely from move sequences. However, in adverse circumstances (small training sets or prediction following long move histories), providing access to board state information during training can yield consistent improvements.
****
The Intrinsic Dimension of Images and Its Impact on Learning
Phil Pope,Chen Zhu,Ahmed Abdelkader,Micah Goldblum,Tom Goldstein
It is widely believed that natural image data exhibits low-dimensional structure despite the high dimensionality of conventional pixel representations. This idea underlies a common intuition for the remarkable success of deep learning in computer vision. In this work, we apply dimension estimation tools to popular datasets and investigate the role of low-dimensional structure in deep learning. We find that common natural image datasets indeed have very low intrinsic dimension relative to the high number of pixels in the images. Additionally, we find that low dimensional datasets are easier for neural networks to learn, and models solving these tasks generalize better from training to test data. Along the way, we develop a technique for validating our dimension estimation tools on synthetic data generated by GANs allowing us to actively manipulate the intrinsic dimension by controlling the image generation process. Code for our experiments may be found \href{https://github.com/ppope/dimensions}{here}.
****
PriorityCut: Occlusion-aware Regularization for Image Animation
Wai Ting Cheung,Gyeongsu Chae
Image animation generates a video of a source image following the motion of a driving video. Self-supervised image animation approaches do not require explicit pose references as inputs, thus offering large flexibility in learning. State-of-the-art self-supervised image animation approaches mostly warp the source image according to the motion of the driving video, and recover the warping artifacts by inpainting. When the source and the driving images have large pose differences, heavy inpainting is necessary. Without guidance, heavily inpainted regions usually suffer from loss of details. While previous data augmentation techniques such as CutMix are effective in regularizing non-warp-based image generation, directly applying them to image animation ignores the difficulty of inpainting on the warped image. We propose PriorityCut, a novel augmentation approach that uses the top-k percent occluded pixels of the foreground to regularize image animation. By taking into account the difficulty of inpainting, PriorityCut preserves better identity than vanilla CutMix and outperforms state-of-the-art image animation models in terms of the pixel-wise difference, low-level similarity, keypoint distance, and feature embedding distance.
****
Decoupling Exploration and Exploitation for Meta-Reinforcement Learning without Sacrifices
Evan Zheran Liu,Aditi Raghunathan,Percy Liang,Chelsea Finn
The goal of meta-reinforcement learning (meta-RL) is to build agents that can quickly learn new tasks by leveraging prior experience on related tasks. Learning a new task often requires both exploring to gather task-relevant information and exploiting this information to solve the task. In principle, optimal exploratio

n and exploitation can be learned end-to-end by simply maximizing task performan ce. However, such meta-RL approaches struggle with local optima due to a chicken -and-egg problem: learning to explore requires good exploitation to gauge the ex ploration's utility, but learning to exploit requires information gathered via e xploration. Optimizing separate objectives for exploration and exploitation can avoid this problem, but prior meta-RL exploration objectives yield suboptimal po licies that gather information irrelevant to the task. We alleviate both concern s by constructing an exploitation objective that automatically identifies task-r elevant information and an exploration objective to recover only this informatio n. This avoids local optima in end-to-end training, without sacrificing optimal exploration. Empirically, DREAM substantially outperforms existing approaches on complex meta-RL problems, such as sparse-reward 3D visual navigation.
****

Hard Attention Control By Mutual Information Maximization
Himanshu Sahni,Charles Lee Isbell
Biological agents have adopted the principle of attention to limit the rate of i ncoming information from the environment. One question that arises is if an arti ficial agent has access to only a limited view of its surroundings, how can it c ontrol its attention to effectively solve tasks? We propose an approach for lear ning how to control a hard attention window by maximizing the mutual information between the environment state and the attention location at each step. The agen t employs an internal world model to make predictions about its state and focuse s attention towards where the predictions may be wrong. Attention is trained joi ntly with a dynamic memory architecture that stores partial observations and kee ps track of the unobserved state. We demonstrate that our approach is effective in predicting the full state from a sequence of partial observations. We also sh ow that the agent's internal representation of the surroundings, a live mental m ap, can be used for control in two partially observable reinforcement learning t asks. Videos of the trained agent can be found at ~\url{https://sites.google.com /view/hard-attention-control}.
****

Conditional Generative Modeling via Learning the Latent Space
Sameera Ramasinghe,Kanchana Nisal Ranasinghe,Salman Khan,Nick Barnes,Stephen Gou ld
Although deep learning has achieved appealing results on several machine learnin g tasks, most of the models are deterministic at inference, limiting their appli cation to single-modal settings. We propose a novel general-purpose framework fo r conditional generation in multimodal spaces, that uses latent variables to mod el generalizable learning patterns while minimizing a family of regression cost functions. At inference, the latent variables are optimized to find solutions co rresponding to multiple output modes. Compared to existing generative solutions , our approach demonstrates faster and more stable convergence, and can learn be tter representations for downstream tasks. Importantly, it provides a simple gen eric model that can perform better than highly engineered pipelines tailored usi ng domain expertise on a variety of tasks, while generating diverse outputs. Cod e available at https://github.com/samgregoost/cGML.
****

Robust Imitation via Decision-Time Planning
Carl Qi,Pieter Abbeel,Aditya Grover
The goal of imitation learning is to mimic expert behavior from demonstrations, without access to an explicit reward signal. A popular class of approaches infer s the (unknown) reward function via inverse reinforcement learning (IRL) followe d by maximizing this reward function via reinforcement learning (RL). The polici es learned via these approaches are however very brittle in practice and deterio rate quickly even with small test-time perturbations due to compounding errors. We propose Imitation with Planning at Test-time (IMPLANT), a new algorithm for i mitation learning that utilizes decision-time planning to correct for compoundin g errors of any base imitation policy. In contrast to existing approaches, we re tain both the imitation policy and the rewards model at decision-time, thereby b enefiting from the learning signal of the two components. Empirically, we demons

trate that IMPLANT significantly outperforms benchmark imitation learning approa
ches on standard control environments and excels at zero-shot generalization whe
n subject to challenging perturbations in test-time dynamics.
****

Distributionally Robust Learning for Unsupervised Domain Adaptation
Haoxuan Wang,Anqi Liu,Zhiding Yu,Yisong Yue,Anima Anandkumar
We propose a distributionally robust learning (DRL) method for unsupervised dom
ain adaptation (UDA) that scales to modern computer-vision benchmarks. DRL can
be naturally formulated as a competitive two-player game between a predictor an
d an adversary that is allowed to corrupt the labels, subject to certain const
raints, and reduces to incorporating a density ratio between the source and tar
get domains (under the standard log loss). This formulation motivates the use o
f two neural networks that are jointly trained --- a discriminative network betw
een the source and target domains for density-ratio estimation, in addition to
the standard classification network. The use of a density ratio in DRL prevents
the model from being overconfident on target inputs far away from the source dom
ain. Thus, DRL provides conservative confidence estimation in the target doma
in, even when the target labels are not available. This conservatism motivates t
he use of DRL in self-training for sample selection, and we term the approach d
istributionally robust self-training (DRST). In our experiments, DRST generates
more calibrated probabilities and achieves state-of-the-art self-training accur
acy on benchmark datasets. We demonstrate that DRST captures shape features mor
e effectively, and reduces the extent of distributional shift during self-traini
ng.
****

Learning with Feature-Dependent Label Noise: A Progressive Approach
Yikai Zhang,Songzhu Zheng,Pengxiang Wu,Mayank Goswami,Chao Chen
Label noise is frequently observed in real-world large-scale datasets. The noise
is introduced due to a variety of reasons; it is heterogeneous and feature-depe
ndent. Most existing approaches to handling noisy labels fall into two categorie
s: they either assume an ideal feature-independent noise, or remain heuristic wi
thout theoretical guarantees. In this paper, we propose to target a new family o
f feature-dependent label noise, which is much more general than commonly used i
.i.d. label noise and encompasses a broad spectrum of noise patterns. Focusing o
n this general noise family, we propose a progressive label correction algorithm
that iteratively corrects labels and refines the model. We provide theoretical
guarantees showing that for a wide variety of (unknown) noise patterns, a classi
fier trained with this strategy converges to be consistent with the Bayes classi
fier. In experiments, our method outperforms SOTA baselines and is robust to var
ious noise types and levels.
****

Privacy-preserving Learning via Deep Net Pruning
YANGSIBO HUANG,Xiaoxiao Li,Yushan Su,Sachin Ravi,Zhao Song,Sanjeev Arora,Kai Li
Neural network pruning has demonstrated its success in significantly improving t
he computational efficiency of deep models while only introducing a small reduct
ion on final accuracy. In this paper, we explore an extra bonus of neural networ
k pruning in terms of enhancing privacy. Specifically, we show a novel connectio
n between magnitude-based pruning and adding differentially private noise to int
ermediate layers under the over-parameterized regime. To the best of our knowled
ge, this is the first work that bridges pruning with the theory of differential
privacy. The paper also presents experimental results by running the model inver
sion attack on two benchmark datasets, which supports the theoretical finding.
****

ScheduleNet: Learn to Solve MinMax mTSP Using Reinforcement Learning with Delaye
d Reward
Junyoung Park,Sanzhar Bakhtiyarov,Jinkyoo Park
Combinatorial Optimization (CO) problems are theoretically challenging yet cruci
al in practice. Numerous works used Reinforcement Learning (RL) to tackle these
CO problems. As current approaches mainly focus on single-worker CO problems suc
h as the famous Travelling Salesman Problem (TSP), we focus on more practical ex

tension of TSP to multi-worker (salesmen) setting, specifically MinMax mTSP. From the RL perspective, Minmax mTSP raises several significant challenges, such as the cooperation of multiple workers and the need for a well-engineered reward function. In this paper, we present the RL framework with (1) worker-task heterograph and type-aware Graph Neural Network, and (2) the RL training method that is stable, has fast convergence speed, and directly optimizes the objective of MinMax mTSP in a delayed reward setting. We achieve comparable performance to a highly optimized meta-heuristic baseline, OR-Tools, and outperforms it in 10% of the cases, both on in-training and out-of-training problem distributions. Moreover, our problem formulation enables us to solve problems with any number of salesmen (workers) and cities.
****

DialoGraph: Incorporating Interpretable Strategy-Graph Networks into Negotiation Dialogues

Rishabh Joshi,Vidhisha Balachandran,Shikhar Vashishth,Alan Black,Yulia Tsvetkov

To successfully negotiate a deal, it is not enough to communicate fluently: pragmatic planning of persuasive negotiation strategies is essential. While modern dialogue agents excel at generating fluent sentences, they still lack pragmatic grounding and cannot reason strategically. We present DialoGraph, a negotiation system that incorporates pragmatic strategies in a negotiation dialogue using graph neural networks. DialoGraph explicitly incorporates dependencies between sequences of strategies to enable improved and interpretable prediction of next optimal strategies, given the dialogue context. Our graph-based method outperforms prior state-of-the-art negotiation models both in the accuracy of strategy/dialogue act prediction and in the quality of downstream dialogue response generation. We qualitatively show further benefits of learned strategy-graphs in providing explicit associations between effective negotiation strategies over the course of the dialogue, leading to interpretable and strategic dialogues.
****

WaNet - Imperceptible Warping-based Backdoor Attack

Tuan Anh Nguyen,Anh Tuan Tran

With the thriving of deep learning and the widespread practice of using pre-trained networks, backdoor attacks have become an increasing security threat drawing many research interests in recent years. A third-party model can be poisoned in training to work well in normal conditions but behave maliciously when a trigger pattern appears. However, the existing backdoor attacks are all built on noise perturbation triggers, making them noticeable to humans. In this paper, we instead propose using warping-based triggers. The proposed backdoor outperforms the previous methods in a human inspection test by a wide margin, proving its stealthiness. To make such models undetectable by machine defenders, we propose a novel training mode, called the ``noise mode. The trained networks successfully attack and bypass the state-ofthe art defense methods on standard classification datasets, including MNIST, CIFAR-10, GTSRB, and CelebA. Behavior analyses show that our backdoors are transparent to network inspection, further proving this novel attack mechanism's efficiency.
****

Nonseparable Symplectic Neural Networks

Shiying Xiong,Yunjin Tong,Xingzhe He,Shuqi Yang,Cheng Yang,Bo Zhu

Predicting the behaviors of Hamiltonian systems has been drawing increasing attention in scientific machine learning. However, the vast majority of the literature was focused on predicting separable Hamiltonian systems with their kinematic and potential energy terms being explicitly decoupled, while building data-driven paradigms to predict nonseparable Hamiltonian systems that are ubiquitous in fluid dynamics and quantum mechanics were rarely explored. The main computational challenge lies in the effective embedding of symplectic priors to describe the inherently coupled evolution of position and momentum, which typically exhibits intricate dynamics. To solve the problem, we propose a novel neural network architecture, Nonseparable Symplectic Neural Networks (NSSNNs), to uncover and embed the symplectic structure of a nonseparable Hamiltonian system from limited observation data. The enabling mechanics of our approach is an augmented symplectic

time integrator to decouple the position and momentum energy terms and facilitat
e their evolution. We demonstrated the efficacy and versatility of our method by
 predicting a wide range of Hamiltonian systems, both separable and nonseparable
, including chaotic vortical flows. We showed the unique computational merits of
 our approach to yield long-term, accurate, and robust predictions for large-sca
le Hamiltonian systems by rigorously enforcing symplectomorphism.
****

Language Models are Open Knowledge Graphs
Chenguang Wang,Xiao Liu,Dawn Song
This paper shows how to construct knowledge graphs (KGs) from pre-trained langua
ge models (e.g., BERT, GPT-2/3), without human supervision. Popular KGs (e.g, Wi
kidata, NELL) are built in either a supervised or semi-supervised manner, requir
ing humans to create knowledge. Recent deep language models automatically acquir
e knowledge from large-scale corpora via pre-training. The stored knowledge has
enabled the language models to improve downstream NLP tasks, e.g., answering que
stions, and writing code and articles. In this paper, we propose an unsupervised
 method to cast the knowledge contained within language models into KGs. We show
 that KGs are constructed with a single forward pass of the pre-trained language
 models (without fine-tuning) over the corpora. We demonstrate the quality of th
e constructed KGs by comparing to two KGs (Wikidata, TAC KBP) created by humans.
 Our KGs also provide open factual knowledge that is new in the existing KGs. Ou
r code and KGs will be made publicly available.
****

Discovering Diverse Multi-Agent Strategic Behavior via Reward Randomization
Zhenggang Tang,Chao Yu,Boyuan Chen,Huazhe Xu,Xiaolong Wang,Fei Fang,Simon Shaole
i Du,Yu Wang,Yi Wu
We propose a simple, general and effective technique, Reward Randomization for d
iscovering diverse strategic policies in complex multi-agent games. Combining re
ward randomization and policy gradient, we derive a new algorithm, Reward-Random
ized Policy Gradient (RPG). RPG is able to discover a set of multiple distinctiv
e human-interpretable strategies in challenging temporal trust dilemmas, includi
ng grid-world games and a real-world game Agar.io, where multiple equilibria exi
st but standard multi-agent policy gradient algorithms always converge to a fixe
d one with a sub-optimal payoff for every player even using state-of-the-art exp
loration techniques. Furthermore, with the set of diverse strategies from RPG, w
e can (1) achieve higher payoffs by fine-tuning the best policy from the set; an
d (2) obtain an adaptive agent by using this set of strategies as its training o
pponents.
****

Partial Rejection Control for Robust Variational Inference in Sequential Latent
Variable Models
Rahul Sharma,Soumya Banerjee,Dootika Vats,Piyush Rai
Effective variational inference crucially depends on a flexible variational fami
ly of distributions. Recent work has explored sequential Monte-Carlo (SMC) metho
ds to construct variational distributions, which can, in principle, approximate
the target posterior arbitrarily well, which is especially appealing for models
with inherent sequential structure. However, SMC, which represents the posterior
 using a weighted set of particles, often suffers from particle weight degenerac
y, leading to a large variance of the resulting estimators. To address this issu
e, we present a novel approach that leverages the idea of \emph{partial} rejecti
on control (PRC) for developing a robust variational inference (VI) framework. I
n addition to developing a superior VI bound, we propose a novel marginal likeli
hood estimator constructed via a dice-enterprise: a generalization of the Bernou
lli factory to construct unbiased estimators for SMC-PRC. The resulting variatio
nal lower bound can be optimized efficiently with respect to the variational par
ameters and generalizes several existing approaches in the VI literature into a
single framework. We show theoretical properties of the lower bound and report e
xperiments on various sequential models, such as the Gaussian state-space model
and variational RNN, on which our approach outperforms existing methods.
****

Improving Self-supervised Pre-training via a Fully-Explored Masked Language Model
Mingzhi Zheng,Dinghan Shen,yelong shen,Weizhu Chen,Lin Xiao
Masked Language Model (MLM) framework has been widely adopted for self-supervised language pre-training. In this paper, we argue that randomly sampled masks in MLM would lead to undesirably large gradient variance. Thus, we theoretically quantify the gradient variance via correlating the gradient covariance with the Hamming distance between two different masks (given a certain text sequence). To reduce the variance due to the sampling of masks, we propose a fully-explored masking strategy, where a text sequence is divided into a certain number of non-overlapping segments. Thereafter, the tokens within one segment are masked for training. We prove, from a theoretical perspective, that the gradients derived from this new masking schema have a smaller variance and can lead to more efficient self-supervised training. We conduct extensive experiments on both continual pre-training and general pre-training from scratch. Empirical results confirm that this new masking strategy can consistently outperform standard random masking. Detailed efficiency analysis and ablation studies further validate the advantages of our fully-explored masking strategy under the MLM framework.
****
Multi-timescale Representation Learning in LSTM Language Models
Shivangi Mahto,Vy Ai Vo,Javier S. Turek,Alexander Huth
Language models must capture statistical dependencies between words at timescales ranging from very short to very long. Earlier work has demonstrated that dependencies in natural language tend to decay with distance between words according to a power law. However, it is unclear how this knowledge can be used for analyzing or designing neural network language models. In this work, we derived a theory for how the memory gating mechanism in long short-term memory (LSTM) language models can capture power law decay. We found that unit timescales within an LSTM, which are determined by the forget gate bias, should follow an Inverse Gamma distribution. Experiments then showed that LSTM language models trained on natural English text learn to approximate this theoretical distribution. Further, we found that explicitly imposing the theoretical distribution upon the model during training yielded better language model perplexity overall, with particular improvements for predicting low-frequency (rare) words. Moreover, the explicit multi-timescale model selectively routes information about different types of words through units with different timescales, potentially improving model interpretability. These results demonstrate the importance of careful, theoretically-motivated analysis of memory and timescale in language models.
****
Explaining the Efficacy of Counterfactually Augmented Data
Divyansh Kaushik,Amrith Setlur,Eduard H Hovy,Zachary Chase Lipton
In attempts to produce machine learning models less reliant on spurious patterns in NLP datasets, researchers have recently proposed curating counterfactually augmented data (CAD) via a human-in-the-loop process in which given some documents and their (initial) labels, humans must revise the text to make a counterfactual label applicable. Importantly, edits that are not necessary to flip the applicable label are prohibited. Models trained on the augmented (original and revised) data appear, empirically, to rely less on semantically irrelevant words and to generalize better out of domain. While this work draws loosely on causal thinking, the underlying causal model (even at an abstract level) and the principles underlying the observed out-of-domain improvements remain unclear. In this paper, we introduce a toy analog based on linear Gaussian models, observing interesting relationships between causal models, measurement noise, out-of-domain generalization, and reliance on spurious signals. Our analysis provides some insights that help to explain the efficacy of CAD. Moreover, we develop the hypothesis that while adding noise to causal features should degrade both in-domain and out-of-domain performance, adding noise to non-causal features should lead to relative improvements in out-of-domain performance. This idea inspires a speculative test for determining whether a feature attribution technique has identified the causal spans. If adding noise (e.g., by random word flips) to the highlighted spans

degrades both in-domain and out-of-domain performance on a battery of challenge datasets, but adding noise to the complement gives improvements out-of-domain, this suggests we have identified causal spans. Thus, we present a large scale empirical study comparing spans edited to create CAD to those selected by attention and saliency maps. Across numerous challenge domains and models, we find that the hypothesized phenomenon is pronounced for CAD.
****
GraphCGAN: Convolutional Graph Neural Network with Generative Adversarial Networks
Sheng Zhang,Rui Song,Wenbin Lu
Graph convolutional networks (GCN) achieved superior performances in graph-based semi-supervised learning (SSL) tasks.
Generative adversarial networks (GAN) also show the ability to increase the performance in SSL.
However, there is still no good way to combine the GAN and GCN in graph-based SSL tasks.
In this work, we present GraphCGAN, a novel framework to incorporate adversarial learning with convolution-based graph neural networks, to operate on graph-structured data.
In GraphCGAN, we show that generator can generate topology structure and features of fake nodes jointly and boost the performance of convolution-based graph neural networks classifier.
In a number of experiments on benchmark datasets, we show that the proposed GraphCGAN outperforms the baseline methods by a significant margin.
****
ACT: Asymptotic Conditional Transport
Huangjie Zheng,Mingyuan Zhou
We propose conditional transport (CT) as a new divergence to measure the difference between two probability distributions. The CT divergence consists of the expected cost of a forward CT, which constructs a navigator to stochastically transport a data point of one distribution to the other distribution, and that of a backward CT which reverses the transport direction. To apply it to the distributions whose probability density functions are unknown but random samples are accessible, we further introduce asymptotic CT (ACT), whose estimation only requires access to mini-batch based discrete empirical distributions. Equipped with two navigators that amortize the computation of conditional transport plans, the ACT divergence comes with unbiased sample gradients that are straightforward to compute, making it amenable to mini-batch stochastic gradient descent based optimization. When applied to train a generative model, the ACT divergence is shown to strike a good balance between mode covering and seeking behaviors and strongly resist mode collapse. To model high-dimensional data, we show that it is sufficient to modify the adversarial game of an existing generative adversarial network (GAN) to a game played by a generator, a forward navigator, and a backward navigator, which try to minimize a distribution-to-distribution transport cost by optimizing both the distribution of the generator and conditional transport plans specified by the navigators, versus a critic that does the opposite by inflating the point-to-point transport cost. On a wide variety of benchmark datasets for generative modeling, substituting the default statistical distance of an existing GAN with the ACT divergence is shown to consistently improve the performance.
****
Outlier Preserving Distribution Mapping Autoencoders
Walter Gerych,Elke Rundensteiner,Emmanuel Agu
State-of-the-art deep outlier detection methods map data into a latent space with the aim of having outliers far away from inliers in this space. Unfortunately, this often fails as the divergence penalty they adopt pushes outliers into the same high-probability regions as inliers. We propose a novel method, OP-DMA, that successfully addresses the above problem. OP-DMA succeeds in mapping outliers to low probability regions in the latent space by leveraging a novel Prior-Weighted Loss (PWL) that utilizes the insight that outliers are likely to have a higher reconstruction error than inliers. Building on this insight, OP-DMA weig

hts the reconstruction error of individual points by a multivariate Gaussian pro bability density function evaluated at each point's latent representation. We de monstrate and provide theoretical proof that this succeeds to map outliers to lo w-probability regions. Our experimental study shows that OP-DMA consistently ou tperforms state-of-art methods on a rich variety of outlier detection benchmar k datasets.
****
Motif-Driven Contrastive Learning of Graph Representations
Shichang Zhang,Ziniu Hu,Arjun Subramonian,Yizhou Sun
Graph motifs are significant subgraph patterns occurring frequently in graphs, a nd they play important roles in representing the whole graph characteristics. Fo r example, in the chemical domain, functional groups are motifs that can determi ne molecule properties. Mining and utilizing motifs, however, is a non-trivial t ask for large graph datasets. Traditional motif discovery approaches mostly rely on exact counting or statistical estimation, which are hard to scale for a larg e number of graphs with continuous and high-dimension features. In light of the significance and challenges of motif mining, we propose : MICRO-Graph: a framewo rk for \underline{M}ot\underline{I}f-driven \underline{C}ontrastive lea\underlin e{R}ning \underline{O}f \underline{G}raph representations to: 1) pre-train Graph Neural Networks (GNNs) in a self-supervised manner to automatically extract gra ph motifs from large graph datasets; 2) leverage learned motifs to guide the con trastive learning of graph representations, which further benefit various graph downstream tasks. Specifically, given a graph dataset, a motif learner cluster s imilar and significant subgraphs into corresponding motif slots. Based on the le arned motifs, a motif-guided subgraph segmenter is trained to generate more info rmative subgraphs, which are used to conduct graph-to-subgraph contrastive learn ing of GNNs. Our discovering strategy is to simutaneously do clustering and cont rastive learning on dynamically sampled subgraphs. The clustering part pull toge ther similar subgraphs across different whole graphs, as the contrastive part pu sh away dissimilar ones. Meanwhile, our learnable sampler will generate subgraph samples better aligned with the discoverying procedure. By pre-training on ogbn -molhiv molecule dataset with our proposed MICRO-Graph, the pre-trained GNN mode l can enhance various chemical property prediction downstream tasks with scarce label by 2.0%, and significantly higher than other state-of-the-art self-supervi sed learning baselines.

****
Revisiting Locally Supervised Learning: an Alternative to End-to-end Training
Yulin Wang,Zanlin Ni,Shiji Song,Le Yang,Gao Huang
Due to the need to store the intermediate activations for back-propagation, end- to-end (E2E) training of deep networks usually suffers from high GPUs memory foo tprint. This paper aims to address this problem by revisiting the locally superv ised learning, where a network is split into gradient-isolated modules and train ed with local supervision. We experimentally show that simply training local mod ules with E2E loss tends to collapse task-relevant information at early layers, and hence hurts the performance of the full model. To avoid this issue, we propo se an information propagation (InfoPro) loss, which encourages local modules to preserve as much useful information as possible, while progressively discard tas k-irrelevant information. As InfoPro loss is difficult to compute in its origina l form, we derive a feasible upper bound as a surrogate optimization objective, yielding a simple but effective algorithm. In fact, we show that the proposed me thod boils down to minimizing the combination of a reconstruction loss and a nor mal cross-entropy/contrastive term. Extensive empirical results on five datasets (i.e., CIFAR, SVHN, STL-10, ImageNet and Cityscapes) validate that InfoPro is c apable of achieving competitive performance with less than 40% memory footprint compared to E2E training, while allowing using training data with higher-resolut ion or larger batch sizes under the same GPU memory constraint. Our method also enables training local modules asynchronously for potential training acceleratio n.
****

Improving the Reconstruction of Disentangled Representation Learners via Multi-Stage Modelling
Akash Srivastava,Yamini Bansal,Yukun Ding,Cole Lincoln Hurwitz,Kai Xu,Bernhard Egger,Prasanna Sattigeri,Joshua B. Tenenbaum,Dan Gutfreund
Current autoencoder-based disentangled representation learning methods achieve disentanglement by penalizing the (aggregate) posterior to encourage statistical independence of the latent factors. This approach introduces a trade-off between disentangled representation learning and reconstruction quality since the model does not have enough capacity to learn correlated latent variables that capture detail information present in most image data. To overcome this trade-off, we present a novel multi-stage modelling approach where the disentangled factors are first learned using a preexisting disentangled representation learning method (such as β-TCVAE); then, the low-quality reconstruction is improved with another deep generative model that is trained to model the missing correlated latent variables, adding detail information while maintaining conditioning on the previously learned disentangled factors. Taken together, our multi-stage modelling approach results in single, coherent probabilistic model that is theoretically justified by the principal of D-separation and can be realized with a variety of model classes including likelihood-based models such as variational autoencoders, implicit models such as generative adversarial networks, and tractable models like normalizing flows or mixtures of Gaussians. We demonstrate that our multi-stage model has much higher reconstruction quality than current state-of-the-art methods with equivalent disentanglement performance across multiple standard benchmarks.
****
How Much Over-parameterization Is Sufficient to Learn Deep ReLU Networks?
Zixiang Chen,Yuan Cao,Difan Zou,Quanquan Gu
A recent line of research on deep learning focuses on the extremely over-parameterized setting, and shows that when the network width is larger than a high degree polynomial of the training sample size n and the inverse of the target error ϵ^{-1}, deep neural networks learned by (stochastic) gradient descent enjoy nice optimization and generalization guarantees. Very recently, it is shown that under certain margin assumptions on the training data, a polylogarithmic width condition suffices for two-layer ReLU networks to converge and generalize (Ji and Telgarsky, 2020). However, whether deep neural networks can be learned with such a mild over-parameterization is still an open question. In this work, we answer this question affirmatively and establish sharper learning guarantees for deep ReLU networks trained by (stochastic) gradient descent. In specific, under certain assumptions made in previous work, our optimization and generalization guarantees hold with network width polylogarithmic in n and ϵ^{-1}. Our results push the study of over-parameterized deep neural networks towards more practical settings.
****
SGD on Neural Networks learns Robust Features before Non-Robust
Vikram Nitin
Neural networks are known to be vulnerable to adversarial attacks - small, imperceptible perturbations that cause the network to misclassify an input. A recent line of work attempts to explain this behavior by positing the existence of non-robust features - well-generalizing but brittle features present in the data distribution that are learned by the network and can be perturbed to cause misclassification.

In this paper, we look at the dynamics of neural network training through the perspective of robust and non-robust features. We find that there are two very distinct pathways that neural network training can follow, depending on the hyperparameters used. In the first pathway, the network initially learns only predictive, robust features and weakly predictive non-robust features, and subsequently learns predictive, non-robust features. On the other hand, a network trained via the second pathway eschews predictive non-robust features altogether, and rapidly overfits the training data. We provide strong empirical evidence to corroborat

e this hypothesis, as well as theoretical analysis in a simplified setting. Key to our analysis is a better understanding of the relationship between predictive non-robust features and adversarial transferability. We present our findings in light of other recent results on the evolution of inductive biases learned by neural networks over the course of training.

Finally, we digress to show that rather than being quirks of the data distribution, predictive non-robust features might actually occur across datasets with different distributions drawn from independent sources, indicating that they perhaps possess some meaning in terms of human semantics.
****
Data augmentation for deep learning based accelerated MRI reconstruction
Zalan Fabian,Reinhard Heckel,Mahdi Soltanolkotabi
Deep neural networks have emerged as very successful tools for image restoration and reconstruction tasks. These networks are often trained end-to-end to directly reconstruct an image from a noisy or corrupted measurement of that image. To achieve state-of-the-art performance, training on large and diverse sets of images is considered critical. However, it is often difficult and/or expensive to collect large amounts of training images. Inspired by the success of Data Augmentation (DA) for classification problems, in this paper, we propose a pipeline for data augmentation for image reconstruction tasks arising in medical imaging and explore its effectiveness at reducing the required training data in a variety of settings. We focus on accelerated magnetic resonance imaging, where the goal is to reconstruct an image from a few under-sampled linear measurements. Our DA pipeline is specifically designed to utilize the invariances present in medical imaging measurements as naive DA strategies that neglect the physics of the problem fail. We demonstrate the effectiveness of our data augmentation pipeline by showing that for some problem regimes, DA can achieve comparable performance to the state of the art on the FastMRI dataset while using significantly fewer training data. Specifically, for 8-fold acceleration we achieve performance comparable to the state of the art with only 10% of the training data for multi-coil reconstruction and with only 33% of the training data for single-coil reconstruction. Our findings show that in the low-data regime DA is beneficial, whereas in the high-data regime it has diminishing returns.
****
Play to Grade: Grading Interactive Coding Games as Classifying Markov Decision Process
Allen Nie,Emma Brunskill,Chris Piech
Contemporary coding education often present students with the task of developing programs that have user interaction and complex dynamic systems, such as mouse based games. While pedagogically compelling, grading such student programs requires dynamic user inputs, therefore they are difficult to grade by unit tests. In this paper we formalize the challenge of grading interactive programs as a task of classifying Markov Decision Processes (MDPs). Each student's program fully specifies an MDP where the agent needs to operate and decide, under reasonable generalization, if the dynamics and reward model of the input MDP conforms to a set of latent MDPs. We demonstrate that by experiencing a handful of latent MDPs millions of times, we can use the agent to sample trajectories from the input MDP and use a classifier to determine membership. Our method drastically reduces the amount of data needed to train an automatic grading system for interactive code assignments and present a challenge to state-of-the-art reinforcement learning generalization methods. Together with Code.org, we curated a dataset of 700k student submissions, one of the largest dataset of anonymized student submissions to a single assignment. This Code.org assignment had no previous solution for automatically providing correctness feedback to students and as such this contribution could lead to meaningful improvement in educational experience.
****
Provably More Efficient Q-Learning in the One-Sided-Feedback/Full-Feedback Settings
Xiao-Yue Gong,David Simchi-Levi

Motivated by the episodic version of the classical inventory control problem, we propose a new Q-learning-based algorithm, Elimination-Based Half-Q-Learning (HQL), that enjoys improved efficiency over existing algorithms for a wide variety of problems in the one-sided-feedback setting. We also provide a simpler variant of the algorithm, Full-Q-Learning (FQL), for the full-feedback setting. We establish that HQL incurs $ \tilde{\mathcal{O}}(H^3\sqrt{ T})$ regret and FQL incurs $\tilde{\mathcal{O}}(H^2\sqrt{ T})$ regret, where H is the length of each episode and T is the total length of the horizon. The regret bounds are not affected by the possibly huge state and action space. Our numerical experiments demonstrate the superior efficiency of HQL and FQL, and the potential to combine reinforcement learning with richer feedback models.
****
Blending MPC & Value Function Approximation for Efficient Reinforcement Learning
Mohak Bhardwaj,Sanjiban Choudhury,Byron Boots
Model-Predictive Control (MPC) is a powerful tool for controlling complex, real-world systems that uses a model to make predictions about future behavior. For each state encountered, MPC solves an online optimization problem to choose a control action that will minimize future cost. This is a surprisingly effective strategy, but real-time performance requirements warrant the use of simple models. If the model is not sufficiently accurate, then the resulting controller can be biased, limiting performance. We present a framework for improving on MPC with model-free reinforcement learning (RL). The key insight is to view MPC as constructing a series of local Q-function approximations. We show that by using a parameter λ, similar to the trace decay parameter in TD(λ), we can systematically trade-off learned value estimates against the local Q-function approximations. We present a theoretical analysis that shows how error from inaccurate models in MPC and value function estimation in RL can be balanced. We further propose an algorithm that changes λ over time to reduce the dependence on MPC as our estimates of the value function improve, and test the efficacy our approach on challenging high-dimensional manipulation tasks with biased models in simulation. We demonstrate that our approach can obtain performance comparable with MPC with access to true dynamics even under severe model bias and is more sample efficient as compared to model-free RL.
****
A framework for learned CountSketch
Simin Liu,Tianrui Liu,Ali Vakilian,Yulin Wan,David Woodruff
Sketching is a compression technique that can be applied to many problems to solve them quickly and approximately. The matrices used to project data to smaller dimensions are called "sketches". In this work, we consider the problem of optimizing sketches to obtain low approximation error over a data distribution.

We introduce a general framework for "learning" and applying CountSketch, a type of sparse sketch. The sketch optimization procedure has two stages: one for optimizing the placements of the sketch's non-zero entries and another for optimizing their values. Next, we provide a way to apply learned sketches that has worst-case guarantees for approximation error.

We instantiate this framework with three sketching applications: least-squares regression, low-rank approximation (LRA), and k-means clustering. Our experiments demonstrate that our approach substantially decreases approximation error compared to classical and naively learned sketches.

Finally, we investigate the theoretical aspects of our approach. For regression and LRA, we show that our method obtains state-of-the art accuracy for fixed time complexity. For LRA, we prove that it is strictly better to include the first optimization stage for two standard input distributions. For k-means, we derive a more straightforward means of retaining approximation guarantees.
****
Environment Predictive Coding for Embodied Agents
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 images, we aim to jointly encode a series of images gathered by an agent as it moves about in 3D environments. We learn these representations via a zone prediction task, where we intelligently mask out portions of an agent's trajectory and predict them from the unmasked portions, conditioned on the agent's camera poses. By learning such representations on a collection of videos, we demonstrate successful transfer to multiple downstream navigation-oriented tasks. Our experiments on the photorealistic 3D environments of Gibson and Matterport3D show that our method outperforms the state-of-the-art on challenging tasks with only a limited budget of experience.
****

Probabilistic Numeric Convolutional Neural Networks
Marc Anton Finzi,Roberto Bondesan,Max Welling
Continuous input signals like images and time series that are irregularly sampled or have missing values are challenging for existing deep learning methods. Coherently defined feature representations must depend on the values in unobserved regions of the input. Drawing from the work in probabilistic numerics, we propose Probabilistic Numeric Convolutional Neural Networks which represent features as Gaussian processes, providing a probabilistic description of discretization error. We then define a convolutional layer as the evolution of a PDE defined on this GP, followed by a nonlinearity. This approach also naturally admits steerable equivariant convolutions under e.g. the rotation group. In experiments we show that our approach yields a $3\times$ reduction of error from the previous state of the art on the SuperPixel-MNIST dataset and competitive performance on the medical time series dataset PhysioNet2012.
****

Three Dimensional Reconstruction of Botanical Trees with Simulatable Geometry
Ed Quigley,Winnie Lin,Yilin Zhu,Ronald Fedkiw
We tackle the challenging problem of creating full and accurate three dimensional reconstructions of botanical trees with the topological and geometric accuracy required for subsequent physical simulation, e.g. in response to wind forces. Although certain aspects of our approach would benefit from various improvements, our results exceed the state of the art especially in geometric and topological complexity and accuracy. Starting with two dimensional RGB image data acquired from cameras attached to drones, we create point clouds, textured triangle meshes, and a simulatable and skinned cylindrical articulated rigid body model. We discuss the pros and cons of each step of our pipeline, and in order to stimulate future research we make the raw and processed data from every step of the pipeline as well as the final geometric reconstructions publicly available.
****

Systematic Evaluation of Causal Discovery in Visual Model Based Reinforcement Learning
Nan Rosemary Ke,Aniket Rajiv Didolkar,Sarthak Mittal,Anirudh Goyal,Guillaume Lajoie,Stefan Bauer,Danilo Jimenez Rezende,Michael Curtis Mozer,Yoshua Bengio,Christopher Pal
Inducing causal relationships from observations is a classic problem in machine learning. Most work in causality starts from the premise that the causal variables themselves have known semantics or are observed. However, for AI agents such as robots trying to make sense of their environment, the only observables are low-level variables like pixels in images. To generalize well, an agent must induce high-level variables, particularly those which are causal or are affected by causal variables. A central goal for AI and causality is thus the joint discovery of abstract representations and causal structure. In this work, we systematically evaluate the agent's ability to learn underlying causal structure. We note that existing environments for studying causal induction are poorly suited for this objective because they have complicated task-specific causal graphs with many confounding factors. Hence, to facilitate research in learning the representation of high-level variables as well as causal structure among these variables, we present a suite of RL environments created to systematically probe the ability

of methods to identify variables as well as causal structure among those variables. We evaluate various representation learning algorithms from literature and found that explicitly incorporating structure and modularity in the model can help causal induction in model-based reinforcement learning.
****

Dependency Structure Discovery from Interventions
Nan Rosemary Ke,Olexa Bilaniuk,Anirudh Goyal,Stefan Bauer,Bernhard Schölkopf,Michael Curtis Mozer,Hugo Larochelle,Christopher Pal,Yoshua Bengio
Promising results have driven a recent surge of interest in continuous optimization methods for Bayesian network structure learning from observational data. However, there are theoretical limitations on the identifiability of underlying structures obtained from observational data alone. Interventional data provides much richer information about the underlying data-generating process. However, the extension and application of methods designed for observational data to include interventions is not straightforward and remains an open problem. In this paper we provide a general framework based on continuous optimization and neural networks to create models for the combination of observational and interventional data. The proposed method is applicable even in the challenging and realistic case that the identity of the intervened upon variable is unknown. We examine the proposed method in the setting of graph recovery both de novo and from a partially-known edge set. We establish strong benchmark results on several structure learning tasks, including structure recovery of both synthetic graphs as well as standard graphs from the Bayesian Network Repository.
****

HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients
Enmao Diao,Jie Ding,Vahid Tarokh
Federated Learning (FL) is a method of training machine learning models on private data distributed over a large number of possibly heterogeneous clients such as mobile phones and IoT devices. In this work, we propose a new federated learning framework named HeteroFL to address heterogeneous clients equipped with very different computation and communication capabilities. Our solution can enable the training of heterogeneous local models with varying computation complexities and still produce a single global inference model. For the first time, our method challenges the underlying assumption of existing work that local models have to share the same architecture as the global model. We demonstrate several strategies to enhance FL training and conduct extensive empirical evaluations, including five computation complexity levels of three model architecture on three datasets. We show that adaptively distributing subnetworks according to clients' capabilities is both computation and communication efficient.
****

A Policy Gradient Algorithm for Learning to Learn in Multiagent Reinforcement Learning
Dong-Ki Kim,Miao Liu,Matthew D Riemer,Chuangchuang Sun,Marwa Abdulhai,Golnaz Habibi,Sebastian Lopez-Cot,Gerald Tesauro,JONATHAN P HOW
A fundamental challenge in multiagent reinforcement learning is to learn beneficial behaviors in a shared environment with other agents that are also simultaneously learning. In particular, each agent perceives the environment as effectively non-stationary due to the changing policies of other agents. Moreover, each agent is itself constantly learning, leading to natural nonstationarity in the distribution of experiences encountered. In this paper, we propose a novel meta-multiagent policy gradient theorem that directly accommodates for the non-stationary policy dynamics inherent to these multiagent settings. This is achieved by modeling our gradient updates to directly consider both an agent’s own non-stationary policy dynamics and the non-stationary policy dynamics of other agents interacting with it in the environment. We find that our theoretically grounded approach provides a general solution to the multiagent learning problem, which inherently combines key aspects of previous state of the art approaches on this topic. We test our method on several multiagent benchmarks and demonstrate a more efficient ability to adapt to new agents as they learn than previous related approach

es across the spectrum of mixed incentive, competitive, and cooperative environm
ents.
****

Learning Blood Oxygen from Respiration Signals
Hao He,Ying-Cong Chen,Yuan Yuan,Dina Katabi
Monitoring blood oxygen is critical in a variety of medical conditions. For almo
st a century, pulse oximetry has been the only non-invasive method for measuring
 blood oxygen. While highly useful, pulse oximetry has important limitations. It
 requires wearable sensors, which can be cumbersome for older patients. It is al
so known to be biased when used for dark-skinned subjects. In this paper, we dem
onstrate, for the first time, the feasibility of predicting oxygen saturation fr
om breathing. By eliminating the dependency on oximetry, we eliminate bias again
st skin color. Further, since breathing can be monitored without body contact by
 analyzing the radio signal in the environment, we show that oxygen too can be m
onitored without any wearable devices. We introduce a new approach for leveragin
g auxiliary variables via a switcher-based multi-headed neural network model. Em
pirical results show that our model achieves good accuracy on multiple medical d
atasets.
****

MSFM: Multi-Scale Fusion Module for Object Detection
Xuesong Wang,Caisheng Wang
Feature fusion is beneficial to object detection tasks in two folds. On one hand
, detail and position information can be combined with semantic information when
 high and low-resolution features from shallow and deep layers are fused. On the
 other hand, objects can be detected in different scales, which improves the rob
ustness of the framework. In this work, we present a Multi-Scale Fusion Module (
MSFM) that extracts both detail and semantical information from a single input b
ut at different scales within the same layer. Specifically, the input of the mod
ule will be resized into different scales on which position and semantic informa
tion will be processed, and then they will be rescaled back and combined with th
e module input. The MSFM is lightweight and can be used as a drop-in layer to ma
ny existing object detection frameworks. Experiments show that MSFM can bring +2
.5% mAP improvement with only 2.4M extra parameters on Faster R-CNN with ResNet-
50 FPN backbone on COCO Object Detection minival set, outperforming that with Re
sNet-101 FPN backbone without the module which obtains +2.0% mAP with 19.0M extr
a parameters. The best resulting model achieves a 45.7% mAP on test-dev set. Cod
e will be available.
****

Semantic Re-tuning with Contrastive Tension
Fredrik Carlsson,Amaru Cuba Gyllensten,Evangelia Gogoulou,Erik Ylipää Hellqvist,
Magnus Sahlgren
Extracting semantically useful natural language sentence representations from pr
e-trained deep neural networks such as Transformers remains a challenge. We firs
t demonstrate that pre-training objectives impose a significant task bias onto t
he final layers of models with a layer-wise survey of the Semantic Textual Simil
arity (STS) correlations for multiple common Transformer language models. We the
n propose a new self-supervised method called Contrastive Tension (CT) to counte
r such biases. CT frames the training objective as a noise-contrastive task betw
een the final layer representations of two independent models, in turn making th
e final layer representations suitable for feature extraction. Results from mult
iple common unsupervised and supervised STS tasks indicate that CT outperforms p
revious State Of The Art (SOTA), and when combining CT with supervised data we i
mprove upon previous SOTA results with large margins.
****

Hard-label Manifolds: Unexpected advantages of query efficiency for finding on-m
anifold adversarial examples
Washington Garcia,Pin-Yu Chen,Somesh Jha,Hamilton Scott Clouse,Kevin Butler
Designing deep networks robust to adversarial examples remains an open problem.
Likewise, recent zeroth order hard-label attacks on image classification tasks h
ave shown comparable performance to their first-order alternatives. It is well k

nown that in this setting, the adversary must search for the nearest decision bo
undary in a query-efficient manner. State-of-the-art (SotA) attacks rely on the
concept of pixel grouping, or super-pixels, to perform efficient boundary search
. It was recently shown in the first-order setting, that regular adversarial exa
mples leave the data manifold, and on-manifold examples are generalization error
s. In this paper, we argue that query efficiency in the zeroth-order setting is
connected to the adversary's traversal through the data manifold. In particular,
 query-efficient hard-label attacks have the unexpected advantage of finding adv
ersarial examples close to the data manifold. We empirically demonstrate that ag
ainst both natural and robustly trained models, an efficient zeroth-order attack
 produces samples with a progressively smaller manifold distance measure. Furthe
r, when a normal zeroth-order attack is made query-efficient through the use of
pixel grouping, it can make up to a two-fold increase in query efficiency, and i
n some cases, reduce a sample's distance to the manifold by an order of magnitud
e.
****

Topic-aware Contextualized Transformers
Ruiying Lu,Bo Chen,Dan dan Guo,Dongsheng Wang,Mingyuan Zhou
Training on disjoint fixed-length segments, Transformers successfully transform
static word embeddings into contextualized word representations. However, they o
ften restrict the context of a token to the segment it resides in and hence negl
ect the flow of contextual information across segments, failing to capture longe
r-term dependencies beyond the predefined segment length. This paper uses a prob
abilistic deep topic model to provide contextualized embeddings at both the toke
n and segment levels. It also introduces topic self-attention and a contextual n
ext-word embedding guided topic select-attention, injecting contextualized topic
 information into Transformer-based architectures. Moving beyond conventional Tr
ansformers that ignore longer-range word dependencies and contextualize their wo
rd representations at the segment level, the proposed method not only captures g
lobal semantic coherence of all segments and global word concurrence patterns, b
ut also enriches the representation of each token by adapting it to its local co
ntext, which is not limited to the segment it resides in and can be flexibly def
ined according to the task. Experiments on various corpora show that adding only
 a few extra parameters, the proposed topic-aware contextualized transformers co
nsistently outperform their conventional counterparts, and can be used to genera
te coherent sentences and paragraphs.
****

Cluster-Former: Clustering-based Sparse Transformer for Question Answering
Shuohang Wang,Luowei Zhou,Zhe Gan,Yen-Chun Chen,Yuwei Fang,Siqi Sun,Yu Cheng,Jin
gjing Liu
Transformer has become ubiquitous in the deep learning field. One of the key ing
redients that destined its success is the self-attention mechanism, which allows
 fully-connected contextual encoding over input tokens.
However, despite its effectiveness in modeling short sequences, self-attention s
uffers when handling inputs with extreme long-range dependencies, as its complex
ity grows quadratically with respect to the sequence length.
Therefore, long sequences are often encoded by Transformer in chunks using a sli
ding window.
In this paper, we propose Cluster-Former, a novel clustering-based sparse Transf
ormer to perform attention across chunked sequences. The proposed framework is p
ivoted on two unique types of Transformer layer: Sliding-Window Layer and Cluste
r-Former Layer, which encode local sequence information and global context joint
ly and iteratively.
This new design allows information integration beyond local windows, which is es
pecially beneficial for question answering (QA) tasks that rely on long-range de
pendencies. Experiments show that Cluster-Former achieves state-of-the-art perfo
rmance on several major QA benchmarks.
****

Dataset Meta-Learning from Kernel Ridge-Regression
Timothy Nguyen,Zhourong Chen,Jaehoon Lee

One of the most fundamental aspects of any machine learning algorithm is the tra
ining data used by the algorithm.
We introduce the novel concept of ϵ-approximation of datasets, obtainin
g datasets which are much smaller than or are significant corruptions of the ori
ginal training data while maintaining similar performance. We introduce a meta-l
earning algorithm Kernel Inducing Points (KIP) for obtaining such remarkable dat
asets, drawing inspiration from recent developments in the correspondence betwee
n infinitely-wide neural networks and kernel ridge-regression (KRR). For KRR tas
ks, we demonstrate that KIP can compress datasets by one or two orders of magnit
ude, significantly improving previous dataset distillation and subset selection
methods while obtaining state of the art results for MNIST and CIFAR10 classific
ation. Furthermore, our KIP-learned datasets are transferable to the training of
 finite-width neural networks even beyond the lazy-training regime. Consequently
, we obtain state of the art results for neural network dataset distillation wit
h potential applications to privacy-preservation.
****

AUXILIARY TASK UPDATE DECOMPOSITION: THE GOOD, THE BAD AND THE NEUTRAL
Lucio M. Dery,Yann Dauphin,David Grangier
While deep learning has been very beneficial in data-rich settings, tasks with s
maller training set
often resort to pre-training or multitask learning to leverage data from other t
asks. In this case,
careful consideration is needed to select tasks and model parameterizations such
 that updates from
the auxiliary tasks actually help the primary task. We seek to alleviate this bu
rden by formulating a model-agnostic framework that performs fine-grained manipu
lation of the auxiliary task gradients. We propose to decompose auxiliary update
s into directions which help, damage or leave the primary task loss unchanged. T
his allows weighting the update directions
differently depending on their impact on the problem of interest. We present a n
ovel and efficient algorithm for that
purpose and show its advantage in practice. Our method leverages efficient autom
atic differentiation
procedures and randomized singular value decomposition for scalability. We show
that our framework is
generic and encompasses some prior work as particular cases. Our approach consis
tently outperforms strong and widely used baselines when leveraging out-of-distr
ibution data for Text and Image classification tasks.
****

A Large-scale Study on Training Sample Memorization in Generative Modeling
Ching-Yuan Bai,Hsuan-Tien Lin,Colin Raffel,Wendy Kan
Many recent developments on generative models for natural images have relied on
heuristically-motivated metrics that can be easily gamed by memorizing a small s
ample from the true distribution or training a model directly to improve the met
ric.
In this work, we critically evaluate the gameability of the benchmarking procedu
re by running a competition which ultimately resulted in participants attempting
 to cheat. Our competition received over 11000 submitted models which allowed us
 to investigate memorization-aware metrics for measuring generative model perfor
mance. Specifically, we propose the Memorization-Informed Frechet Inception Dist
ance (MiFID) and discuss ways to ensure that winning submissions were based on g
enuine improvements in perceptual quality. We evaluate the effectiveness of our
benchmark by manually inspecting the code for the 1000 top-performing models and
 labeling different forms of memorization that were intentionally or unintention
ally used. To facilitate future work on benchmarking generative models, we relea
se generated images and our labels for these models as well as code to compute t
he MiFID metric.
****

AWAC: Accelerating Online Reinforcement Learning with Offline Datasets
Ashvin Nair,Murtaza Dalal,Abhishek Gupta,Sergey Levine

Reinforcement learning provides an appealing formalism for learning control poli
cies from experience. However, the classic active formulation of reinforcement l
earning necessitates a lengthy active exploration process for each behavior, mak
ing it difficult to apply in real-world settings. If we can instead allow reinfo
rcement learning to effectively use previously collected data to aid the online
learning process, where the data could be expert demonstrations or more generall
y any prior experience, we could make reinforcement learning a substantially mor
e practical tool. While a number of recent methods have sought to learn offline
from previously collected data, it remains exceptionally difficult to train a po
licy with offline data and improve it further with online reinforcement learning
. In this paper we systematically analyze why this problem is so challenging, an
d propose an algorithm that combines sample-efficient dynamic programming with m
aximum likelihood policy updates, providing a simple and effective framework tha
t is able to leverage large amounts of offline data and then quickly perform onl
ine fine-tuning of reinforcement learning policies. We show that our method enab
les rapid learning of skills with a combination of prior demonstration data and
online experience across a suite of difficult dexterous manipulation and benchma
rk tasks.
****
Fast And Slow Learning Of Recurrent Independent Mechanisms
Kanika Madan,Nan Rosemary Ke,Anirudh Goyal,Bernhard Schölkopf,Yoshua Bengio
Decomposing knowledge into interchangeable pieces promises a generalization adva
ntage when there are changes in distribution. A learning agent interacting with
its environment is likely to be faced with situations requiring novel combinatio
ns of existing pieces of knowledge. We hypothesize that such a decomposition of
knowledge is particularly relevant for being able to generalize in a systematic
way to out-of-distribution changes. To study these ideas, we propose a particula
r training framework in which we assume that the pieces of knowledge an agent ne
eds and its reward function are stationary and can be re-used across tasks. An a
ttention mechanism dynamically selects which modules can be adapted to the curre
nt task, and the parameters of the \textit{selected} modules are allowed to chan
ge quickly as the learner is confronted with variations in what it experiences,
while the parameters of the attention mechanisms act as stable, slowly changing,
 meta-parameters. We focus on pieces of knowledge captured by an ensemble of mo
dules sparsely communicating with each other via a bottleneck of attention. We f
ind that meta-learning the modular aspects of the proposed system greatly helps
 in achieving faster adaptation in a reinforcement learning setup involving navi
gation in a partially observed grid world with image-level input. We also find
that reversing the role of parameters and meta-parameters does not work nearly a
s well, suggesting a particular role for fast adaptation of the dynamically sele
cted modules.
****
Detecting Hallucinated Content in Conditional Neural Sequence Generation
Chunting Zhou,Jiatao Gu,Mona T. Diab,Paco Guzmán,Luke Zettlemoyer,Marjan Ghazvin
inejad
Neural sequence models can generate highly fluent sentences but recent studies h
ave also shown that they are also prone to hallucinate additional content not su
pported by the input, which can cause a lack of trust in the model.
To better assess the faithfulness of the machine outputs, we propose a new task
to predict whether each token in the output sequence is hallucinated conditioned
 on the source input, and collect new manually annotated evaluation sets for thi
s task.
We also introduce a novel method for learning to model hallucination detection,
based on pretrained language models fine tuned on synthetic data that includes a
utomatically inserted hallucinations.
Experiments on machine translation and abstract text summarization demonstrate t
he effectiveness of our proposed approach -- we obtain an average F1 of around 0
.6 across all the benchmark datasets.
Furthermore, we demonstrate how to use the token-level hallucination labels to d
efine a fine-grained loss over the target sequence in the low-resource machine t

ranslation and achieve significant improvements over strong baseline methods. We will also release our annotated data and code for future research.
****
Auction Learning as a Two-Player Game
Jad Rahme,Samy Jelassi,S. Matthew Weinberg
Designing an incentive compatible auction that maximizes expected revenue is a central problem in Auction Design. While theoretical approaches to the problem have hit some limits, a recent research direction initiated by Duetting et al. (2019) consists in building neural network architectures to find optimal auctions. We propose two conceptual deviations from their approach which result in enhanced performance. First, we use recent results in theoretical auction design to introduce a time-independent Lagrangian. This not only circumvents the need for an expensive hyper-parameter search (as in prior work), but also provides a single metric to compare the performance of two auctions (absent from prior work). Second, the optimization procedure in previous work uses an inner maximization loop to compute optimal misreports. We amortize this process through the introduction of an additional neural network. We demonstrate the effectiveness of our approach by learning competitive or strictly improved auctions compared to prior work. Both results together further imply a novel formulation of Auction Design as a two-player game with stationary utility functions.
****
Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies
Yae Jee Cho,Jianyu Wang,Gauri Joshi
Federated learning is a distributed optimization paradigm that enables a large number of resource-limited client nodes to cooperatively train a model without data sharing. Several works have analyzed the convergence of federated learning by accounting of data heterogeneity, communication and computation limitations, and partial client participation. However, they assume unbiased client participation, where clients are selected at random or in proportion of their data sizes. In this paper, we present the first convergence analysis of federated optimization for biased client selection strategies, and quantify how the selection bias affects convergence speed. We reveal that biasing client selection towards clients with higher local loss achieves faster error convergence. Using this insight, we propose Power-of-Choice, a communication- and computation-efficient client selection framework that can flexibly span the trade-off between convergence speed and solution bias. We also propose an extension of Power-of-Choice that is able to maintain convergence speed improvement while diminishing the selection skew. Our experiments demonstrate that Power-of-Choice strategies converge up to 3 $\times$ faster and give 10% higher test accuracy than the baseline random selection.
****
Provable Robustness by Geometric Regularization of ReLU Networks
Chester Holtz,Changhao Shi,Gal Mishne
Recent work has demonstrated that neural networks are vulnerable to small, adversarial perturbations of their input. In this paper, we propose an efficient regularization scheme inspired by convex geometry and barrier methods to improve the robustness of feedforward ReLU networks. Since such networks are piecewise linear, they partition the input space into polyhedral regions (polytopes). Our regularizer is designed to minimize the distance between training samples and the \textit{analytical centers} of their respective polytopes so as to push points away from the boundaries. Our regularizer \textit{provably} improves a lower bound on the necessary adversarial perturbation required to switch an example's label. The addition of a second regularizer that encourages linear decision boundaries improves robustness while avoiding over-regularization of the classifier. We demonstrate the robustness of our approach with respect to ℓ_∞ and ℓ_2 adversarial perturbations on multiple datasets. Our method is competitive with state-of-the-art algorithms for learning robust networks. Moreover, applying our algorithm in conjunction with adversarial training boosts the robustness of classifiers even further.

****
A PAC-Bayesian Approach to Generalization Bounds for Graph Neural Networks
Renjie Liao,Raquel Urtasun,Richard Zemel
In this paper, we derive generalization bounds for two primary classes of graph neural networks (GNNs), namely graph convolutional networks (GCNs) and message passing GNNs (MPGNNs), via a PAC-Bayesian approach. Our result reveals that the maximum node degree and the spectral norm of the weights govern the generalization bounds of both models. We also show that our bound for GCNs is a natural generalization of the results developed in \citep{neyshabur2017pac} for fully-connected and convolutional neural networks. For MPGNNs, our PAC-Bayes bound improves over the Rademacher complexity based bound \citep{garg2020generalization}, showing a tighter dependency on the maximum node degree and the maximum hidden dimension. The key ingredients of our proofs are a perturbation analysis of GNNs and the generalization of PAC-Bayes analysis to non-homogeneous GNNs. We perform an empirical study on several synthetic and real-world graph datasets and verify that our PAC-Bayes bound is tighter than others.
****
Online Continual Learning Under Domain Shift
Quang Pham,Chenghao Liu,Steven HOI
Existing continual learning benchmarks often assume each task's training and test data are from the same distribution, which may not hold in practice. Towards making continual learning practical, in this paper, we introduce a novel setting of online continual learning under conditional domain shift, in which domain shift exists between training and test data of all tasks: $P^{tr}(X, Y) \neq P^{te}(X,Y)$, and the model is required to generalize to unseen domains at test time. To address this problem, we propose \emph{Conditional Invariant Experience Replay (CIER)} that can simultaneously retain old knowledge, acquire new information, and generalize to unseen domains. CIER employs an adversarial training to correct the shift in $P(X,Y)$ by matching $P(X|Y)$, which results in an invariant representation that can generalize to unseen domains during inference. Our extensive experiments show that CIER can bridge the domain gap in continual learning and significantly outperforms state-of-the-art methods. We will release our benchmarks and implementation upon acceptance.
****
You Only Sample (Almost) Once: Linear Cost Self-Attention Via Bernoulli Sampling
Zhanpeng Zeng,Yunyang Xiong,Sathya N. Ravi,Shailesh Acharya,Glenn Fung,Vikas Singh
Transformer-based models have come to dominate the landscape in a wide range of natural language processing (NLP) applications. The heart of the transformer model is the self-attention mechanism, which captures the interactions of token pairs in the input sequences and consequently, depends quadratically on the input sequence length. It is known that training such models on longer sequences is quite expensive, and often, prohibitively so. We show that a Bernoulli sampling attention mechanism based on Locality Sensitive Hashing (LSH), decreases the quadratic complexity to linear. We bypass the quadratic cost by considering self-attention as a sum of individual tokens associated with Bernoulli random variables that can, in principle, be sampled at once by a single hash (although in practice, this number may be a small constant). This leads to an efficient sampling scheme to estimate self-attention which relies on specific modifications of LSH (based on feasibility of deployment on GPU architectures). We evaluate our proposed algorithm on the GLUE benchmark with standard 512 sequence length and our method achieves comparable or even slightly better performance than a standard pretrained Transformer. To evaluate whether our method can indeed handle longer sequences, we conduct experiments on long sequence (4096) language model pretraining and achieve consistent results as standard self-attention, while observing sizable inference speed-ups and memory savings.
****
DEMI: Discriminative Estimator of Mutual Information
Ruizhi Liao,Daniel Moyer,Polina Golland,William M Wells

Estimating mutual information between continuous random variables is often intra
ctable and extremely challenging for high-dimensional data. Recent progress has
leveraged neural networks to optimize variational lower bounds on mutual informa
tion. Although showing promise for this difficult problem, the variational metho
ds have been theoretically and empirically proven to have serious statistical li
mitations: 1) many methods struggle to produce accurate estimates when the under
lying mutual information is either low or high; 2) the resulting estimators may
suffer from high variance. Our approach is based on training a classifier that p
rovides the probability that a data sample pair is drawn from the joint distribu
tion rather than from the product of its marginal distributions. Moreover, we es
tablish a direct connection between mutual information and the average log odds
estimate produced by the classifier on a test set, leading to a simple and accur
ate estimator of mutual information. We show theoretically that our method and o
ther variational approaches are equivalent when they achieve their optimum, whil
e our method sidesteps the variational bound. Empirical results demonstrate high
 accuracy of our approach and the advantages of our estimator in the context of
representation learning.

****
Contextual Transformation Networks for Online Continual Learning
Quang Pham,Chenghao Liu,Doyen Sahoo,Steven HOI
Continual learning methods with fixed architectures rely on a single network to
learn models that can perform well on all tasks.
As a result, they often only accommodate common features of those tasks but negl
ect each task's specific features. On the other hand, dynamic architecture metho
ds can have a separate network for each task, but they are too expensive to trai
n and not scalable in practice, especially in online settings.
To address this problem, we propose a novel online continual learning method nam
ed ``Contextual Transformation Networks'' (CTN) to efficiently model the \emph{ta
sk-specific features} while enjoying neglectable complexity overhead compared to
 other fixed architecture methods.
Moreover, inspired by the Complementary Learning Systems (CLS) theory, we propos
e a novel dual memory design and an objective to train CTN that can address both
 catastrophic forgetting and knowledge transfer simultaneously.
Our extensive experiments show that CTN is competitive with a large scale dynami
c architecture network and consistently outperforms other fixed architecture met
hods under the same standard backbone. Our implementation can be found at \url{h
ttps://github.com/phquang/Contextual-Transformation-Network}.
****
Imitation with Neural Density Models
Kuno Kim,Akshat Jindal,Yang Song,Jiaming Song,Yanan Sui,Stefano Ermon
We propose a new framework for Imitation Learning (IL) via density estimation of
 the expert's occupancy measure followed by Maximum Occupancy Entropy Reinforcem
ent Learning (RL) using the density as a reward. Our approach maximizes a non-ad
versarial model-free RL objective that provably lower bounds reverse Kullback–Le
ibler divergence between occupancy measures of the expert and imitator. We prese
nt a practical IL algorithm, Neural Density Imitation (NDI), which obtains state
-of-the-art demonstration efficiency on benchmark control tasks.
****
Continual Lifelong Causal Effect Inference with Real World Evidence
Zhixuan Chu,Stephen Rathbun,Sheng Li
The era of real world evidence has witnessed an increasing availability of obser
vational data, which much facilitates the development of causal effect inference
. Although significant advances have been made to overcome the challenges in cau
sal effect estimation, such as missing counterfactual outcomes and selection bia
s, they only focus on source-specific and stationary observational data. In this
 paper, we investigate a new research problem of causal effect inference from in
crementally available observational data, and present three new evaluation crite
ria accordingly, including extensibility, adaptability, and accessibility. We pr
opose a Continual Causal Effect Representation Learning method for estimating ca

usal effect with observational data, which are incrementally available from non-stationary data distributions. Instead of having access to all seen observational data, our method only stores a limited subset of feature representations learned from previous data. Combining the selective and balanced representation learning, feature representation distillation, and feature transformation, our method achieves the continual causal effect estimation for new data without compromising the estimation capability for original data. Extensive experiments demonstrate the significance of continual causal effect inference and the effectiveness of our method.
****

Improving Learning to Branch via Reinforcement Learning
Haoran Sun,Wenbo Chen,Hui Li,Le Song
Branch-and-Bound~(B\&B) is a general and widely used algorithm paradigm for solving Mixed Integer Programming~(MIP).
Recently there is a surge of interest in designing learning-based branching policies as a fast approximation of strong branching, a human-designed heuristic. In this work, we argue strong branching is not a good expert to imitate for its poor decision quality when turning off its side effects in solving linear programming. To obtain more effective and non-myopic policies than a local heuristic, we formulate the branching process in MIP as reinforcement learning~(RL) and design a policy characterization for the B\&B process to improve our agent by novelty search evolutionary strategy. Across a range of NP-hard problems, our trained RL agent significantly outperforms expert-designed branching rules and the state-of-the-art learning-based branching methods in terms of both speed and effectiveness. Our results suggest that with carefully designed policy networks and learning algorithms, reinforcement learning has the potential to advance algorithms for solving MIPs.
****

Offline Policy Optimization with Variance Regularization
Riashat Islam,Samarth Sinha,Homanga Bharadhwaj,Samin Yeasar Arnob,Zhuoran Yang,Zhaoran Wang,Animesh Garg,Lihong Li,Doina Precup
Learning policies from fixed offline datasets is a key challenge to scale up reinforcement learning (RL) algorithms towards practical applications. This is often because off-policy RL algorithms suffer from distributional shift, due to mismatch between dataset and the target policy, leading to high variance and over-estimation of value functions. In this work, we propose variance regularization for offline RL algorithms, using stationary distribution corrections. We show that by using Fenchel duality, we can avoid double sampling issues for computing the gradient of the variance regularizer. The proposed algorithm for offline variance regularization can be used to augment any existing offline policy optimization algorithms. We show that the regularizer leads to a lower bound to the offline policy optimization objective, which can help avoid over-estimation errors, and explains the benefits of our approach across a range of continuous control domains when compared to existing algorithms.
****

Synthetic Petri Dish: A Novel Surrogate Model for Rapid Architecture Search
Aditya Rawal,Joel Lehman,Felipe Petroski Such,Jeff Clune,Kenneth Stanley
Neural Architecture Search (NAS) explores a large space of architectural motifs --
a compute-intensive process that often involves ground-truth evaluation of each motif by instantiating it within a large network, and training and evaluating the network with thousands or more data samples. Inspired by how biological motifs such as cells are sometimes extracted from their natural environment and studied in an artificial Petri dish setting, this paper proposes the Synthetic Petri Dish model for evaluating architectural motifs. In the Synthetic Petri Dish, architectural motifs are instantiated in very small networks and evaluated using very few learned synthetic data samples (to effectively approximate performance in the full problem). The relative performance of motifs in the Synthetic Petri Dish can substitute for their ground-truth performance, thus accelerating the most expensive step of NAS. Unlike other neural network-based prediction models that

parse the structure of the motif to estimate its performance, the Synthetic Petri Dish predicts motif performance by training the actual motif in an artificial setting, thus deriving predictions from its true intrinsic properties. Experiments in this paper demonstrate that the Synthetic Petri Dish can therefore predict the performance of new motifs with significantly higher accuracy, especially when insufficient ground truth data is available.
Our hope is that this work can inspire a new research direction in studying the performance of extracted components of models in a synthetic diagnostic setting optimized to provide informative evaluations.
****

Causal Inference Q-Network: Toward Resilient Reinforcement Learning
Chao-Han Huck Yang,Danny I-Te Hung,Yi Ouyang,Pin-Yu Chen
Deep reinforcement learning (DRL) has demonstrated impressive performance in various gaming simulators and real-world applications. In practice, however, a DRL agent may receive faulty observation by abrupt interferences such as black-out, frozen-screen, and adversarial perturbation. How to design a resilient DRL algorithm against these rare but mission-critical and safety-crucial scenarios is an important yet challenging task. In this paper, we consider a resilient DRL framework with observational interferences. Under this framework, we discuss the importance of the causal relation and propose a causal inference based DRL algorithm called causal inference Q-network (CIQ). We evaluate the performance of CIQ in several benchmark DRL environments with different types of interferences. Our experimental results show that the proposed CIQ method could achieve higher performance and more resilience against observational interferences.
****

Learning-Augmented Sketches for Hessians
Yi Li,Honghao Lin,David Woodruff
We study learning-based sketching for Hessians, which is known to provide considerable speedups to second order optimization. A number of works have shown how to sketch or subsample the Hessian to speed up each iteration, but such sketches are usually specific to the matrix at hand, rather than being learned from a distribution. We extend such schemes to learned sketches, where we learn different potentially different sketches for the different iterations, and show empirically that learned sketches, compared with their "non-learned" counterparts, improve the approximation accuracy for a large number of important problems, including LASSO, SVM, and matrix estimation with nuclear norm constraints.
****

A self-explanatory method for the black box problem on discrimination part of CNN
Jinwei Zhao,Qizhou Wang,Wanli Qiu,Guo Xie,Wei Wang,Xinhong Hei,Deyu Meng
Recently, for finding inherent causality implied in CNN, the black box problem of its discrimination part, which is composed of all fully connected layers of the CNN, has been studied by different scientific communities. Many methods were proposed, which can extract various interpretable models from the optimal discrimination part based on inputs and outputs of the part for finding the inherent causality implied in the part. However, the inherent causality cannot readily be found. We think that the problem could be solved by shrinking an interpretable distance which can evaluate the degree for the discrimination part to be easily explained by an interpretable model. This paper proposes a lightweight interpretable model, Deep Cognitive Learning Model(DCLM). And then, a game method between the DCLM and the discrimination part is implemented for shrinking the interpretation distance. Finally, the proposed self-explanatory method was evaluated by some contrastive experiments with certain baseline methods on some standard image processing benchmarks. These experiments indicate that the proposed method can effectively find the inherent causality implied in the discrimination part of the CNN without largely reducing its generalization performance. Moreover, the generalization performance of the DCLM also can be improved.
****

Counterfactual Thinking for Long-tailed Information Extraction
Guoshun Nan,Jiaqi Zeng,Rui Qiao,Wei Lu

Information Extraction (IE) aims to extract structured information from unstructured texts. However, in practice, the long-tailed and imbalanced data may lead to severe bias issues for deep learning models, due to very few training instances available for the tail classes. Existing works are mainly from computer vision society, leveraging re-balancing, decoupling, transfer learning and causal inference to address this problem on image classification and scene graph generation. However, these approaches may not achieve good performance on textual data, which involves complex language structures that have been proven crucial for the IE tasks. To this end, we propose a novel framework (named CFIE) based on language structure and causal reasoning with three key ingredients. First, by fusing the syntax information to various structured causal models for mainstream IE tasks including relation extraction (RE), named entity recognition (NER), and event detection (ED), our approach is able to learn the direct effect for classification from an imbalanced dataset. Second, counterfactuals are generated based on an explicit language structure to better calculate the direct effect during the inference stage. Third, we propose a flexible debiasing approach for more robust prediction during the inference stage. Experimental results on three IE tasks across five public datasets show that our model significantly outperforms the state-of-the-art models by a large margin in terms of Mean Recall and Macro F1, achieving a relative 30% improvement in Mean Recall for 7 tail classes on the ACE2005 dataset. We also discuss some interesting findings based on our observations.
****
Disentangled Recurrent Wasserstein Autoencoder
Jun Han,Martin Renqiang Min,Ligong Han,Li Erran Li,Xuan Zhang
Learning disentangled representations leads to interpretable models and facilitates data generation with style transfer, which has been extensively studied on static data such as images in an unsupervised learning framework. However, only a few works have explored unsupervised disentangled sequential representation learning due to challenges of generating sequential data. In this paper, we propose recurrent Wasserstein Autoencoder (R-WAE), a new framework for generative modeling of sequential data. R-WAE disentangles the representation of an input sequence into static and dynamic factors (i.e., time-invariant and time-varying parts). Our theoretical analysis shows that, R-WAE minimizes an upper bound of a penalized form of the Wasserstein distance between model distribution and sequential data distribution, and simultaneously maximizes the mutual information between input data and different disentangled latent factors, respectively. This is superior to (recurrent) VAE which does not explicitly enforce mutual information maximization between input data and disentangled latent representations. When the number of actions in sequential data is available as weak supervision information, R-WAE is extended to learn a categorical latent representation of actions to improve its disentanglement. Experiments on a variety of datasets show that our models outperform other baselines with the same settings in terms of disentanglement and unconditional video generation both quantitatively and qualitatively.
****
Adaptive and Generative Zero-Shot Learning
Yu-Ying Chou,Hsuan-Tien Lin,Tyng-Luh Liu
We address the problem of generalized zero-shot learning (GZSL) where the task is to predict the class label of a target image whether its label belongs to the seen or unseen category. Similar to ZSL, the learning setting assumes that all class-level semantic features are given, while only the images of seen classes are available for training. By exploring the correlation between image features and the corresponding semantic features, the main idea of the proposed approach is to enrich the semantic-to-visual (S2V) embeddings via a seamless fusion of adaptive and generative learning. To this end, we extend the semantic features of each class by supplementing image-adaptive attention so that the learned S2V embedding can account for not only inter-class but also intra-class variations. In addition, to break the limit of training with images only from seen classes, we design a generative scheme to simultaneously generate virtual class labels and their visual features by sampling and interpolating over seen counterparts. In inference, a testing image will give rise to two different S2V embeddings, seen and

virtual. The former is used to decide whether the underlying label is of the unseen category or otherwise a specific seen class; the latter is to predict an unseen class label. To demonstrate the effectiveness of our method, we report state-of-the-art results on four standard GZSL datasets, including an ablation study of the proposed modules.
****
Prior Knowledge Representation for Self-Attention Networks
Kehai Chen,Rui Wang,Masao Utiyama,Eiichiro Sumita
Self-attention networks (SANs) have shown promising empirical results in various natural language processing tasks. Typically, it gradually learning language knowledge on the whole training dataset in parallel and stacked ways, thereby modeling language representation. In this paper, we propose a simple and general representation method to consider prior knowledge related to language representation from the beginning of training. Also, the proposed method allows SANs to leverage prior knowledge in a universal way compatible with neural networks. Furthermore, we apply it to one prior word frequency knowledge for the monolingual data and other prior translation lexicon knowledge for the bilingual data, respectively, thereby enhancing the language representation. Experimental results on WMT14 English-to-German and WMT17 Chinese-to-English translation tasks demonstrate the effectiveness and universality of the proposed method over a strong Transformer-based baseline.
****
Impact-driven Exploration with Contrastive Unsupervised Representations
Min Jae Song,Dan Kushnir
Procedurally-generated sparse reward environments pose significant challenges for many RL algorithms. The recently proposed impact-driven exploration method (RIDE) by Raileanu & Rocktäschel (2020), which rewards actions that lead to large changes (measured by ℓ_2-distance) in the observation embedding, achieves state-of-the-art performance on such procedurally-generated MiniGrid tasks. Yet, the definition of "impact" in RIDE is not conceptually clear because its learned embedding space is not inherently equipped with any similarity measure, let alone ℓ_2-distance. We resolve this issue in RIDE via contrastive learning. That is, we train the embedding with respect to cosine similarity, where we define two observations to be similar if the agent can reach one observation from the other within a few steps, and define impact in terms of this similarity measure. Experimental results show that our method performs similarly to RIDE on the MiniGrid benchmarks while learning a conceptually clear embedding space equipped with the cosine similarity measure. Our modification of RIDE also provides a new perspective which connects RIDE and episodic curiosity (Savinov et al., 2019), a different exploration method which rewards the agent for visiting states that are unfamiliar to the agent's episodic memory. By incorporating episodic memory into our method, we outperform RIDE on the MiniGrid benchmarks.
****
Contrastive Self-Supervised Learning of Global-Local Audio-Visual Representations
Shuang Ma,Zhaoyang Zeng,Daniel McDuff,Yale Song
Contrastive self-supervised learning has delivered impressive results in many audio-visual recognition tasks. However, existing approaches optimize for learning either global representations useful for high-level understanding tasks such as classification, or local representations useful for tasks such as audio-visual source localization and separation. While they produce satisfactory results in their intended downstream scenarios, they often fail to generalize to tasks that they were not originally designed for. In this work, we propose a versatile self-supervised approach to learn audio-visual representations that can generalize to both the tasks which require global semantic information (e.g., classification) and the tasks that require fine-grained spatio-temporal information (e.g. localization). We achieve this by optimizing two cross-modal contrastive objectives that together encourage our model to learn discriminative global-local visual information given audio signals. To show that our approach learns generalizable video representations, we evaluate it on various downstream scenarios including ac

tion/sound classification, lip reading, deepfake detection, and sound source localization.
****
Safe Reinforcement Learning with Natural Language Constraints
Tsung-Yen Yang,Michael Hu,Yinlam Chow,Peter Ramadge,Karthik R Narasimhan
In this paper, we tackle the problem of learning control policies for tasks when provided with constraints in natural language. In contrast to instruction following, language here is used not to specify goals, but rather to describe situations that an agent must avoid during its exploration of the environment. Specifying constraints in natural language also differs from the predominant paradigm in safe reinforcement learning, where safety criteria are enforced by hand-defined cost functions. While natural language allows for easy and flexible specification of safety constraints and budget limitations, its ambiguous nature presents a challenge when mapping these specifications into representations that can be used by techniques for safe reinforcement learning. To address this, we develop a model that contains two components: (1) a constraint interpreter to encode natural language constraints into vector representations capturing spatial and temporal information on forbidden states, and (2) a policy network that uses these representations to output a policy with minimal constraint violations. Our model is end-to-end differentiable and we train it using a recently proposed algorithm for constrained policy optimization. To empirically demonstrate the effectiveness of our approach, we create a new benchmark task for autonomous navigation with crowd-sourced free-form text specifying three different types of constraints. Our method outperforms several baselines by achieving 6-7 times higher returns and 76% fewer constraint violations on average. Dataset and code to reproduce our experiments are available at https://sites.google.com/view/polco-hazard-world/.
****
Iterated learning for emergent systematicity in VQA
Ankit Vani,Max Schwarzer,Yuchen Lu,Eeshan Dhekane,Aaron Courville
Although neural module networks have an architectural bias towards compositionality, they require gold standard layouts to generalize systematically in practice. When instead learning layouts and modules jointly, compositionality does not arise automatically and an explicit pressure is necessary for the emergence of layouts exhibiting the right structure. We propose to address this problem using iterated learning, a cognitive science theory of the emergence of compositional languages in nature that has primarily been applied to simple referential games in machine learning. Considering the layouts of module networks as samples from an emergent language, we use iterated learning to encourage the development of structure within this language. We show that the resulting layouts support systematic generalization in neural agents solving the more complex task of visual question-answering. Our regularized iterated learning method can outperform baselines without iterated learning on SHAPES-SyGeT (SHAPES Systematic Generalization Test), a new split of the SHAPES dataset we introduce to evaluate systematic generalization, and on CLOSURE, an extension of CLEVR also designed to test systematic generalization. We demonstrate superior performance in recovering ground-truth compositional program structure with limited supervision on both SHAPES-SyGeT and CLEVR.
****
Online Adversarial Purification based on Self-supervised Learning
Changhao Shi,Chester Holtz,Gal Mishne
Deep neural networks are known to be vulnerable to adversarial examples, where a perturbation in the input space leads to an amplified shift in the latent network representation. In this paper, we combine canonical supervised learning with self-supervised representation learning, and present Self-supervised Online Adversarial Purification (SOAP), a novel defense strategy that uses a self-supervised loss to purify adversarial examples at test-time. Our approach leverages the label-independent nature of self-supervised signals and counters the adversarial perturbation with respect to the self-supervised tasks. SOAP yields competitive robust accuracy against state-of-the-art adversarial training and purification methods, with considerably less training complexity. In addition, our approach is

robust even when adversaries are given the knowledge of the purification defens e strategy. To the best of our knowledge, our paper is the first that generalize s the idea of using self-supervised signals to perform online test-time purifica tion.
****
Optimal Neural Program Synthesis from Multimodal Specifications
Xi Ye,Qiaochu Chen,Isil Dillig,Greg Durrett
Multimodal program synthesis, which leverages different types of user input to s ynthesize a desired program, is an attractive way to scale program synthesis to challenging settings; however, it requires integrating noisy signals from the us er (like natural language) with hard constraints on the program's behavior. This paper proposes an optimal neural synthesis approach where the goal is to find a program that satisfies user-provided constraints while also maximizing the prog ram's score with respect to a neural model. Specifically, we focus on multimodal synthesis tasks in which the user intent is expressed using combination of natu ral language (NL) and input-output examples. At the core of our method is a top-down recurrent neural model that places distributions over abstract syntax trees conditioned on the NL input. This model not only allows for efficient search ov er the space of syntactically valid programs, but it allows us to leverage autom ated program analysis techniques for pruning the search space based on infeasibi lity of partial programs with respect to the user's constraints. The experimenta l results on a multimodal synthesis dataset (StructuredRegex) show that our meth od substantially outperforms prior state-of-the-art techniques in terms of accur acy and explores fewer states during search.
****
FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders
Pengyu Cheng,Weituo Hao,Siyang Yuan,Shijing Si,Lawrence Carin
Pretrained text encoders, such as BERT, have been applied increasingly in variou s natural language processing (NLP) tasks, and have recently demonstrated signif icant performance gains. However, recent studies have demonstrated the existence of social bias in these pretrained NLP models. Although prior works have made p rogress on word-level debiasing, improved sentence-level fairness of pretrained encoders still lacks exploration. In this paper, we proposed the first neural de biasing method for a pretrained sentence encoder, which transforms the pretraine d encoder outputs into debiased representations via a fair filter (FairFil) netw ork. To learn the FairFil, we introduce a contrastive learning framework that no t only minimizes the correlation between filtered embeddings and bias words but also preserves rich semantic information of the original sentences. On real-worl d datasets, our FairFil effectively reduces the bias degree of pretrained text e ncoders, while continuously showing desirable performance on downstream tasks. M oreover, our post hoc method does not require any retraining of the text encoder s, further enlarging FairFil's application space.
****
Demystifying Learning of Unsupervised Neural Machine Translation
Guanlin Li,lemao liu,Taro Watanabe,Conghui Zhu,Tiejun Zhao
Unsupervised Neural Machine Translation or UNMT has received great attention in recent years. Though tremendous empirical improvements have been achieved, there still lacks theory-oriented investigation and thus some fundamental questions like \textit{why} certain training protocol can work or not under \textit{what} circumstances have not yet been well understood. This paper attempts to provide theoretical insights for the above questions. Specifically, following the methodology of comparative study, we leverage two perspectives, i) \textit{marginal likelihood maximization} and ii) \textit{mutual information} from information theory, to understand the different learning effects from the standard training protocol and its variants. Our detailed analyses reveal several critical conditions for the successful training of UNMT.
****
Searching for Convolutions and a More Ambitious NAS
Nicholas Carl Roberts,Mikhail Khodak,Tri Dao,Liam Li,Nina Balcan,Christopher Re, Ameet Talwalkar

An important goal of neural architecture search (NAS) is to automate-away the de
sign of neural networks on new tasks in under-explored domains, thus helping to
democratize machine learning. However, current NAS research largely focuses on s
earch spaces consisting of existing operations---such as different types of conv
olution---that are already known to work well on well-studied problems---often i
n computer vision. Our work is motivated by the following question: can we enabl
e users to build their own search spaces and discover the right neural operation
s given data from their specific domain? We make progress towards this broader v
ision for NAS by introducing a space of operations generalizing the convolution
that enables search over a large family of parameterizable linear-time matrix-ve
ctor functions. Our flexible construction allows users to design their own searc
h spaces adapted to the nature and shape of their data, to warm-start search met
hods using convolutions when they are known to perform well, or to discover new
operations from scratch when they do not. We evaluate our approach on several no
vel search spaces over vision and text data, on all of which simple NAS search a
lgorithms can find operations that perform better than baseline layers.
****

not-so-big-GAN: Generating High-Fidelity Images on Small Compute with Wavelet-ba
sed Super-Resolution
Seungwook Han,Akash Srivastava,Cole Lincoln Hurwitz,Prasanna Sattigeri,David Dan
iel Cox
State-of-the-art models for high-resolution image generation, such as BigGAN and
 VQVAE-2, require an incredible amount of compute resources and/or time (512 TPU
-v3 cores) to train, putting them out of reach for the larger research community
. On the other hand, GAN-based image super-resolution models, such as ESRGAN, ca
n not only upscale images to high dimensions, but also are efficient to train. I
n this paper, we present not-so-big-GAN (nsb-GAN), a simple yet cost-effective t
wo-step training framework for deep generative models (DGMs) of high-dimensional
 natural images. First, we generate images in low-frequency bands by training a
sampler in the wavelet domain. Then, we super-resolve these images from the wave
let domain back to the pixel-space with our novel wavelet super-resolution decod
er network. Wavelet-based down-sampling method preserves more structural informa
tion than pixel-based methods, leading to significantly better generative qualit
y of the low-resolution sampler (e.g., 64×64). Since the sampler and decoder can
 be trained in parallel and operate on much lower dimensional spaces than end-to
-end models, the training cost is substantially reduced. On ImageNet 512×512, ou
r model achieves a Fréchet Inception Distance (FID) of 10.59 – beating the basel
ine BigGAN model – at half the compute (256 TPU-v3 cores).
****

Uncertainty in Neural Processes
Saeid Naderiparizi,Kenny Chiu,Benjamin Bloem-Reddy,Frank Wood
We explore the effects of architecture and training objective choice on amortize
d posterior predictive inference in probabilistic conditional generative models.
  We aim this work to be a counterpoint to a recent trend in the literature that
 stresses achieving good samples when the amount of conditioning data is large.
 We instead focus our attention on the case where the amount of conditioning dat
a is small. We highlight specific architecture and objective choices that we fi
nd lead to qualitative and quantitative improvement to posterior inference in th
is low data regime. Specifically we explore the effects of choices of pooling o
perator and variational family on posterior quality in neural processes. Superi
or posterior predictive samples drawn from our novel neural process architecture
s are demonstrated via image completion/in-painting experiments.
****

Reset-Free Lifelong Learning with Skill-Space Planning
Kevin Lu,Aditya Grover,Pieter Abbeel,Igor Mordatch
The objective of \textit{lifelong} reinforcement learning (RL) is to optimize ag
ents which can continuously adapt and interact in changing environments. However
, current RL approaches fail drastically when environments are non-stationary an
d interactions are non-episodic. We propose \textit{Lifelong Skill Planning} (Li
SP), an algorithmic framework for lifelong RL based on planning in an abstract s

pace of higher-order skills. We learn the skills in an unsupervised manner using intrinsic rewards and plan over the learned skills using a learned dynamics model. Moreover, our framework permits skill discovery even from offline data, thereby reducing the need for excessive real-world interactions. We demonstrate empirically that LiSP successfully enables long-horizon planning and learns agents that can avoid catastrophic failures even in challenging non-stationary and non-episodic environments derived from gridworld and MuJoCo benchmarks.
****
Outlier Robust Optimal Transport
Debarghya Mukherjee,Aritra Guha,Justin Solomon,Yuekai Sun,Mikhail Yurochkin
Optimal transport (OT) provides a way of measuring distances between distributions that depends on the geometry of the sample space. In light of recent advances in solving the OT problem, OT distances are widely used as loss functions in minimum distance estimation. Despite its prevalence and advantages, however, OT is extremely sensitive to outliers. A single adversarially-picked outlier can increase OT distance arbitrarily. To address this issue, in this work we propose an outlier-robust OT formulation. Our formulation is convex but challenging to scale at a first glance. We proceed by deriving an equivalent formulation based on cost truncation that is easy to incorporate into modern stochastic algorithms for regularized OT. We demonstrate our model applied to mean estimation under the Huber contamination model in simulation as well as outlier detection on real data.
****
An Attention Free Transformer
Shuangfei Zhai,Walter Talbott,Nitish Srivastava,Chen Huang,Hanlin Goh,Joshua M. Susskind
We introduce Attention Free Transformer (AFT), an efficient variant of Transformers \citep{transformer} that eliminates the need for dot product attention. AFT offers great simplicity and efficiency compared with standard Transformers, where the multi-head attention operation is replaced with the composition of element-wise multiplications/divisions and global/local pooling. During training time, AFT has linear time and space complexity w.r.t. both the sequence length and feature dimension; in the autoregressive decoding mode, AFT has constant memory and time complexity per step. We show that, surprisingly, we are able to train AFT effectively on challenging benchmarks, and also to match or surpass the standard Transformer counterparts and other efficient variants. In particular, AFT achieves the state-of-the-art result on CIFAR10 autoregressive modeling with much reduced complexity, and also outperforms several efficient Transformer variants on Enwik8.
****
Private Split Inference of Deep Networks
Mohammad Samragh,Hossein Hosseini,Kambiz Azarian,Joseph Soriaga
Splitting network computations between the edge device and the cloud server is a promising approach for enabling low edge-compute and private inference of neural networks. Current methods for providing the privacy train the model to minimize information leakage for a given set of private attributes. In practice, however, the test queries might contain private attributes that are not foreseen during training.
We propose an alternative solution, in which, instead of obfuscating the information corresponding to a set of attributes, the edge device discards the information irrelevant to the main task. To this end, the edge device runs the model up to a split layer determined based on its computational capacity and then removes the activation content that is in the null space of the next layer of the model before sending it to the server. It can further remove the low-energy components of the remaining signal to improve the privacy at the cost of reducing the accuracy. The experimental results show that our methods provide privacy while maintaining the accuracy and introducing only a small computational overhead.
****
Uniform-Precision Neural Network Quantization via Neural Channel Expansion
Seongmin Park,Beomseok Kwon,Kyuyoung Sim,Jieun Lim,Tae-Ho Kim,Jungwook Choi

Uniform-precision neural network quantization has gained popularity thanks to its simple arithmetic unit densely packed for high computing capability. However, it ignores heterogeneous sensitivity to the impact of quantization across the layers, resulting in sub-optimal inference accuracy. This work proposes a novel approach to adjust the network structure to alleviate the impact of uniform-precision quantization. The proposed neural architecture search selectively expands channels for the quantization sensitive layers while satisfying hardware constraints (e.g., FLOPs). We provide substantial insights and empirical evidence that the proposed search method called neural channel expansion can adapt several popular networks' channels to achieve superior 2-bit quantization accuracy on CIFAR10 and ImageNet. In particular, we demonstrate the best-to-date Top-1/Top-5 accuracy for 2-bit ResNet50 with smaller FLOPs and the parameter size.
****
Efficient Empowerment Estimation for Unsupervised Stabilization
Ruihan Zhao,Kevin Lu,Pieter Abbeel,Stas Tiomkin
Intrinsically motivated artificial agents learn advantageous behavior without externally-provided rewards. Previously, it was shown that maximizing mutual information between agent actuators and future states, known as the empowerment principle, enables unsupervised stabilization of dynamical systems at upright positions, which is a prototypical intrinsically motivated behavior for upright standing and walking. This follows from the coincidence between the objective of stabilization and the objective of empowerment. Unfortunately, sample-based estimation of this kind of mutual information is challenging. Recently, various variational lower bounds (VLBs) on empowerment have been proposed as solutions; however, they are often biased, unstable in training, and have high sample complexity. In this work, we propose an alternative solution based on a trainable representation of a dynamical system as a Gaussian channel, which allows us to efficiently calculate an unbiased estimator of empowerment by convex optimization. We demonstrate our solution for sample-based unsupervised stabilization on different dynamical control systems and show the advantages of our method by comparing it to the existing VLB approaches. Specifically, we show that our method has a lower sample complexity, is more stable in training, possesses the essential properties of the empowerment function, and allows estimation of empowerment from images. Consequently, our method opens a path to wider and easier adoption of empowerment for various applications.
****
Efficient Graph Neural Architecture Search
Huan Zhao,Lanning Wei,quanming yao,Zhiqiang He
Recently, graph neural networks (GNN) have been demonstrated effective in various graph-based tasks.
To obtain state-of-the-art (SOTA) data-specific GNN architectures, researchers turn to the neural architecture search (NAS) methods.
However, it remains to be a challenging problem to conduct efficient architecture search for GNN.
In this work, we present a novel framework for Efficient GrAph Neural architecture search (EGAN).
By designing a novel and expressive search space, an efficient one-shot NAS method based on stochastic relaxation and natural gradient is proposed.
Further, to enable architecture search in large graphs, a transfer learning paradigm is designed.
Extensive experiments, including node-level and graph-level tasks, are conducted. The results show that the proposed EGAN can obtain SOTA data-specific architectures, and reduce the search cost by two orders of magnitude compared to existing NAS baselines.
****
Self-supervised and Supervised Joint Training for Resource-rich Machine Translation
Yong Cheng,Wei Wang.,Lu Jiang,Wolfgang Macherey
Self-supervised pre-training of text representations has been successfully applied to low-resource Neural Machine Translation (NMT). However, it usually fails t

o achieve notable gains on resource-rich NMT. In this paper, we propose a joint training approach, F_2-XEnDec, to combine self-supervised and supervised learning to optimize NMT models. To exploit complementary self-supervised signals for supervised learning, NMT models are trained on examples that are interbred from monolingual and parallel sentences through a new process called crossover encoder-decoder. Experiments on two resource-rich translation bench-marks, WMT'14 English-German and WMT'14 English-French, demonstrate that our approach achieves substantial improvements over a vanilla Transformer and obtains a new state of the art of 46 BLEU on English-French. Results also show that our approach is capable of improving model robustness against input perturbations which is known as a key weakness in contemporary NMT systems.

****

Speeding up Deep Learning Training by Sharing Weights and Then Unsharing

Shuo Yang,Le Hou,Xiaodan Song,qiang liu,Denny Zhou

It has been widely observed that increasing deep learning model sizes often leads to significant performance improvements on a variety of natural language processing and computer vision tasks. In the meantime, however, computational costs and training time would dramatically increase when models get larger. In this paper, we propose a simple approach to speed up training for a particular kind of deep networks which contain repeated structures, such as the transformer module. In our method, we first train such a deep network with the weights shared across all the repeated layers till some point. We then stop weight sharing and continue training until convergence. The untying point is automatically determined by monitoring gradient statistics. Our adaptive untying criterion is obtained from a theoretic analysis over deep linear networks. Empirical results show that our method is able to reduce the training time of BERT by 50%.

****

Robust Multi-Agent Reinforcement Learning Driven by Correlated Equilibrium

Yizheng Hu,Kun Shao,Dong Li,Jianye HAO,Wulong Liu,Yaodong Yang,Jun Wang,Zhanxing Zhu

In this paper we deal with robust cooperative multi-agent reinforcement learning (CMARL). While CMARL has many potential applications, only a trained policy that is robust enough can be confidently deployed in real world. Existing works on robust MARL mainly apply vanilla adversarial training in centralized training and decentralized execution paradigm. We, however, find that if a CMARL environment contains an adversarial agent, the performance of decentralized equilibrium might perform significantly poor for achieving such adversarial robustness. To tackle this issue, we suggest that when execution the non-adversarial agents must jointly make the decision to improve the robustness, therefore solving correlated equilibrium instead. We theoretically demonstrate the superiority of correlated equilibrium over the decentralized one in adversarial MARL settings. Therefore, to achieve robust CMARL, we introduce novel strategies to encourage agents to learn correlated equilibrium while maximally preserving the convenience of the decentralized execution. The global variables with mutual information are proposed to help agents learn robust policies with MARL algorithms. The experimental results show that our method can dramatically boost performance on the SMAC environments.

****

Driving through the Lens: Improving Generalization of Learning-based Steering using Simulated Adversarial Examples

Yu Shen,Laura Yu Zheng,Manli Shu,Weizi Li,Tom Goldstein,Ming Lin

To ensure the wide adoption and safety of autonomous driving, the vehicles need to be able to drive under various lighting, weather, and visibility conditions in different environments. These external and environmental factors, along with internal factors associated with sensors, can pose significant challenges to perceptual data processing, hence affecting the decision-making of the vehicle. In this work, we address this critical issue by analyzing the sensitivity of the learning algorithm with respect to varying quality in the image input for autonomous driving. Using the results of sensitivity analysis, we further propose an algorithm to improve the overall performance of the task of ``learning to steer''.

The results show that our approach is able to enhance the learning outcomes up t
o 48%. A comparative study drawn between our approach and other related techniqu
es, such as data augmentation and adversarial training, confirms the effectivene
ss of our algorithm as a way to improve the robustness and generalization of neu
ral network training for self-driving cars.
****

An Open Review of OpenReview: A Critical Analysis of the Machine Learning Confer
ence Review Process
David Tran,Alexander V Valtchanov,Keshav R Ganapathy,Raymond Feng,Eric Victor Sl
ud,Micah Goldblum,Tom Goldstein
Mainstream machine learning conferences have seen a dramatic increase in the num
ber of participants, along with a growing range of perspectives, in recent years
. Members of the machine learning community are likely to overhear allegations
ranging from randomness of acceptance decisions to institutional bias. In this
work, we critically analyze the review process through a comprehensive study of
papers submitted to ICLR between 2017 and 2020. We quantify reproducibility/ran
domness in review scores and acceptance decisions, and examine whether scores co
rrelate with paper impact. Our findings suggest strong institutional bias in ac
cept/reject decisions, even after controlling for paper quality. Furthermore, we
 find evidence for a gender gap, with female authors receiving lower scores, low
er acceptance rates, and fewer citations per paper than their male counterparts.
  We conclude our work with recommendations for future conference organizers.
****

MixKD: Towards Efficient Distillation of Large-scale Language Models
Kevin J Liang,Weituo Hao,Dinghan Shen,Yufan Zhou,Weizhu Chen,Changyou Chen,Lawre
nce Carin
Large-scale language models have recently demonstrated impressive empirical perf
ormance. Nevertheless, the improved results are attained at the price of bigger
models, more power consumption, and slower inference, which hinder their applica
bility to low-resource (both memory and computation) platforms. Knowledge distil
lation (KD) has been demonstrated as an effective framework for compressing such
 big models. However, large-scale neural network systems are prone to memorize t
raining instances, and thus tend to make inconsistent predictions when the data
distribution is altered slightly. Moreover, the student model has few opportunit
ies to request useful information from the teacher model when there is limited t
ask-specific data available. To address these issues, we propose MixKD, a data-a
gnostic distillation framework that leverages mixup, a simple yet efficient data
 augmentation approach, to endow the resulting model with stronger generalizatio
n ability. Concretely, in addition to the original training examples, the studen
t model is encouraged to mimic the teacher's behavior on the linear interpolatio
n of example pairs as well. We prove from a theoretical perspective that under r
easonable conditions MixKD gives rise to a smaller gap between the generalizatio
n error and the empirical error. To verify its effectiveness, we conduct experim
ents on the GLUE benchmark, where MixKD consistently leads to significant gains
over the standard KD training, and outperforms several competitive baselines. Ex
periments under a limited-data setting and ablation studies further demonstrate
the advantages of the proposed approach.
****

AR-ELBO: Preventing Posterior Collapse Induced by Oversmoothing in Gaussian VAE
Yuhta Takida,Wei-Hsiang Liao,Toshimitsu Uesaka,Shusuke Takahashi,Yuki Mitsufuji
Variational autoencoders (VAEs) often suffer from posterior collapse, which is a
 phenomenon that the learned latent space becomes uninformative. This is related
 to local optima introduced by a fixed hyperparameter resembling the data varian
ce in the objective function. We suggest that this variance parameter regularize
s the VAE and affects its smoothness, which is the magnitude of its gradient. An
 inappropriate choice of this parameter causes oversmoothness and leads to poste
rior collapse. This is shown theoretically by analysis on the linear approximate
d objective function and empirically in general cases. We propose AR-ELBO, which
 stands for adaptively regularized ELBO~(Evidence Lower BOund). It controls the
strength of regularization by adapting the variance parameter, and thus avoids o

versmoothing the model. Generation models trained by proposed objectives show improved Fréchet inception distance~(FID) of images generated from the MNIST and CelebA datasets.
****
Consistency and Monotonicity Regularization for Neural Knowledge Tracing
Seewoo Lee,Youngduck Choi,Juneyoung Park,Byungsoo Kim,Jinwoo Shin
Knowledge Tracing (KT), tracking a human's knowledge acquisition, is a central component in online learning and AI in Education. In this paper, we present a simple, yet effective strategy to improve the generalization ability of KT models: we propose three types of novel data augmentation, coined replacement, insertion, and deletion, along with corresponding regularization losses that impose certain consistency or monotonicity bias on model's predictions for the original and augmented sequence. Extensive experiments on various KT benchmarks show that our regularization scheme significantly improve the prediction performances, under 3 widely-used neural networks and 4 public benchmarks for KT, e.g., it yields 6.3% improvement in AUC under the DKT model and the ASSISTmentsChall dataset.
****
DCT-SNN: Using DCT to Distribute Spatial Information over Time for Learning Low-Latency Spiking Neural Networks
Isha Garg,Sayeed Shafayet Chowdhury,Kaushik Roy
Spiking Neural Networks (SNNs) offer a promising alternative to traditional deep learning frameworks, since they provide higher
computational efficiency due to event-driven information processing. SNNs distribute the analog values of pixel intensities into binary spikes over time. However, the most widely used input coding schemes, such as Poisson based rate-coding, do not leverage the additional temporal learning capability of SNNs effectively. Moreover, these SNNs suffer from high inference latency which is a major bottleneck to their deployment. To overcome this, we propose a scalable time-based encoding scheme that utilizes the Discrete Cosine Transform (DCT) to reduce the number of timesteps required for inference. DCT decomposes an image into a weighted sum of sinusoidal basis images. At each time step, a single frequency base, taken in order and modulated
by its corresponding DCT coefficient, is input to an accumulator that generates spikes upon crossing a threshold. We use the proposed scheme to learn DCT-SNN, a low-latency deep SNN with leaky-integrate-and-fire neurons, trained using surrogate gradient descent based backpropagation. We achieve top-1 accuracy of 89.94%, 68.3% and 52.43% on CIFAR-10, CIFAR-100 and TinyImageNet, respectively using VGG architectures. Notably, DCT-SNN performs inference with 2-14X reduced latency compared to other state-of-the-art SNNs, while achieving comparable accuracy to their standard deep learning counterparts. The dimension of the transform allows us to control the number of timesteps required for inference. Additionally, we can trade-off accuracy with latency in a principled manner by dropping the highest frequency components during inference.
****
What's new? Summarizing Contributions in Scientific Literature
Hiroaki Hayashi,Wojciech Maciej Kryscinski,Bryan McCann,Nazneen Rajani,Caiming Xiong
With thousands of academic articles shared on a daily basis, it has become increasingly difficult to keep up with the latest scientific findings. To overcome this problem, we introduce a new task of $\textit{disentangled paper summarization}$, which seeks to generate separate summaries for the paper contributions and the context of the work, making it easier to identify the key findings shared in articles. For this purpose, we extend the S2ORC corpus of academic articles, which spans a diverse set of domains ranging from economics to psychology, by adding disentangled "contribution" and "context" reference labels. Together with the dataset, we introduce and analyze three baseline approaches: 1) a unified model controlled by input code prefixes, 2) a model with separate generation heads specialized in generating the disentangled outputs, and 3) a training strategy that guides the model using additional supervision coming from inbound and outbound citations. We also propose a comprehensive automatic evaluation protocol which r

eports the $\textit{relevance}$, $\textit{novelty}$, and $\textit{disentanglement}$ of generated outputs. Through a human study involving expert annotators, we show that in 79%, of cases our new task is considered more helpful than traditional scientific paper summarization.

****
CaPC Learning: Confidential and Private Collaborative Learning
Christopher A. Choquette-Choo,Natalie Dullerud,Adam Dziedzic,Yunxiang Zhang,Somesh Jha,Nicolas Papernot,Xiao Wang
Machine learning benefits from large training datasets, which may not always be possible to collect by any single entity, especially when using privacy-sensitive data. In many contexts, such as healthcare and finance, separate parties may wish to collaborate and learn from each other's data but are prevented from doing so due to privacy regulations. Some regulations prevent explicit sharing of data between parties by joining datasets in a central location (confidentiality). Others also limit implicit sharing of data, e.g., through model predictions (privacy). There is currently no method that enables machine learning in such a setting, where both confidentiality and privacy need to be preserved, to prevent both explicit and implicit sharing of data. Federated learning only provides confidentiality, not privacy, since gradients shared still contain private information. Differentially private learning assumes unreasonably large datasets. Furthermore, both of these learning paradigms produce a central model whose architecture was previously agreed upon by all parties rather than enabling collaborative learning where each party learns and improves their own local model. We introduce Confidential and Private Collaborative (CaPC) learning, the first method provably achieving both confidentiality and privacy in a collaborative setting. We leverage secure multi-party computation (MPC), homomorphic encryption (HE), and other techniques in combination with privately aggregated teacher models. We demonstrate how CaPC allows participants to collaborate without having to explicitly join their training sets or train a central model. Each party is able to improve the accuracy and fairness of their model, even in settings where each party has a model that performs well on their own dataset or when datasets are not IID and model architectures are heterogeneous across parties.
****
Multiplicative Filter Networks
Rizal Fathony,Anit Kumar Sahu,Devin Willmott,J Zico Kolter
Although deep networks are typically used to approximate functions over high dimensional inputs, recent work has increased interest in neural networks as function approximators for low-dimensional-but-complex functions, such as representing images as a function of pixel coordinates, solving differential equations, or representing signed distance fields or neural radiance fields. Key to these recent successes has been the use of new elements such as sinusoidal nonlinearities, or Fourier features in positional encodings, which vastly outperform simple ReLU networks. In this paper, we propose and empirically demonstrate that an arguably simpler class of function approximators can work just as well for such problems: multiplicative filter networks. In these networks, we avoid traditional compositional depth altogether, and simply multiply together (linear functions of) sinusoidal or Gabor wavelet functions applied to the input. This representation has the notable advantage that the entire function can simply be viewed as a linear function approximator over an exponential number of Fourier or Gabor basis functions, respectively. Despite this simplicity, when compared to recent approaches that use Fourier features with ReLU networks or sinusoidal activation networks, we show that these multiplicative filter networks largely outperform or match the performance of these recent approaches on the domains highlighted in these past works.
****
On the Robustness of Sentiment Analysis for Stock Price Forecasting
Gabriel Deza,Colin Rowat,Nicolas Papernot
Machine learning (ML) models are known to be vulnerable to attacks both at training and test time. Despite the extensive literature on adversarial ML, prior eff

orts focus primarily on applications of computer vision to object recognition or sentiment analysis to movie reviews. In these settings, the incentives for adversaries to manipulate the model's prediction are often unclear and attacks require extensive control of direct inputs to the model. This makes it difficult to evaluate how severe the impact of vulnerabilities exposed is on systems deploying ML with little provenance guarantees for the input data. In this paper, we study adversarial ML with stock price forecasting. Adversarial incentives are clear and may be quantified experimentally through a simulated portfolio. We replicate an industry standard pipeline, which performs a sentiment analysis of Twitter data to forecast trends in stock prices. We show that an adversary can exploit the lack of provenance to indirectly use tweets to manipulate the model's perceived sentiment about a target company and in turn force the model to forecast price erroneously. Our attack is mounted at test time and does not modify the training data. Given past market anomalies, we conclude with a series of recommendations for the use of machine learning as input signal to trading algorithms.

****

Signal Coding and Reconstruction using Spike Trains

Anik Chattopadhyay,Arunava Banerjee

In many animal sensory pathways, the transformation from external stimuli to spike trains is essentially deterministic. In this context, a new mathematical framework for coding and reconstruction, based on a biologically plausible model of the spiking neuron, is presented. The framework considers encoding of a signal through spike trains generated by an ensemble of neurons via a standard convolve-then-threshold mechanism, albeit with a wide variety of convolution kernels. Neurons are distinguished by their convolution kernels and threshold values. Reconstruction is posited as a convex optimization minimizing energy. Formal conditions under which perfect reconstruction of the signal from the spike trains is possible are then identified. Coding experiments on a large audio dataset are presented to demonstrate the strength of the framework.

****

XMixup: Efficient Transfer Learning with Auxiliary Samples by Cross-Domain Mixup

Xingjian Li,Haoyi Xiong,Haozhe An,Cheng-zhong Xu,Dejing Dou

Transferring knowledge from large source datasets is an effective way to fine-tune the deep neural networks of the target task with a small sample size. A great number of algorithms have been proposed to facilitate deep transfer learning, and these techniques could be generally categorized into two groups – Regularized Learning of the target task using models that have been pre-trained from source datasets, and Multitask Learning with both source and target datasets to train a shared backbone neural network. In this work, we aim to improve the multitask paradigm for deep transfer learning via Cross-domain Mixup (XMixup). While the existing multitask learning algorithms need to run backpropagation over both the source and target datasets and usually consume a higher gradient complexity, XMixup transfers the knowledge from source to target tasks more efficiently: for every class of the target task, XMixup selects the auxiliary samples from the source dataset and augments training samples via the simple mixup strategy. We evaluate XMixup over six real world transfer learning datasets. Experiment results show that XMixup improves the accuracy by 1.9% on average. Compared with other state-of-the-art transfer learning approaches, XMixup costs much less training time while still obtains higher accuracy.

****

Generating Plannable Lifted Action Models for Visually Generated Logical Predicates

Masataro Asai

We propose FOSAE++, an unsupervised end-to-end neural system that generates a compact discrete state transition model (dynamics / action model) from raw visual observations. Our representation can be exported to Planning Domain Description Language (PDDL), allowing symbolic state-of-the-art classical planners to perform high-level task planning on raw observations. FOSAE++ expresses states and actions in First Order Logic (FOL), a superset of so-called object-centric representation. It is the first unsupervised neural system that fully supports FOL in PD

DL action modeling, while existing systems are limited to continuous, propositio
nal, or property-based representations, and/or require manually labeled input fo
r actions/predicates/propositions.
****

Planning from Pixels using Inverse Dynamics Models
Keiran Paster,Sheila A. McIlraith,Jimmy Ba
Learning dynamics models in high-dimensional observation spaces can be challengi
ng for model-based RL agents. We propose a novel way to learn models in a latent
 space by learning to predict sequences of future actions conditioned on task co
mpletion. These models track task-relevant environment dynamics over a distribut
ion of tasks, while simultaneously serving as an effective heuristic for plannin
g with sparse rewards. We evaluate our method on challenging visual goal complet
ion tasks and show a substantial increase in performance compared to prior model
-free approaches.
****

Semi-supervised Keypoint Localization
Olga Moskvyak,Frederic Maire,Feras Dayoub,Mahsa Baktashmotlagh
Knowledge about the locations of keypoints of an object in an image can assist i
n fine-grained classification and identification tasks, particularly for the cas
e of objects that exhibit large variations in poses that greatly influence their
 visual appearance, such as wild animals. However, supervised training of a keyp
oint detection network requires annotating a large image dataset for each animal
 species, which is a labor-intensive task. To reduce the need for labeled data,
we propose to learn simultaneously keypoint heatmaps and pose invariant keypoint
 representations in a semi-supervised manner using a small set of labeled images
 along with a larger set of unlabeled images. Keypoint representations are learn
t with a semantic keypoint consistency constraint that forces the keypoint detec
tion network to learn similar features for the same keypoint across the dataset.
 Pose invariance is achieved by making keypoint representations for the image an
d its augmented copies closer together in feature space. Our semi-supervised app
roach significantly outperforms previous methods on several benchmarks for human
 and animal body landmark localization.
****

Influence Estimation for Generative Adversarial Networks
Naoyuki Terashita,Hiroki Ohashi,Yuichi Nonaka,Takashi Kanemaru
Identifying harmful instances, whose absence in a training dataset improves mode
l performance, is important for building better machine learning models.
Although previous studies have succeeded in estimating harmful instances under s
upervised settings, they cannot be trivially extended to generative adversarial
networks (GANs).
This is because previous approaches require that (i) the absence of a training i
nstance directly affects the loss value and that (ii) the change in the loss dir
ectly measures the harmfulness of the instance for the performance of a model.
In GAN training, however, neither of the requirements is satisfied.
This is because, (i) the generator’s loss is not directly affected by the traini
ng instances as they are not part of the generator's training steps, and (ii) th
e values of GAN's losses normally do not capture the generative performance of a
 model.
To this end, (i) we propose an influence estimation method that uses the Jacobia
n of the gradient of the generator's loss with respect to the discriminator’s pa
rameters (and vice versa) to trace how the absence of an instance in the discrim
inator’s training affects the generator’s parameters, and (ii) we propose a nove
l evaluation scheme, in which we assess harmfulness of each training instance on
 the basis of how GAN evaluation metric (e.g., inception score) is expected to c
hange due to the removal of the instance.
We experimentally verified that our influence estimation method correctly inferr
ed the changes in GAN evaluation metrics.
We also demonstrated that the removal of the identified harmful instances effect
ively improved the model’s generative performance with respect to various GAN ev
aluation metrics.

****
On the Neural Tangent Kernel of Equilibrium Models
Zhili Feng,J Zico Kolter
Existing analyses of the neural tangent kernel (NTK) for infinite-depth networks show that the kernel typically becomes degenerate as the number of layers grows. This raises the question of how to apply such methods to practical "infinite depth" architectures such as the recently-proposed deep equilibrium (DEQ) model, which directly computes the infinite-depth limit of a weight-tied network via root-finding. In this work, we show that because of the input injection component of these networks, DEQ models have non-degenerate NTKs even in the infinite depth limit. Furthermore, we show that these kernels themselves can be computed by an analogous root-finding problem as in traditional DEQs, and highlight methods for computing the NTK for both fully-connected and convolutional variants. We evaluate these models empirically, showing they match or improve upon the performance of existing regularized NTK methods.
****
A Coach-Player Framework for Dynamic Team Composition
Bo Liu,qiang liu,Peter Stone,Animesh Garg,Yuke Zhu,Anima Anandkumar
In real-world multi-agent teams, agents with different capabilities may join or leave "on the fly" without altering the team's overarching goals. Coordinating teams with such dynamic composition remains a challenging problem: the optimal team strategy may vary with its composition. Inspired by real-world team sports, we propose a coach-player framework to tackle this problem. We assume that the players only have a partial view of the environment, while the coach has a complete view. The coach coordinates the players by distributing individual strategies. Specifically, we 1) propose an attention mechanism for both the players and the coach; 2) incorporate a variational objective to regularize learning; and 3) design an adaptive communication method to let the coach decide when to communicate with different players. Our attention mechanism on the players and the coach allows for a varying number of heterogeneous agents, and can thus tackle the dynamic team composition. We validate our methods on resource collection tasks in multi-agent particle environment. We demonstrate zero-shot generalization to new team compositions with varying numbers of heterogeneous agents. The performance of our method is comparable or even better than the setting where all players have a full view of the environment, but no coach. Moreover, we see that the performance stays nearly the same even when the coach communicates as little as 13% of the time using our adaptive communication strategy. These results demonstrate the significance of a coach to coordinate players in dynamic teams.
****
Emergent Road Rules In Multi-Agent Driving Environments
Avik Pal,Jonah Philion,Yuan-Hong Liao,Sanja Fidler
For autonomous vehicles to safely share the road with human drivers, autonomous vehicles must abide by specific "road rules" that human drivers have agreed to follow. "Road rules" include rules that drivers are required to follow by law – such as the requirement that vehicles stop at red lights – as well as more subtle social rules – such as the implicit designation of fast lanes on the highway. In this paper, we provide empirical evidence that suggests that – instead of hard-coding road rules into self-driving algorithms – a scalable alternative may be to design multi-agent environments in which road rules emerge as optimal solutions to the problem of maximizing traffic flow. We analyze what ingredients in driving environments cause the emergence of these road rules and find that two crucial factors are noisy perception and agents' spatial density. We provide qualitative and quantitative evidence of the emergence of seven social driving behaviors, ranging from obeying traffic signals to following lanes, all of which emerge from training agents to drive quickly to destinations without colliding. Our results add empirical support for the social road rules that countries worldwide have agreed on for safe, efficient driving.
****
Prior Preference Learning From Experts: Designing A Reward with Active Inference
Jin Young Shin,Cheolhyeong Kim,Hyung Ju Hwang

Active inference may be defined as Bayesian modeling of a brain with a biologica
lly plausible model of the agent. Its primary idea relies on the free energy pri
nciple and the prior preference of the agent. An agent will choose an action tha
t leads to its prior preference for a future observation. In this paper, we clai
m that active inference can be interpreted using reinforcement learning (RL) alg
orithms and find a theoretical connection between them. We extend the concept of
 expected free energy (EFE), which is a core quantity in active inference, and c
laim that EFE can be treated as a negative value function. Motivated by the conc
ept of prior preference and a theoretical connection, we propose a simple but no
vel method for learning a prior preference from experts. This illustrates that t
he problem with RL can be approached with a new perspective of active inference.
 Experimental results of prior preference learning show the possibility of activ
e inference with EFE-based rewards and its application to an inverse RL problem.
****

SSD: A Unified Framework for Self-Supervised Outlier Detection

Vikash Sehwag,Mung Chiang,Prateek Mittal

We ask the following question: what training information is required to design a
n effective outlier/out-of-distribution (OOD) detector, i.e., detecting samples
that lie far away from training distribution? Since unlabeled data is easily acc
essible for many applications, the most compelling approach is to develop detect
ors based on only unlabeled in-distribution data. However, we observe that most
existing detectors based on unlabeled data perform poorly, often equivalent to a
 random prediction. In contrast, existing state-of-the-art OOD detectors achieve
 impressive performance but require access to fine-grained data labels for super
vised training. We propose SSD, an outlier detector based on only unlabeled in-d
istribution data. We use self-supervised representation learning followed by a M
ahalanobis distance based detection in the feature space. We demonstrate that SS
D outperforms most existing detectors based on unlabeled data by a large margin.
 Additionally, SSD even achieves performance on par, and sometimes even better,
with supervised training based detectors. Finally, we expand our detection fram
ework with two key extensions. First, we formulate few-shot OOD detection, in wh
ich the detector has access to only one to five samples from each class of the t
argeted OOD dataset. Second, we extend our framework to incorporate training dat
a labels, if available. We find that our novel detection framework based on SSD
displays enhanced performance with these extensions, and achieves state-of-the-a
rt performance. Our code is publicly available at https://github.com/inspire-gro
up/SSD.
****

ECONOMIC HYPERPARAMETER OPTIMIZATION WITH BLENDED SEARCH STRATEGY

Chi Wang,Qingyun Wu,Silu Huang,Amin Saied

We study the problem of using low cost to search for hyperparameter configuratio
ns in a large search space with heterogeneous evaluation cost and model quality.
We propose a blended search strategy to combine the strengths of global and loca
l search, and prioritize them on the fly with the goal of minimizing the total c
ost spent in finding good configurations. Our approach demonstrates robust perfo
rmance for tuning both tree-based models and deep neural networks on a large Aut
oML benchmark, as well as superior performance in model quality, time, and resou
rce consumption for a production transformer-based NLP model fine-tuning task.
****

Precondition Layer and Its Use for GANs

Tiantian Fang,Alex Schwing,Ruoyu Sun

One of the major challenges when training generative adversarial nets (GANs) is
instability. To address this instability spectral normalization (SN) is remarka
bly successful. However, SN-GAN still suffers from training instabilities, espe
cially when working with higher-dimensional data. We find that those instabilit
ies are accompanied by large condition numbers of the discriminator weight matri
ces. To improve training stability we study common linear-algebra practice and e
mploy preconditioning. Specifically, we introduce a preconditioning layer (
PC-layer)that performs a low-degree polynomial preconditioning. We use this
PC-layer in two ways: 1) fixed preconditioning (FPC) adds a fixed PC-layer

to all layers, and 2) adaptive preconditioning (APC) adaptively controls the strength of preconditioning. Empirically, we show that FPC and APC stabilize the training of un-conditional GANs using classical architectures. On LSUN256×256 data, APC improves FID scores by around 5 points over baselines.
****
When Do Curricula Work?
Xiaoxia Wu,Ethan Dyer,Behnam Neyshabur
Inspired by human learning, researchers have proposed ordering examples during training based on their difficulty. Both curriculum learning, exposing a network to easier examples early in training, and anti-curriculum learning, showing the most difficult examples first, have been suggested as improvements to the standard i.i.d. training. In this work, we set out to investigate the relative benefits of ordered learning. We first investigate the implicit curricula resulting from architectural and optimization bias and find that samples are learned in a highly consistent order. Next, to quantify the benefit of explicit curricula, we conduct extensive experiments over thousands of orderings spanning three kinds of learning: curriculum, anti-curriculum, and random-curriculum -- in which the size of the training dataset is dynamically increased over time, but the examples are randomly ordered. We find that for standard benchmark datasets, curricula have only marginal benefits, and that randomly ordered samples perform as well or better than curricula and anti-curricula, suggesting that any benefit is entirely due to the dynamic training set size. Inspired by common use cases of curriculum learning in practice, we investigate the role of limited training time budget and noisy data in the success of curriculum learning. Our experiments demonstrate that curriculum, but not anti-curriculum or random ordering can indeed improve the performance either with limited training time budget or in the existence of noisy data.
****
Orthogonalizing Convolutional Layers with the Cayley Transform
Asher Trockman,J Zico Kolter
Recent work has highlighted several advantages of enforcing orthogonality in the weight layers of deep networks, such as maintaining the stability of activations, preserving gradient norms, and enhancing adversarial robustness by enforcing low Lipschitz constants. Although numerous methods exist for enforcing the orthogonality of fully-connected layers, those for convolutional layers are more heuristic in nature, often focusing on penalty methods or limited classes of convolutions. In this work, we propose and evaluate an alternative approach to directly parameterize convolutional layers that are constrained to be orthogonal. Specifically, we propose to apply the Cayley transform to a skew-symmetric convolution in the Fourier domain, so that the inverse convolution needed by the Cayley transform can be computed efficiently. We compare our method to previous Lipschitz-constrained and orthogonal convolutional layers and show that it indeed preserves orthogonality to a high degree even for large convolutions. Applied to the problem of certified adversarial robustness, we show that networks incorporating the layer outperform existing deterministic methods for certified defense against ℓ_2-norm-bounded adversaries, while scaling to larger architectures than previously investigated. Code is available at https://github.com/locuslab/orthogonal-convolutions.
****
Inductive Representation Learning in Temporal Networks via Causal Anonymous Walks
Yanbang Wang,Yen-Yu Chang,Yunyu Liu,Jure Leskovec,Pan Li
Temporal networks serve as abstractions of many real-world dynamic systems. These networks typically evolve according to certain laws, such as the law of triadic closure, which is universal in social networks. Inductive representation learning of temporal networks should be able to capture such laws and further be applied to systems that follow the same laws but have not been unseen during the training stage. Previous works in this area depend on either network node identities or rich edge attributes and typically fail to extract these laws. Here, we propose {\em Causal Anonymous Walks (CAWs)} to inductively represent a temporal net

work. CAWs are extracted by temporal random walks and work as automatic retrieva l of temporal network motifs to represent network dynamics while avoiding the ti me-consuming selection and counting of those motifs. CAWs adopt a novel anonymiz ation strategy that replaces node identities with the hitting counts of the node s based on a set of sampled walks to keep the method inductive, and simultaneous ly establish the correlation between motifs. We further propose a neural-network model CAW-N to encode CAWs, and pair it with a CAW sampling strategy with const ant memory and time cost to support online training and inference. CAW-N is eval uated to predict links over 6 real temporal networks and uniformly outperforms p revious SOTA methods by averaged 15\% AUC gain in the inductive setting. CAW-N a lso outperforms previous methods in 5 out of the 6 networks in the transductive setting.
****

ALT-MAS: A Data-Efficient Framework for Active Testing of Machine Learning Algor ithms

Huong Ha,Sunil Gupta,Santu Rana,Svetha Venkatesh

Machine learning models are being used extensively in many important areas, but there is no guarantee that a model will always perform well or as its developers intended. Understanding the correctness of a model is crucial to prevent potent ial failures that may have significant detrimental impact in critical applicatio n areas. In this paper, we propose a novel framework to efficiently test a machi ne learning model using only a small amount of labelled test data. The core idea is to efficiently estimate the metrics of interest for a model-under-test using Bayesian neural network. We develop a methodology to efficiently train the Baye sian neural network from the limited number of labelled data. We also devise an entropy-based sampling strategy to sample the data point such that the proposed framework can give accurate estimations for the metrics of interest. Finally, we conduct an extensive set of experiments to test various machine learning models for different types of metrics. Our experiments with multiple datasets show tha t given a testing budget, the estimation of the metrics by our method is signifi cantly better compared to existing state-of-the-art approaches.
****

Inferring Principal Components in the Simplex with Multinomial Variational Autoe ncoders

James Morton,Justin Silverman,Gleb Tikhonov,Harri Lähdesmäki,Rich Bonneau

Covariance estimation on high-dimensional data is a central challenge across mul tiple scientific disciplines. Sparse high-dimensional count data, frequently enc ountered in biological applications such as DNA sequencing and proteomics, are o ften well modeled using multinomial logistic normal models. In many cases, thes e datasets are also compositional, presented item-wise as fractions of a normali zed total, due to measurement and instrument constraints. In compositional setti ngs, three key factors limit the ability of these models to estimate covariance: (1) the computational complexity of inverting high-dimensional covariance matri ces, (2) the non-exchangeability introduced from the summation constraint on mul tinomial parameters, and (3) the irreducibility of the component multinomial log istic normal distribution that necessitates the use of parameter augmentation, o r similar techniques, during inference. We show that a variational autoencoder a ugmented with a fast isometric log-ratio (ILR) transform can address these issue s and accurately estimate principal components from multinomially logistic norma l distributed data.

This model can be optimized on GPUs and modified to handle mini-batching, with t he ability to scale across thousands of dimensions and thousands of samples.
****

Accelerating DNN Training through Selective Localized Learning

Sarada Krithivasan,Sanchari Sen,Swagath Venkataramani,Anand Raghunathan

Training Deep Neural Networks (DNNs) places immense compute requirements on the underlying hardware platforms, expending large amounts of time and energy. We pr oposeLoCal+SGD, a new algorithmic approach to accelerate DNN train-ing by select ively combining localized or Hebbian learning within a StochasticGradient Descen t (SGD) based training framework. Back-propagation is a computationally expensiv

e process that requires 2 Generalized Matrix Multiply (GEMM)operations to compute the error and weight gradients for each layer. We alleviate this by selectively updating some layers' weights using localized learning rules that require only 1 GEMM operation per layer. Further, since the weight update is performed during the forward pass itself, the layer activations for the mini-batch do not need to be stored until the backward pass, resulting in a reduced memory footprint. Localized updates can substantially boost training speed, but need to be used selectively and judiciously in order to preserve accuracy and convergence. We address this challenge through the design of a Learning Mode Selection Algorithm, where all layers start with SGD, and as epochs progress, layers gradually transition to localized learning. Specifically, for each epoch, the algorithm identifies a Localized→SGDtransition layer, which delineates the network into two regions. Layers before the transition layer use localized updates, while the transition layer and later layers use gradient-based updates. The trend in the weight updates made to the transition layer across epochs is used to determine how the boundary betweenSGD and localized updates is shifted in future epochs. We also propose a low-cost weak supervision mechanism by controlling the learning rate of localized updates based on the overall training loss. We applied LoCal+SGDto 8 image recognition CNNs (including ResNet50 and MobileNetV2) across 3 datasets (Cifar10, Cifar100and ImageNet). Our measurements on a Nvidia GTX 1080Ti GPU demonstrate upto 1.5×improvement in end-to-end training time with~0.5% loss in Top-1classification accuracy.

****

Nonconvex Continual Learning with Episodic Memory

Sungyeob Han,Yeongmo Kim,Jungwoo Lee

Continual learning aims to prevent catastrophic forgetting while learning a new task without accessing data of previously learned tasks.

The memory for such learning scenarios build a small subset of the data for previous tasks and is used in various ways such as quadratic programming and sample selection.

Current memory-based continual learning algorithms are formulated as a constrained optimization problem and rephrase constraints as a gradient-based approach.

However, previous works have not provided the theoretical proof on convergence to previously learned tasks.

In this paper, we propose a theoretical convergence analysis of continual learning based on stochastic gradient descent method.

Our method, nonconvex continual learning (NCCL), can achieve the same convergence rate when the proposed catastrophic forgetting term is suppressed at each iteration.

We also show that memory-based approaches have an inherent problem of overfitting to memory, which degrades the performance on previously learned tasks, namely catastrophic forgetting.

We empirically demonstrate that NCCL successfully performs continual learning with episodic memory by scaling learning rates adaptive to mini-batches on several image classification tasks.

****

Robust Overfitting may be mitigated by properly learned smoothening

Tianlong Chen,Zhenyu Zhang,Sijia Liu,Shiyu Chang,Zhangyang Wang

A recent study (Rice et al., 2020) revealed overfitting to be a dominant phenomenon in adversarially robust training of deep networks, and that appropriate early-stopping of adversarial training (AT) could match the performance gains of most recent algorithmic improvements. This intriguing problem of robust overfitting motivates us to seek more remedies. As a pilot study, this paper investigates two empirical means to inject more learned smoothening during AT: one leveraging knowledge distillation and self-training to smooth the logits, the other performing stochastic weight averaging (Izmailov et al., 2018) to smooth the weights. Despite the embarrassing simplicity, the two approaches are surprisingly effective and hassle-free in mitigating robust overfitting. Experiments demonstrate that by plugging in them to AT, we can simultaneously boost the standard accuracy by $3.72\%\sim6.68\%$ and robust accuracy by $0.22\%\sim2 .03\%$, across multiple

datasets (STL-10, SVHN, CIFAR-10, CIFAR-100, and Tiny ImageNet), perturbation types (ℓ_{∞} and ℓ_2), and robustified methods (PGD, TRADES, and FSGM), establishing the new state-of-the-art bar in AT. We present systematic visualizations and analyses to dive into their possible working mechanisms. We also carefully exclude the possibility of gradient masking by evaluating our models' robustness against transfer attacks. Codes are available at https://github.com/VITA-Group/Alleviate-Robust-Overfitting.

****

Predicting Infectiousness for Proactive Contact Tracing

Yoshua Bengio,Prateek Gupta,Tegan Maharaj,Nasim Rahaman,Martin Weiss,Tristan Deleu,Eilif Benjamin Muller,Meng Qu,victor schmidt,Pierre-Luc St-Charles,hannah alsdurf,Olexa Bilaniuk,david buckeridge,gaetan caron,pierre luc carrier,Joumana Ghosn,satya ortiz gagne,Christopher Pal,Irina Rish,Bernhard Schölkopf,abhinav sharma,Jian Tang,andrew williams

The COVID-19 pandemic has spread rapidly worldwide, overwhelming manual contact tracing in many countries and resulting in widespread lockdowns for emergency containment. Large-scale digital contact tracing (DCT) has emerged as a potential solution to resume economic and social activity while minimizing spread of the virus. Various DCT methods have been proposed, each making trade-offs be-tween privacy, mobility restrictions, and public health. The most common approach, binary contact tracing (BCT), models infection as a binary event, informed only by an individual’s test results, with corresponding binary recommendations that either all or none of the individual’s contacts quarantine. BCT ignores the inherent uncertainty in contacts and the infection process, which could be used to tailor messaging to high-risk individuals, and prompt proactive testing or earlier warnings. It also does not make use of observations such as symptoms or pre-existing medical conditions, which could be used to make more accurate infectiousness predictions. In this paper, we use a recently-proposed COVID-19 epidemiological simulator to develop and test methods that can be deployed to a smartphone to locally and proactively predict an individual’s infectiousness (risk of infecting others) based on their contact history and other information, while respecting strong privacy constraints. Predictions are used to provide personalized recommendations to the individual via an app, as well as to send anonymized messages to the individual’s contacts, who use this information to better predict their own infectiousness, an approach we call proactive contact tracing (PCT). Similarly to other works, we find that compared to no tracing, all DCT methods tested are able to reduce spread of the disease and thus save lives, even at low adoption rates, strongly supporting a role for DCT methods in managing the pandemic. Further, we find a deep-learning based PCT method which improves over BCT for equivalent average mobility, suggesting PCT could help in safe re-opening and second-wave prevention.

****

Local Search Algorithms for Rank-Constrained Convex Optimization

Kyriakos Axiotis,Maxim Sviridenko

We propose greedy and local search algorithms for rank-constrained convex optimization, namely solving $\underset{\mathrm{rank}(A)\leq r^*}{\min}\, R(A)$ given a convex function $R:\mathbb{R}^{m\times n}\rightarrow \mathbb{R}$ and a parameter r^*. These algorithms consist of repeating two steps: (a) adding a new rank-1 matrix to A and (b) enforcing the rank constraint on A. We refine and improve the theoretical analysis of Shalev-Shwartz et al. (2011), and show that if the rank-restricted condition number of R is κ, a solution A with rank $O(r^*\cdot \min\{\kappa \log \frac{R(\mathbf{0})-R(A^*)}{\epsilon}, \kappa^2\})$ and $R(A) \leq R(A^*) + \epsilon$ can be recovered, where A^* is the optimal solution. This significantly generalizes associated results on sparse convex optimization, as well as rank-constrained convex optimization for smooth functions. We then introduce new practical variants of these algorithms that have superior runtime and recover better solutions in practice. We demonstrate the versatility of these methods on a wide range of applications involving matrix completion and robust principal component analysis.

****
Learning Task Decomposition with Ordered Memory Policy Network
Yuchen Lu,Yikang Shen,Siyuan Zhou,Aaron Courville,Joshua B. Tenenbaum,Chuang Gan
Many complex real-world tasks are composed of several levels of subtasks. Humans leverage these hierarchical structures to accelerate the learning process and achieve better generalization. In this work, we study the inductive bias and propose Ordered Memory Policy Network (OMPN) to discover subtask hierarchy by learning from demonstration. The discovered subtask hierarchy could be used to perform task decomposition, recovering the subtask boundaries in an unstructured demonstration. Experiments on Craft and Dial demonstrate that our model can achieve higher task decomposition performance under both unsupervised and weakly supervised settings, comparing with strong baselines. OMPN can also be directly applied to partially observable environments and still achieve higher task decomposition performance. Our visualization further confirms that the subtask hierarchy can emerge in our model 1.
****
Variational saliency maps for explaining model's behavior
Jae Myung Kim,Eunji Kim,Seokhyeon Ha,Sungroh Yoon,Jungwoo Lee
Saliency maps have been widely used to explain the behavior of an image classifier. We introduce a new interpretability method which considers a saliency map as a random variable and aims to calculate the posterior distribution over the saliency map. The likelihood function is designed to measure the distance between the classifier's predictive probability of an image and that of locally perturbed image. For the prior distribution, we make attributions of adjacent pixels have a positive correlation. We use a variational approximation, and show that the approximate posterior is effective in explaining the classifier's behavior. It also has benefits of providing uncertainty over the explanation, giving auxiliary information to experts on how much the explanation is trustworthy.
****
Federated Learning's Blessing: FedAvg has Linear Speedup
Zhaonan Qu,Kaixiang Lin,Zhaojian Li,Jiayu Zhou,Zhengyuan Zhou
Federated learning (FL) learns a model jointly from a set of participating devices without sharing each other's privately held data. The characteristics of non-\textit{i.i.d.} data across the network, low device participation, high communication costs, and the mandate that data remain private bring challenges in understanding the convergence of FL algorithms, particularly in regards to how convergence scales with the number of participating devices. In this paper, we focus on Federated Averaging (FedAvg)--arguably the most popular and effective FL algorithm class in use today--and provide a unified and comprehensive study of its convergence rate. Although FedAvg has recently been studied by an emerging line of literature, it remains open as to how FedAvg's convergence scales with the number of participating devices in the fully heterogeneous FL setting--a crucial question whose answer would shed light on the performance of FedAvg in large FL systems. We fill this gap by providing a unified analysis that establishes convergence guarantees for FedAvg under three classes of problems: strongly convex smooth, convex smooth, and overparameterized strongly convex smooth problems. We show that FedAvg enjoys linear speedup in each case, although with different convergence rates and communication efficiencies. While there have been linear speedup results from distributed optimization that assumes full participation, ours are the first to establish linear speedup for FedAvg under both statistical and system heterogeneity. For strongly convex and convex problems, we also characterize the corresponding convergence rates for the Nesterov accelerated FedAvg algorithm, which are the first linear speedup guarantees for momentum variants of FedAvg in the convex setting. To provably accelerate FedAvg, we design a new momentum-based FL algorithm that further improves the convergence rate in overparameterized linear regression problems. Empirical studies of the algorithms in various settings have supported our theoretical results.
****
Sandwich Batch Normalization
Xinyu Gong,Wuyang Chen,Tianlong Chen,Zhangyang Wang

We present Sandwich Batch Normalization ($\textbf{SaBN}$), a frustratingly easy improvement of Batch Normalization (BN) with only a few lines of code changes. SaBN is motivated by addressing the inherent $\textit{feature distribution heterogeneity}$ that one can be identified in many tasks, which can arise from model heterogeneity (dynamic architectures, model conditioning, etc.), or data heterogeneity (multiple input domains). A SaBN factorizes the BN affine layer into one shared $\textit{sandwich affine}$ layer, cascaded by several parallel $\textit{independent affine}$ layers. Its variants include further decomposing the normalization layer into multiple parallel ones, and extending similar ideas to instance normalization. We demonstrate the prevailing effectiveness of SaBN (as well as its variants) as a $\textbf{drop-in replacement in four tasks}$: neural architecture search (NAS), image generation, adversarial training, and style transfer. Leveraging SaBN immediately boosts two state-of-the-art weight-sharing NAS algorithms significantly on NAS-Bench-201; achieves better Inception Score and FID on CIFAR-10 and ImageNet conditional image generation with three state-of-the art GANs; substantially improves the robust and standard accuracy for adversarial defense; and produces superior arbitrary stylized results. We also provide visualizations and analysis to help understand why SaBN works. All our codes and pre-trained models will be released upon acceptance.

****

Conditional Networks

Anthony Ortiz,Kris Sankaran,Olac Fuentes,Christopher Kiekintveld,Pascal Vincent, Yoshua Bengio,Doina Precup

In this work we tackle the problem of out-of-distribution generalization through conditional computation. Real-world applications often exhibit a larger distributional shift between training and test data than most datasets used in research. On the other hand, training data in such applications often comes with additional annotation. We propose a method for leveraging this extra information by using an auxiliary network that modulates activations of the main network. We show that this approach improves performance over a strong baseline on the Inria Aerial Image Labeling and the Tumor-Infiltrating Lymphocytes (TIL) Datasets, which by design evaluate out-of-distribution generalization in both semantic segmentation and image classification.

****

Noisy Agents: Self-supervised Exploration by Predicting Auditory Events

Chuang Gan,Xiaoyu Chen,Phillip Isola,Antonio Torralba,Joshua B. Tenenbaum

Humans integrate multiple sensory modalities (e.g., visual and audio) to build a causal understanding of the physical world. In this work, we propose a novel type of intrinsic motivation for Reinforcement Learning (RL) that encourages the agent to understand the causal effect of its actions through auditory event prediction. First, we allow the agent to collect a small amount of acoustic data and use K-means to discover underlying auditory event clusters. We then train a neural network to predict the auditory events and use the prediction errors as intrinsic rewards to guide RL exploration. We first conduct an in-depth analysis of our module using a set of Atari games. We then apply our model to audio-visual exploration using the Habitat simulator and active learning using the TDW simulator. Experimental results demonstrate the advantages of using audio signals over vision-based models as intrinsic rewards to guide RL explorations.

****

Exploiting Verified Neural Networks via Floating Point Numerical Error

Kai Jia,Martin Rinard

Motivated by the need to reliably characterize the robustness of deep neural networks, researchers have developed verification algorithms for deep neural networks. Given a neural network, the verifiers aim to answer whether certain properties are guaranteed with respect to all inputs in a space. However, little attention has been paid to floating point numerical error in neural network verification.

We exploit floating point errors in the inference and verification implementations to construct adversarial examples for neural networks that a verifier claims

to be robust with respect to certain inputs. We argue that, to produce sound ver
ification results, any verification system must accurately (or conservatively) m
odel the effects of any float point computations in the network inference or ver
ification system.
****

Approximation Algorithms for Sparse Principal Component Analysis
Agniva Chowdhury,Petros Drineas,David Woodruff,Samson Zhou
Principal component analysis (PCA) is a widely used dimension reduction techniqu
e in machine learning and multivariate statistics. To improve the interpretabili
ty of PCA, various approaches to obtain sparse principal direction loadings have
 been proposed, which are termed Sparse Principal Component Analysis (SPCA). In
this paper, we present three provably accurate, polynomial time, approximation a
lgorithms for the SPCA problem, without imposing any restrictive assumptions on
the input covariance matrix. The first algorithm is based on randomized matrix m
ultiplication; the second algorithm is based on a novel deterministic thresholdi
ng scheme; and the third algorithm is based on a semidefinite programming relaxa
tion of SPCA. All algorithms come with provable guarantees and run in low-degree
 polynomial time. Our empirical evaluations confirm our theoretical findings.
****

VilNMN: A Neural Module Network approach to Video-Grounded Language Tasks
Hung Le,Nancy F. Chen,Steven Hoi
Neural module networks (NMN) have achieved success in image-grounded tasks such
as question answering (QA) on synthetic images. However, very limited work on NM
N has been studied in the video-grounded language tasks. These tasks extend the
complexity of traditional visual tasks with the additional visual temporal varia
nce. Motivated by recent NMN approaches on image-grounded tasks, we introduce V
isio-Linguistic Neural Module Network (VilNMN) to model the information retrieva
l process in video-grounded language tasks as a pipeline of neural modules. VilN
MN first decomposes all language components to explicitly resolves entity refere
nces and detect corresponding action-based inputs from the question. Detected en
tities and actions are used as parameters to instantiate neural module networks
and extract visual cues from the video. Our experiments show that VilNMN can ach
ieve promising performance on two video-grounded language tasks: video QA and vi
deo-grounded dialogues.
****

On Linear Identifiability of Learned Representations
Geoffrey Roeder,Luke Metz,Diederik P Kingma
Identifiability is a desirable property of a statistical model: it implies that
the true model parameters may be estimated to any desired precision, given suffi
cient computational resources and data. We study identifiability in the context
of representation learning: discovering nonlinear data representations that are
optimal with respect to some downstream task. When parameterized as deep neural
networks, such representation functions lack identifiability in parameter space,
 because they are overparameterized by design. In this paper, building on recent
 advances in nonlinear Independent Components Analysis, we aim to rehabilitate i
dentifiability by showing that a large family of discriminative models are in fa
ct identifiable in function space, up to a linear indeterminacy. Many models for
 representation learning in a wide variety of domains have been identifiable in
this sense, including text, images and audio, state-of-the-art at time of public
ation. We derive sufficient conditions for linear identifiability and provide em
pirical support for the result on both simulated and real-world data.
****

Recall Loss for Imbalanced Image Classification and Semantic Segmentation
Junjiao Tian,Niluthpol Chowdhury Mithun,Zachary Seymour,Han-pang Chiu,Zsolt Kira
Class imbalance is a fundamental problem in computer vision applications such as
 semantic segmentation and image classification. Specifically, uneven class dist
ributions in a training dataset often result in unsatisfactory performance on un
der-represented classes. Many works have proposed to weigh the standard cross en
tropy loss function with pre-computed weights based on class statistics such as
the number of samples and class margins. There are two major drawbacks to these

methods: 1) constantly up-weighing minority classes can introduce excessive false positives especially in semantic segmentation; 2) many recent works discovered that pre-computed weights have adversarial effects on representation learning. In this regard, we propose a hard-class mining loss by reshaping the vanilla cross entropy loss such that it weights the loss for each class dynamically based on changing recall performance. We show mathematically that the novel recall loss changes gradually between the standard cross entropy loss and the well-known inverse frequency cross entropy loss and balances precision and accuracy. We first demonstrate that the proposed loss effectively balances precision and accuracy on semantic segmentation datasets, and leads to significant performance improvement over other state-of-the-art loss functions used in semantic segmentation, especially on shallow networks. On image classification, we design a simple two-head training strategy to show that the novel loss function improves representation learning on imbalanced datasets. We outperform the previously best performing method by 5.7% on Place365-LT and by 1.1% on iNaturalist.
****

TwinDNN: A Tale of Two Deep Neural Networks
Hyunmin Jeong,Deming Chen
Compression technologies for deep neural networks (DNNs), such as weight quantization, have been widely investigated to reduce the DNN model size so that they can be implemented on hardware with strict resource restrictions. However, one major downside of model compression is accuracy degradation. To deal with this problem effectively, we propose a new compressed network inference scheme, with a high accuracy but slower DNN coupled with its highly compressed DNN version that typically delivers much faster inference speed but with a lower accuracy. During inference, we determine the confidence of the prediction of the compressed DNN, and infer the original DNN for the inputs that are considered not confident by the compressed DNN. The proposed design can deliver overall accuracy close to the high accuracy model, but with the latency closer to the compressed DNN. We demonstrate our design on two image classification tasks: CIFAR-10 and ImageNet. Our experiments show that our design can recover up to 94% of accuracy drop caused by extreme network compression, with more than 90% increase in throughput compared to just using the original DNN. This is the first work that considers using a highly compressed DNN along with the original DNN in parallel to improve latency significantly while effectively maintaining the original model accuracy.
****

Hidden Markov models are recurrent neural networks: A disease progression modeling application
Matthew Baucum,Anahita Khojandi,Theodore Papamarkou
Hidden Markov models (HMMs) are commonly used for disease progression modeling when the true state of a patient is not fully known. Since HMMs may have multiple local optima, performance can be improved by incorporating additional patient covariates to inform parameter estimation. To allow for this, we formulate a special case of recurrent neural networks (RNNs), which we name hidden Markov recurrent neural networks (HMRNNs), and prove that each HMRNN has the same likelihood function as a corresponding discrete-observation HMM. As a neural network, the HMRNN can also be combined with any other predictive neural networks that take patient covariate information as input. We first show that parameter estimates from HMRNNs are numerically close to those obtained from HMMs via the Baum-Welch algorithm, thus empirically validating their theoretical equivalence. We then demonstrate how the HMRNN can be combined with other neural networks to improve parameter estimation and prediction, using an Alzheimer's disease dataset. The HMRNN yields parameter estimates that improve disease forecasting performance and offer a novel clinical interpretation compared with a standard HMM.
****

VideoFlow: A Framework for Building Visual Analysis Pipelines
Yue Wu,Jianqiang Huang,Jiangjie Zhen,Guokun Wang,Chen Shen,Chang Zhou,Xian-Sheng Hua
The past years have witnessed an explosion of deep learning frameworks like PyTorch and TensorFlow since the success of deep neural networks. These frameworks h

ave significantly facilitated algorithm development in multimedia research and p
roduction. However, how to easily and efficiently build an end-to-end visual ana
lysis pipeline with these algorithms is still an open issue. In most cases, deve
lopers have to spend a huge amount of time tackling data input and output, optim
izing computation efficiency, or even debugging exhausting memory leaks together
 with algorithm development. VideoFlow aims to overcome these challenges by prov
iding a flexible, efficient, extensible, and secure visual analysis framework fo
r both the academia and industry. With VideoFlow, developers can focus on the im
provement of algorithms themselves, as well as the construction of a complete vi
sual analysis workflow. VideoFlow has been incubated in the practices of smart c
ity innovation for more than three years. It has been widely used in tens of int
elligent visual analysis systems. VideoFlow will be open-sourced at \url{https:/
/github.com/xxx/videoflow}.
****

Property Controllable Variational Autoencoder via Invertible Mutual Dependence
Xiaojie Guo,Yuanqi Du,Liang Zhao
Deep generative models have made important progress towards modeling complex, hi
gh dimensional data via learning latent representations. Their usefulness is nev
ertheless often limited by a lack of control over the generative process or a po
or understanding of the latent representation. To overcome these issues, attenti
on is now focused on discovering latent variables correlated to the data propert
ies and ways to manipulate these properties. This paper presents the new Propert
y controllable VAE (PCVAE), where a new Bayesian model is proposed to inductivel
y bias the latent representation using explicit data properties via novel group-
wise and property-wise disentanglement. Each data property corresponds seamlessl
y to a latent variable, by innovatively enforcing invertible mutual dependence b
etween them. This allows us to move along the learned latent dimensions to contr
ol specific properties of the generated data with great precision. Quantitative
and qualitative evaluations confirm that the PCVAE outperforms the existing mode
ls by up to 28% in capturing and 65% in manipulating the desired properties.
****

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 case. Our analysis is based on the stability framework. W
e 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 in
stead of euclidean stability in weights. For neural networks, the measure of dis
tance we consider is invariant to rescaling the weights of each layer. Furtherm
ore, we exploit the notion of on-average stability in order to obtain a data-dep
endent quantity in the bound. This data dependent quantity is seen to be more fa
vorable when training with larger learning rates in our numerical experiments.Th
is might help to shed some light on why larger learning rates can lead to better
 generalization in some practical scenarios.
****

Test-Time Adaptation and Adversarial Robustness
Xi Wu,Yang Guo,Tianqi Li,Jiefeng Chen,Qicheng Lao,Yingyu Liang,Somesh Jha
This paper studies test-time adaptation in the context of adversarial robustness
. We formulate an adversarial threat model for test-time adaptation, where the d
efender may have a unique advantage as the adversarial game becomes a maximin ga
me, instead of a minimax game as in the classic adversarial robustness threat mo
del. We then study whether the maximin threat model admits more ``good solutions
'' than the minimax threat model, and is thus \emph{strictly weaker}. For this p
urpose, we first present a provable separation between the two threat models in
a natural Gaussian data model. For deep learning, while we do not have a proof,
we propose a candidate, Domain Adversarial Neural Networks (${\sf DANN}$), an al
gorithm designed for unsupervised domain adaptation, by showing that it provides
 nontrivial robustness in the test-time maximin threat model against strong tran
sfer attacks and adaptive attacks. This is somewhat surprising since ${\sf DANN}
$ is not designed specifically for adversarial robustness (e.g., against norm-ba

sed attacks), and provides no robustness in the minimax model. Complementing the se results, we show that recent data-oblivious test-time adaptations can be easi ly attacked even with simple transfer attacks. We conclude the paper with variou s future directions of studying adversarially robust test-time adaptation.
****
Deep symbolic regression: Recovering mathematical expressions from data via risk -seeking policy gradients
Brenden K Petersen,Mikel Landajuela Larma,Terrell N. Mundhenk,Claudio Prata Sant iago,Soo Kyung Kim,Joanne Taery Kim
Discovering the underlying mathematical expressions describing a dataset is a co re challenge for artificial intelligence. This is the problem of $\textit{symbol ic regression}$. Despite recent advances in training neural networks to solve co mplex tasks, deep learning approaches to symbolic regression are underexplored. We propose a framework that leverages deep learning for symbolic regression via a simple idea: use a large model to search the space of small models. Specifical ly, we use a recurrent neural network to emit a distribution over tractable math ematical expressions and employ a novel risk-seeking policy gradient to train th e network to generate better-fitting expressions. Our algorithm outperforms seve ral baseline methods (including Eureqa, the gold standard for symbolic regressio n) in its ability to exactly recover symbolic expressions on a series of benchma rk problems, both with and without added noise. More broadly, our contributions include a framework that can be applied to optimize hierarchical, variable-lengt h objects under a black-box performance metric, with the ability to incorporate constraints in situ, and a risk-seeking policy gradient formulation that optimiz es for best-case performance instead of expected performance.
****
Modifying Memories in Transformer Models
Chen Zhu,Ankit Singh Rawat,Manzil Zaheer,Srinadh Bhojanapalli,Daliang Li,Felix Y u,Sanjiv Kumar
Large Transformer models have achieved impressive performance in many natural la nguage tasks. In particular, Transformer based language models have been shown t o have great capabilities in encoding factual knowledge in their vast amount of parameters. While the tasks of improving the memorization and generalization of Transformers have been widely studied, it is not well known how to make transfor mers forget specific old facts and memorize new ones. In this paper, we propose a new task of explicitly modifying specific factual knowledge in Transformer mod els while ensuring the model performance does not degrade on the unmodified fact s. This task is useful in many scenarios, such as updating stale knowledge, prot ecting privacy, and eliminating unintended biases stored in the models. We bench marked several approaches that provide natural baseline performances on this tas k. This leads to the discovery of key components of a Transformer model that are especially effective for knowledge modifications. The work also provides insigh ts into the role that different training phases (such as pretraining and fine-tu ning) play towards memorization and knowledge modification.
****
Grounding Physical Concepts of Objects and Events Through Dynamic Visual Reasoni ng
Zhenfang Chen,Jiayuan Mao,Jiajun Wu,Kwan-Yee Kenneth Wong,Joshua B. Tenenbaum,Ch uang Gan
We study the problem of dynamic visual reasoning on raw videos. This is a challe nging problem; currently, state-of-the-art models often require dense supervisio n on physical object properties and events from simulation, which are impractica l to obtain in real life. In this paper, we present the Dynamic Concept Learner (DCL), a unified framework that grounds physical objects and events from video a nd language. DCL first adopts a trajectory extractor to track each object over t ime and to represent it as a latent, object-centric feature vector. Building upo n this object-centric representation, DCL learns to approximate the dynamic inte raction among objects using graph networks. DCL further incorporates a semantic parser to parse question into semantic programs and, finally, a program executor to run the program to answer the question, levering the learned dynamics model.

After training, DCL can detect and associate objects across the frames, ground visual properties and physical events, understand the causal relationship between events, make future and counterfactual predictions, and leverage these extracted presentations for answering queries. DCL achieves state-of-the-art performance on CLEVRER, a challenging causal video reasoning dataset, even without using ground-truth attributes and collision labels from simulations for training. We further test DCL on a newly proposed video-retrieval and event localization dataset derived from CLEVRER, showing its strong generalization capacity.
****
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 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 a blackbox simulator. 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 empirically test these methods on a popular Atari game (Montezuma's Revenge) and modified versions 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.
****
On Single-environment Extrapolations in Graph Classification and Regression Tasks
Beatrice Bevilacqua,Yangze Zhou,Ryan L Murphy,Bruno Ribeiro
Extrapolation in graph classification/regression remains an underexplored area of an otherwise rapidly developing field. Our work contributes to a growing literature by providing the first systematic counterfactual modeling framework for extrapolations in graph classification/regression tasks. To show that extrapolation from a single training environment is possible, we develop a connection between certain extrapolation tasks on graph sizes and Lovasz's characterization of graph limits. For these extrapolations, standard graph neural networks (GNNs) will fail, while classifiers using induced homomorphism densities succeed, but mostly on unattributed graphs. Generalizing these density features through a GNN subgraph decomposition allows them to also succeed in more complex attributed graph extrapolation tasks. Finally, our experiments validate our theoretical results and showcase some shortcomings of common (interpolation) methods in the literature.
****
Learning a Latent Simplex in Input Sparsity Time
Ainesh Bakshi,Chiranjib Bhattacharyya,Ravi Kannan,David Woodruff,Samson Zhou
We consider the problem of learning a latent k-vertex simplex $K\in\mathbb{R}^d$, given $\mathbf{A}\in\mathbb{R}^{d\times n}$, which can be viewed as n data points that are formed by randomly perturbing some latent points in K, possibly beyond K. A large class of latent variable models, such as adversarial clustering, mixed membership stochastic block models, and topic models can be cast in this view of learning a latent simplex. Bhattacharyya and Kannan (SODA 2020) give an algorithm for learning such a k-vertex latent simplex in time roughly $O(k\cdot\text{nnz}(\mathbf{A}))$, where $\text{nnz}(\mathbf{A})$ is the number of non-zeros in $\mathbf{A}$. We show that the dependence on k in the running time is unnecessary given a natural assumption about the mass of the top k singular values of $\mathbf{A}$, which holds in many of these applications. Further, we show this assumption is necessary, as otherwise an algorithm for learning a latent simplex would imply a better low rank approximation algorithm than what is known.

We obtain a spectral low-rank approximation to $\mathbf{A}$ in input-sparsity time and show that the column space thus obtained has small $\sin\Theta$ (angular) distance to the right top-k singular space of $\mathbf{A}$. Our algorithm then selects k points in the low-rank subspace with the largest inner product (in absolute value) with k carefully chosen random vectors. By working in the low-rank subspace, we avoid reading the entire matrix in each iteration and thus circumvent the $\Theta(k\cdot\text{nnz}(\mathbf{A}))$ running time.

****

Emergent Properties of Foveated Perceptual Systems

Arturo Deza,Talia Konkle

We introduce foveated perceptual systems -- a hybrid architecture inspired by human vision, to explore the role of a \textit{texture-based} foveation stage on the nature and robustness of subsequently learned visual representation in machines. Specifically, these two-stage perceptual systems first foveate an image, inducing a texture-like encoding of peripheral information -- mimicking the effects of \textit{visual crowding} -- which is then relayed through a convolutional neural network (CNN) trained to perform scene categorization. We find that these foveated perceptual systems learn a visual representation that is \textit{distinct} from their non-foveated counterpart through experiments that probe: 1) i.i.d and o.o.d generalization; 2) robustness to occlusion; 3) a center image bias; and 4) high spatial frequency sensitivity. In addition, we examined the impact of this foveation transform with respect to two additional models derived with a rate-distortion optimization procedure to compute matched-resource systems: a lower resolution non-foveated system, and a foveated system with adaptive Gaussian blurring. The properties of greater i.i.d generalization, high spatial frequency sensitivity, and robustness to occlusion emerged exclusively in our foveated texture-based models, independent of network architecture and learning dynamics. Altogether, these results demonstrate that foveation -- via peripheral texture-based computations -- yields a distinct and robust representational format of scene information relative to standard machine vision approaches, and also provides symbiotic computational support that texture-based peripheral encoding has important representational consequences for processing in the human visual system.

****

SACoD: Sensor Algorithm Co-Design Towards Efficient CNN-powered Intelligent PhlatCam

Yonggan Fu,Yang Zhang,Yue Wang,Zhihan Lu,Vivek Boominathan,Ashok Veeraraghavan,Yingyan Lin

There has been a booming demand for integrating Convolutional Neural Networks (CNNs) powered functionalities into Internet-of-Thing (IoT) devices to enable ubiquitous intelligent "IoT cameras". However, more extensive applications of such IoT systems are still limited by two challenges. First, some applications, especially medicine- and wearable-related ones, impose stringent requirements on the camera form factor. Second, powerful CNNs often require considerable storage and energy cost, whereas IoT devices often suffer from limited resources. PhlatCam, with its form factor potentially reduced by orders of magnitude, has emerged as a promising solution to the first aforementioned challenge, while the second one remains a bottleneck. Existing compression techniques, which can potentially tackle the second challenge, are far from realizing the full potential in storage and energy reduction, because they mostly focus on the CNN algorithm itself. To this end, this work proposes SACoD, a Sensor Algorithm Co-Design framework to develop more efficient CNN-powered PhlatCam. In particular, the mask coded in the PhlatCam sensor and the backend CNN model are jointly optimized in terms of both model parameters and architectures via differential neural architecture search. Extensive experiments including both simulation and physical measurement on manufactured masks show that the proposed SACoD framework achieves aggressive model compression and energy savings while maintaining or even boosting the task accuracy, when benchmarking over two state-of-the-art (SOTA) designs with six data sets on four different tasks. We also perform visualization for better understan

ding the superiority of SACoD generated designs. All the codes will be released publicly upon acceptance.
****
gradSim: Differentiable simulation for system identification and visuomotor control
J. Krishna Murthy,Miles Macklin,Florian Golemo,Vikram Voleti,Linda Petrini,Martin Weiss,Breandan Considine,Jérôme Parent-Lévesque,Kevin Xie,Kenny Erleben,Liam Paull,Florian Shkurti,Derek Nowrouzezahrai,Sanja Fidler
In this paper, we tackle the problem of estimating object physical properties such as mass, friction, and elasticity directly from video sequences. Such a system identification problem is fundamentally ill-posed due to the loss of information during image formation. Current best solutions to the problem require precise 3D labels which are labor intensive to gather, and infeasible to create for many systems such as deformable solids or cloth. In this work we present gradSim, a framework that overcomes the dependence on 3D supervision by combining differentiable multiphysics simulation and differentiable rendering to jointly model the evolution of scene dynamics and image formation. This unique combination enables backpropagation from pixels in a video sequence through to the underlying physical attributes that generated them. Furthermore, our unified computation graph across dynamics and rendering engines enables the learning of challenging visuomotor control tasks, without relying on state-based (3D) supervision, while obtaining performance competitive to/better than techniques that require precise 3D labels.
****
A Strong On-Policy Competitor To PPO
Xiangxiang Chu
As a recognized variant and improvement for Trust Region Policy Optimization (TRPO), proximal policy optimization (PPO) has been widely used with several advantages: efficient data utilization, easy implementation and good parallelism. In this paper, a first-order gradient on-policy learning algorithm called Policy Optimization with Penalized Point Probability Distance (POP3D), which is a lower bound to the square of total variance divergence is proposed as another powerful variant. The penalty item has dual effects, prohibiting policy updates from overshooting and encouraging more explorations. Carefully controlled experiments on both discrete and continuous benchmarks verify our approach is highly competitive to PPO.
****
Batch Inverse-Variance Weighting: Deep Heteroscedastic Regression
Vincent Mai,Waleed Khamies,Liam Paull
In model learning, when the training dataset on which the parameters are optimized and the testing dataset on which the model is evaluated are not sampled from identical distributions, we say that the datasets are misaligned. It is well-known that this misalignment can negatively impact model performance. A common source of misalignment is that the inputs are sampled from different distributions. Another source for this misalignment is that the label generating process used to create the training dataset is imperfect. In this work, we consider this setting and additionally assume that the label generating process is able to provide us with a quantity for the role of each label in the misalignment between the datasets, which we consider to be privileged information. Specifically, we consider the task of regression with labels corrupted by heteroscedastic noise and we assume that we have access to an estimate of the variance over each sample. We propose a general approach to include this privileged information in the loss function together with dataset statistics inferred from the mini-batch to mitigate the impact of the dataset misalignment. Subsequently, we propose a specific algorithm for the heteroscedastic regression case, called Batch Inverse-Variance weighting, which adapts inverse-variance weighting for linear regression to the case of neural network function approximation. We demonstrate that this approach achieves a significant improvement in network training performances compared to baselines when confronted with high, input-independent noise.
****

Solving Min-Max Optimization with Hidden Structure via Gradient Descent Ascent
Emmanouil-Vasileios Vlatakis-Gkaragkounis,Lampros Flokas,Georgios Piliouras
Many recent AI architectures are inspired by zero-sum games, however, the behavi
or of their dynamics is still not well understood. Inspired by this, we study st
andard gradient descent ascent (GDA) dynamics in a specific class of non-convex
non-concave zero-sum games, that we call hidden zero-sum games. In this class, p
layers control the inputs of smooth but possibly non-linear functions whose outp
uts are being applied as inputs to a convex-concave game. Unlike general min-max
 games, these games have a well-defined notion of solution; outcomes that implem
ent the von-Neumann equilibrium of the ``hidden convex-concave game. We prove th
at if the hidden game is strictly convex-concave then vanilla GDA converges not
merely to local Nash, but typically to the von-Neumann solution. If the game lac
ks strict convexity properties, GDA may fail to converge to any equilibrium, how
ever, by applying standard regularization techniques we can prove convergence to
 a von-Neumann solution of a slightly perturbed min-max game. Our convergence gu
arantees are non-local, which as far as we know is a first-of-its-kind type of r
esult in non-convex non-concave games. Finally, we discuss connections of our fr
amework with generative adversarial networks.

****
Generative Scene Graph Networks
Fei Deng,Zhuo Zhi,Donghun Lee,Sungjin Ahn
Human perception excels at building compositional hierarchies of parts and objec
ts from unlabeled scenes that help systematic generalization. Yet most work on g
enerative scene modeling either ignores the part-whole relationship or assumes a
ccess to predefined part labels. In this paper, we propose Generative Scene Grap
h Networks (GSGNs), the first deep generative model that learns to discover the
primitive parts and infer the part-whole relationship jointly from multi-object
scenes without supervision and in an end-to-end trainable way. We formulate GSGN
 as a variational autoencoder in which the latent representation is a tree-struc
tured probabilistic scene graph. The leaf nodes in the latent tree correspond to
 primitive parts, and the edges represent the symbolic pose variables required f
or recursively composing the parts into whole objects and then the full scene. T
his allows novel objects and scenes to be generated both by sampling from the pr
ior and by manual configuration of the pose variables, as we do with graphics en
gines. We evaluate GSGN on datasets of scenes containing multiple compositional
objects, including a challenging Compositional CLEVR dataset that we have develo
ped. We show that GSGN is able to infer the latent scene graph, generalize out o
f the training regime, and improve data efficiency in downstream tasks.
****
Adaptive Single-Pass Stochastic Gradient Descent in Input Sparsity Time
Sepideh Mahabadi,David Woodruff,Samson Zhou
We study sampling algorithms for variance reduction methods for stochastic optim
ization. Although stochastic gradient descent (SGD) is widely used for large sca
le machine learning, it sometimes experiences slow convergence rates due to the
high variance from uniform sampling. In this paper, we introduce an algorithm th
at approximately samples a gradient from the optimal distribution for a common f
inite-sum form with n terms, while just making a single pass over the data, us
ing input sparsity time, and $\tO{Td}$ space. Our algorithm can be implemented i
n big data models such as the streaming and distributed models. Moreover, we sho
w that our algorithm can be generalized to approximately sample Hessians and thu
s provides variance reduction for second-order methods as well. We demonstrate t
he efficiency of our algorithm on large-scale datasets.
****
Weights Having Stable Signs Are Important: Finding Primary Subnetworks and Kerne
ls to Compress Binary Weight Networks
Zhaole Sun,Anbang Yao
Binary Weight Networks (BWNs) have significantly lower computational and memory
costs compared to their full-precision counterparts. To address the non-differen
tiable issue of BWNs, existing methods usually use the Straight-Through-Estimato

r (STE). In the optimization, they learn optimal binary weight outputs represent ed as a combination of scaling factors and weight signs to approximate 32-bit fl oating-point weight values, usually with a layer-wise quantization scheme. In th is paper, we begin with an empirical study of training BWNs with STE under the s ettings of using common techniques and tricks. We show that in the context of us ing batch normalization after convolutional layers, adapting scaling factors wit h either hand-crafted or learnable methods brings marginal or no accuracy gain t o final model, while the change of weight signs is crucial in the training of BW Ns. Furthermore, we observe two astonishing training phenomena. Firstly, the tra ining of BWNs demonstrates the process of seeking primary binary sub-networks wh ose weight signs are determined and fixed at the early training stage, which is akin to recent findings on the lottery ticket hypothesis for efficient learning of sparse neural networks. Secondly, we find binary kernels in the convolutional layers of final models tend to be centered on a limited number of the most freq uent binary kernels, showing binary weight networks may has the potential to be further compressed, which breaks the common wisdom that representing each weight with a single bit puts the quantization to the extreme compression. To testify this hypothesis, we additionally propose a binary kernel quantization method, an d we call resulting models Quantized Binary-Kernel Networks (QBNs). We hope thes e new experimental observations would shed new design insights to improve the tr aining and broaden the usages of BWNs.
****

Model Selection for Cross-Lingual Transfer using a Learned Scoring Function
Yang Chen,Alan Ritter
Transformers that are pre-trained on multilingual text corpora, such as, mBERT a nd XLM-RoBERTa, have achieved impressive cross-lingual transfer learning results . In the zero-shot cross-lingual transfer setting, only English training data i s assumed, and the fine-tuned model is evaluated on another target language. No target-language validation data is assumed in this setting, however substantial variance has been observed in target language performance between different fin e-tuning runs. Prior work has relied on English validation/development data to select among models that are fine-tuned with different learning rates, number of steps and other hyperparameters, often resulting in suboptimal choices. In thi s paper, we show that it is possible to select consistently better models when s mall amounts of annotated data are available in an auxiliary pivot language.
We propose a machine learning approach to model selection that uses the fine-tun ed model's own internal representations to predict its cross-lingual capabilitie s. In extensive experiments we find that our approach consistently selects bett er models than English validation data across five languages and five well-studi ed NLP tasks, achieving results that are comparable to small amounts of target l anguage development data.
****

Triple-Search: Differentiable Joint-Search of Networks, Precision, and Accelerat ors
Yonggan Fu,Yongan Zhang,Haoran You,Yingyan Lin
The record-breaking performance and prohibitive complexity of deep neural networ ks (DNNs) have ignited a substantial need for customized DNN accelerators which have the potential to boost DNN acceleration efficiency by orders-of-magnitude. While it has been recognized that maximizing DNNs' acceleration efficiency requi res a joint design/search for three different yet highly coupled aspects, includ ing the networks, adopted precision, and their accelerators, the challenges asso ciated with such a joint search have not yet been fully discussed and addressed. First, to jointly search for a network and its precision via differentiable sea rch, there exists a dilemma of whether to explode the memory consumption or achi eve sub-optimal designs. Second, a generic and differentiable joint search of th e networks and their accelerators is non-trivial due to (1) the discrete nature of the accelerator space and (2) the difficulty of obtaining operation-wise hard ware cost penalties because some accelerator parameters are determined by the wh ole network. To this end, we propose a Triple-Search (TRIPS) framework to addres s the aforementioned challenges towards jointly searching for the network struct

ure, precision, and accelerator in a differentiable manner, to efficiently and e ffectively explore the huge joint search space. Our TRIPS addresses the first ch allenge above via a heterogeneous sampling strategy to achieve unbiased search w ith constant memory consumption, and tackles the latter one using a novel co-sea rch pipeline that integrates a generic differentiable accelerator search engine. Extensive experiments and ablation studies validate that both TRIPS generated n etworks and accelerators consistently outperform state-of-the-art (SOTA) designs (including co-search/exploration techniques, hardware-aware NAS methods, and DN N accelerators), in terms of search time, task accuracy, and accelerator efficie ncy. All codes will be released upon acceptance.
****

Decentralized Attribution of Generative Models
Changhoon Kim,Yi Ren,Yezhou Yang
Growing applications of generative models have led to new threats such as malici ous personation and digital copyright infringement.
One solution to these threats is model attribution, i.e., the identification of user-end models where the contents under question are generated.
Existing studies showed empirical feasibility of attribution through a centraliz ed classifier trained on all existing user-end models.
However, this approach is not scalable in a reality where the number of models e ver grows. Neither does it provide an attributability guarantee.
To this end, this paper studies decentralized attribution, which relies on binar y classifiers associated with each user-end model.
Each binary classifier is parameterized by a user-specific key and distinguishes its associated model distribution from the authentic data distribution.
We develop sufficient conditions of the keys that guarantee an attributability l ower bound.
Our method is validated on MNIST, CelebA, and FFHQ datasets. We also examine the trade-off between generation quality and robustness of attribution against adve rsarial post-processes.
****

Learn Goal-Conditioned Policy with Intrinsic Motivation for Deep Reinforcement L earning
Jinxin Liu,Donglin Wang,Qiangxing Tian,Zhengyu Chen
It is of significance for an agent to learn a widely applicable and general-purp ose policy that can achieve diverse goals including images and text descriptions . Considering such perceptually-specific goals, the frontier of deep reinforcem ent learning research is to learn a goal-conditioned policy without hand-crafted rewards. To learn this kind of policy, recent works usually take as the reward the non-parametric distance to a given goal in an explicit embedding space. From a different viewpoint, we propose a novel unsupervised learning approach named goal-conditioned policy with intrinsic motivation (GPIM), which jointly learns b oth an abstract-level policy and a goal-conditioned policy. The abstract-level p olicy is conditioned on a latent variable to optimize a discriminator and discov ers diverse states that are further rendered into perceptually-specific goals fo r the goal-conditioned policy. The learned discriminator serves as an intrinsic reward function for the goal-conditioned policy to imitate the trajectory induce d by the abstract-level policy. Experiments on various robotic tasks demonstrat e the effectiveness and efficiency of our proposed GPIM method which substantial ly outperforms prior techniques.
****

Flow Neural Network for Traffic Flow Modelling in IP Networks
Xiangle Cheng,Yuchen He,Feifei Long,Shihan Xiao,Fenglin Li
This paper presents and investigates a novel and timely application domain for d eep learning: sub-second traffic flow modelling in IP networks. Traffic flows ar e the most fundamental components in an IP based networking system. The accurate modelling of the generative patterns of these flows is crucial for many practic al network applications. However, the high nonlinearity and dynamics of both the traffic and network conditions make this task challenging, particularly at the time granularity of sub-second. In this paper, we cast this problem as a represe

ntation learning task to model the intricate patterns in data traffic according to the IP network structure and working mechanism. Accordingly, we propose a customized Flow Neural Network, which works in a self-supervised way to extract the domain-specific data correlations. We report the state-of-the-art performances on both synthetic and realistic traffic patterns on multiple practical network applications, which provides a good testament to the strength of our approach.
****

Improved Autoregressive Modeling with Distribution Smoothing
Chenlin Meng,Jiaming Song,Yang Song,Shengjia Zhao,Stefano Ermon
While autoregressive models excel at image compression, their sample quality is often lacking. Although not realistic, generated images often have high likelihood according to the model, resembling the case of adversarial examples. Inspired by a successful adversarial defense method, we incorporate randomized smoothing into autoregressive generative modeling. We first model a smoothed version of the data distribution, and then reverse the smoothing process to recover the original data distribution. This procedure drastically improves the sample quality of existing autoregressive models on several synthetic and real-world image datasets while obtaining competitive likelihoods on synthetic datasets.
****

Pointwise Binary Classification with Pairwise Confidence Comparisons
Lei Feng,Senlin Shu,Nan Lu,Bo Han,Miao Xu,Gang Niu,Bo An,Masashi Sugiyama
Ordinary (pointwise) binary classification aims to learn a binary classifier from pointwise labeled data. However, such pointwise labels may not be directly accessible due to privacy, confidentiality, or security considerations. In this case, can we still learn an accurate binary classifier? This paper proposes a novel setting, namely pairwise comparison (Pcomp) classification, where we are given only pairs of unlabeled data that we know one is more likely to be positive than the other, instead of pointwise labeled data. Compared with pointwise labels, pairwise comparisons are easier to collect, and Pcomp classification is useful for subjective classification tasks. To solve this problem, we present a mathematical formulation for the generation process of pairwise comparison data, based on which we exploit an unbiased risk estimator (URE) to train a binary classifier by empirical risk minimization and establish an estimation error bound. We first prove that a URE can be derived and improve it using correction functions. Then, we start from the noisy-label learning perspective to introduce a progressive URE and improve it by imposing consistency regularization. Finally, experiments validate the effectiveness of our proposed solutions for Pcomp classification.
****

Optimizing Transformers with Approximate Computing for Faster, Smaller and more Accurate NLP Models
Amrit Nagarajan,Sanchari Sen,Jacob R. Stevens,Anand Raghunathan
Transformer models have garnered a lot of interest in recent years by delivering state-of-the-art performance in a range of Natural Language Processing (NLP) tasks. However, these models can have over a hundred billion parameters, presenting very high computational and memory requirements. We address this challenge through Approximate Computing, specifically targeting the use of Transformers in NLP tasks. Transformers are typically pre-trained and subsequently specialized for specific tasks through transfer learning. We observe that pre-trained Transformers are often over-parameterized for several downstream NLP tasks and propose a framework to create smaller and faster models with comparable accuracy. The key cornerstones of the framework are a Significance Analysis (SA) method to identify important components in a pre-trained Transformer for a given task, and techniques to approximate the less significant components. Our framework can be adapted to produce models that are faster, smaller and/or more accurate, depending on the user's constraints. We apply our framework to multiple Transformer models and different downstream tasks, including previously proposed optimized models like DistilBERT and Q8BERT. We demonstrate that our framework produces models that are up to 4$\times$ faster and up to 14$\times$ smaller (with less than 0.5% relative accuracy degradation), or up to 5.5% more accurate with simultaneous model size and speed improvements of up to 9.8$\times$ and 2.9$\times$, respectively.

****
Analysis of Alignment Phenomenon in Simple Teacher-student Networks with Finite Width
Hanlin Zhu,Chengyang Ying,Song Zuo
Recent theoretical analysis suggests that ultra-wide neural networks always converge to global minima near the initialization under first order methods. However, the convergence property of neural networks with finite width could be very different. The simplest experiment with two-layer teacher-student networks shows that the input weights of student neurons eventually align with one of the teacher neurons. This suggests a distinct convergence nature for ``not-too-wide'' neural networks that there might not be any local minima near the initialization. As the theoretical justification, we prove that under the most basic settings, all student neurons must align with the teacher neuron at any local minima. The methodology is extendable to more general cases, where the proof can be reduced to analyzing the properties of a special class of functions that we call {\em Angular Distance (AD) function}. Finally, we demonstrate that these properties can be easily verified numerically.
****
Adaptive Automotive Radar data Acquisition
Madhumitha Sakthi,Ahmed Tewfik
In an autonomous driving scenario, it is vital to acquire and efficiently process data from various sensors to obtain a complete and robust perspective of the surroundings. Many studies have shown the importance of having radar data in addition to images since radar improves object detection performance. We develop a novel algorithm motivated by the hypothesis that with a limited sampling budget, allocating more sampling budget to areas with the object as opposed to a uniform sampling budget ultimately improves relevant object detection and classification. In order to identify the areas with objects, we develop an algorithm to process the object detection results from the Faster R-CNN object detection algorithm and the previous radar frame and use these as prior information to adaptively allocate more bits to areas in the scene that may contain relevant objects. We use previous radar frame information to mitigate the potential information loss of an object missed by the image or the object detection network. Also, in our algorithm, the error of missing relevant information in the current frame due to the limited budget sampling of the previous radar frame did not propagate across frames. We also develop an end-to-end transformer-based 2D object detection network using the NuScenes radar and image data. Finally, we compare the performance of our algorithm against that of standard CS and adaptive CS using radar on the Oxford Radar RobotCar dataset.
****
CPT: Efficient Deep Neural Network Training via Cyclic Precision
Yonggan Fu,Han Guo,Meng Li,Xin Yang,Yining Ding,Vikas Chandra,Yingyan Lin
Low-precision deep neural network (DNN) training has gained tremendous attention as reducing precision is one of the most effective knobs for boosting DNNs' training time/energy efficiency. In this paper, we attempt to explore low-precision training from a new perspective as inspired by recent findings in understanding DNN training: we conjecture that DNNs' precision might have a similar effect as the learning rate during DNN training, and advocate dynamic precision along the training trajectory for further boosting the time/energy efficiency of DNN training. Specifically, we propose Cyclic Precision Training (CPT) to cyclically vary the precision between two boundary values which can be identified using a simple precision range test within the first few training epochs. Extensive simulations and ablation studies on five datasets and eleven models demonstrate that CPT's effectiveness is consistent across various models/tasks (including classification and language modeling). Furthermore, through experiments and visualization we show that CPT helps to (1) converge to a wider minima with a lower generalization error and (2) reduce training variance which we believe opens up a new design knob for simultaneously improving the optimization and efficiency of DNN training.
****

Individually Fair Rankings
Amanda Bower,Hamid Eftekhari,Mikhail Yurochkin,Yuekai Sun
We develop an algorithm to train individually fair learning-to-rank (LTR) models . The proposed approach ensures items from minority groups appear alongside simi lar items from majority groups. This notion of fair ranking is based on the defi nition of individual fairness from supervised learning and is more nuanced than prior fair LTR approaches that simply ensure the ranking model provides underrep resented items with a basic level of exposure. The crux of our method is an opti mal transport-based regularizer that enforces individual fairness and an efficie nt algorithm for optimizing the regularizer. We show that our approach leads to certifiably individually fair LTR models and demonstrate the efficacy of our met hod on ranking tasks subject to demographic biases.
****
Watch-And-Help: A Challenge for Social Perception and Human-AI Collaboration
Xavier Puig,Tianmin Shu,Shuang Li,Zilin Wang,Yuan-Hong Liao,Joshua B. Tenenbaum, Sanja Fidler,Antonio Torralba
In this paper, we introduce Watch-And-Help (WAH), a challenge for testing social intelligence in agents. In WAH, an AI agent needs to help a human-like agent pe rform a complex household task efficiently. To succeed, the AI agent needs to i) understand the underlying goal of the task by watching a single demonstration o f the human-like agent performing the same task (social perception), and ii) coo rdinate with the human-like agent to solve the task in an unseen environment as fast as possible (human-AI collaboration). For this challenge, we build VirtualH ome-Social, a multi-agent household environment, and provide a benchmark includi ng both planning and learning based baselines. We evaluate the performance of AI agents with the human-like agent as well as and with real humans using objectiv e metrics and subjective user ratings. Experimental results demonstrate that our challenge and virtual environment enable a systematic evaluation on the importa nt aspects of machine social intelligence at scale.
****
Interpretable Meta-Reinforcement Learning with Actor-Critic Method
Xingyuan Liang,Xu-Ying Liu
Meta-reinforcement learning (meta-RL) algorithms have successfully trained agent systems to perform well on different tasks within only few updates. However, in gradient-based meta-RL algorithms, the Q-function at adaptation step is mainly estimated by the return of few trajectories, which can lead to high variance in Q-value and biased meta-gradient estimation, and the adaptation uses a large num ber of batched trajectories. To address these challenges, we propose a new meta-RL algorithm that can reduce the variance and bias of the meta-gradient estimati on and perform few-shot task data sampling, which makes the meta-policy more int erpretable. We reformulate the meta-RL objective, and introduce contextual Q-fun ction as a meta-policy critic during task adaptation step and learn the Q-functi on under a soft actor-critic (SAC) framework. The experimental results on 2D nav igation task and meta-RL benchmarks show that our approach can learn an more int erpretable meta-policy to explore unknown environment and the performance are co mparable to previous gradient-based algorithms.
****
Adaptive Federated Optimization
Sashank J. Reddi,Zachary Charles,Manzil Zaheer,Zachary Garrett,Keith Rush,Jakub Kone■ný,Sanjiv Kumar,Hugh Brendan McMahan
Federated learning is a distributed machine learning paradigm in which a large n umber of clients coordinate with a central server to learn a model without shari ng their own training data. Standard federated optimization methods such as Fede rated Averaging (FedAvg) are often difficult to tune and exhibit unfavorable con vergence behavior. In non-federated settings, adaptive optimization methods have had notable success in combating such issues. In this work, we propose federate d versions of adaptive optimizers, including Adagrad, Adam, and Yogi, and analy ze their convergence in the presence of heterogeneous data for general non-conve x settings. Our results highlight the interplay between client heterogeneity and communication efficiency. We also perform extensive experiments on these method

s and show that the use of adaptive optimizers can significantly improve the per formance of federated learning.
****
Unconditional Synthesis of Complex Scenes Using a Semantic Bottleneck
Samaneh Azadi,Michael Tschannen,Eric Tzeng,Sylvain Gelly,Trevor Darrell,Mario Lu cic
Coupling the high-fidelity generation capabilities of label-conditional image sy nthesis methods with the flexibility of unconditional generative models, we prop ose a semantic bottleneck GAN model for unconditional synthesis of complex scene s. We assume pixel-wise segmentation labels are available during training and us e them to learn the scene structure through an unconditional progressive segment ation generation network. During inference, our model first synthesizes a realis tic segmentation layout from scratch, then synthesizes a realistic scene conditi oned on that layout through a conditional segmentation-to-image synthesis networ k. When trained end-to-end, the resulting model outperforms state-of-the-art gen erative models in unsupervised image synthesis on two challenging domains in ter ms of the Frechet Inception Distance and perceptual evaluations. Moreover, we de monstrate that the end-to-end training significantly improves the segmentation-t o-image synthesis sub-network, which results in superior performance over the st ate-of-the-art when conditioning on real segmentation layouts.
****
Near-Optimal Glimpse Sequences for Training Hard Attention Neural Networks
William Harvey,Michael Teng,Frank Wood
Hard visual attention is a promising approach to reduce the computational burden of modern computer vision methodologies. Hard attention mechanisms are typicall y non-differentiable. They can be trained with reinforcement learning but the hi gh-variance training this entails hinders more widespread application. We show h ow hard attention for image classification can be framed as a Bayesian optimal e xperimental design (BOED) problem. From this perspective, the optimal locations to attend to are those which provide the greatest expected reduction in the entr opy of the classification distribution. We introduce methodology from the BOED l iterature to approximate this optimal behaviour, and use it to generate `near-op timal' sequences of attention locations. We then show how to use such sequences to partially supervise, and therefore speed up, the training of a hard attention mechanism. Although generating these sequences is computationally expensive, th ey can be reused by any other networks later trained on the same task.
****
GANs Can Play Lottery Tickets Too
Xuxi Chen,Zhenyu Zhang,Yongduo Sui,Tianlong Chen
Deep generative adversarial networks (GANs) have gained growing popularity in nu merous scenarios, while usually suffer from high parameter complexities for reso urce-constrained real-world applications. However, the compression of GANs has l ess been explored. A few works show that heuristically applying compression tech niques normally leads to unsatisfactory results, due to the notorious training i nstability of GANs. In parallel, the lottery ticket hypothesis shows prevailing success on discriminative models, in locating sparse matching subnetworks capabl e of training in isolation to full model performance. In this work, we for the f irst time study the existence of such trainable matching subnetworks in deep GAN s. For a range of GANs, we certainly find matching subnetworks at 67%-74% sparsity. We observe that with or without pruning discriminator has a minor effe ct on the existence and quality of matching subnetworks, while the initializatio n weights used in the discriminator plays a significant role. We then show the p owerful transferability of these subnetworks to unseen tasks. Furthermore, exten sive experimental results demonstrate that our found subnetworks substantially o utperform previous state-of-the-art GAN compression approaches in both image gen eration (e.g. SNGAN) and image-to-image translation GANs (e.g. CycleGAN). Codes available at https://github.com/VITA-Group/GAN-LTH.
****
Does Adversarial Transferability Indicate Knowledge Transferability?
Kaizhao Liang,Jacky Y. Zhang,Oluwasanmi O Koyejo,Bo Li

Despite the immense success that deep neural networks (DNNs) have achieved, \emph{adversarial examples}, which are perturbed inputs that aim to mislead DNNs to make mistakes, have recently led to great concerns. On the other hand, adversarial examples exhibit interesting phenomena, such as \emph{adversarial transferability}. DNNs also exhibit knowledge transfer, which is critical to improving learning efficiency and learning in domains that lack high-quality training data. To uncover the fundamental connections between these phenomena, we investigate and give an affirmative answer to the question: \emph{does adversarial transferability indicate knowledge transferability?} We theoretically analyze the relationship between adversarial transferability and knowledge transferability, and outline easily checkable sufficient conditions that identify when adversarial transferability indicates knowledge transferability. In particular, we show that composition with an affine function is sufficient to reduce the difference between two models when they possess high adversarial transferability. Furthermore, we provide empirical evaluation for different transfer learning scenarios on diverse datasets, showing a strong positive correlation between the adversarial transferability and knowledge transferability, thus illustrating that our theoretical insights are predictive of practice.

****

Score-Based Generative Modeling through Stochastic Differential Equations

Yang Song,Jascha Sohl-Dickstein,Diederik P Kingma,Abhishek Kumar,Stefano Ermon,Ben Poole

Creating noise from data is easy; creating data from noise is generative modeling. We present a stochastic differential equation (SDE) that smoothly transforms a complex data distribution to a known prior distribution by slowly injecting noise, and a corresponding reverse-time SDE that transforms the prior distribution back into the data distribution by slowly removing the noise.
Crucially, the reverse-time SDE depends only on the time-dependent gradient field (a.k.a., score) of the perturbed data distribution. By leveraging advances in score-based generative modeling, we can accurately estimate these scores with neural networks, and use numerical SDE solvers to generate samples. We show that this framework encapsulates previous approaches in score-based generative modeling and diffusion probabilistic modeling, allowing for new sampling procedures and new modeling capabilities. In particular, we introduce a predictor-corrector framework to correct errors in the evolution of the discretized reverse-time SDE. We also derive an equivalent neural ODE that samples from the same distribution as the SDE, but additionally enables exact likelihood computation, and improved sampling efficiency. In addition, we provide a new way to solve inverse problems with score-based models, as demonstrated with experiments on class-conditional generation, image inpainting, and colorization. Combined with multiple architectural improvements, we achieve record-breaking performance for unconditional image generation on CIFAR-10 with an Inception score of 9.89 and FID of 2.20, a competitive likelihood of 2.99 bits/dim, and demonstrate high fidelity generation of 1024×1024 images for the first time from a score-based generative model.

****

Improving Relational Regularized Autoencoders with Spherical Sliced Fused Gromov Wasserstein

Khai Nguyen,Son Nguyen,Nhat Ho,Tung Pham,Hung Bui

Relational regularized autoencoder (RAE) is a framework to learn the distribution of data by minimizing a reconstruction loss together with a relational regularization on the prior of latent space. A recent attempt to reduce the inner discrepancy between the prior and aggregated posterior distributions is to incorporate sliced fused Gromov-Wasserstein (SFG) between these distributions. That approach has a weakness since it treats every slicing direction similarly, meanwhile several directions are not useful for the discriminative task. To improve the discrepancy and consequently the relational regularization, we propose a new relational discrepancy, named spherical sliced fused Gromov Wasserstein (SSFG), that can find an important area of projections characterized by a von Mises-Fisher distribution. Then, we introduce two variants of SSFG to improve its performance. T

he first variant, named mixture spherical sliced fused Gromov Wasserstein (MSSFG), replaces the vMF distribution by a mixture of von Mises-Fisher distributions to capture multiple important areas of directions that are far from each other. The second variant, named power spherical sliced fused Gromov Wasserstein (PSSFG), replaces the vMF distribution by a power spherical distribution to improve th e sampling time of the vMF distribution in high dimension settings. We then appl y the new discrepancies to the RAE framework to achieve its new variants. Finall y, we conduct extensive experiments to show that the new autoencoders have favor able performance in learning latent manifold structure, image generation, and re construction.
****

Learning Reasoning Paths over Semantic Graphs for Video-grounded Dialogues
Hung Le,Nancy F. Chen,Steven Hoi
Compared to traditional visual question answering, video-grounded dialogues requ ire additional reasoning over dialogue context to answer questions in a multi-tu rn setting. Previous approaches to video-grounded dialogues mostly use dialogue context as a simple text input without modelling the inherent information flows at the turn level. In this paper, we propose a novel framework of Reasoning Path s in Dialogue Context (PDC). PDC model discovers information flows among dialogu e turns through a semantic graph constructed based on lexical components in each question and answer. PDC model then learns to predict reasoning paths over this semantic graph. Our path prediction model predicts a path from the current turn through past dialogue turns that contain additional visual cues to answer the c urrent question. Our reasoning model sequentially processes both visual and text ual information through this reasoning path and the propagated features are used to generate the answer. Our experimental results demonstrate the effectiveness of our method and provide additional insights on how models use semantic depende ncies in a dialogue context to retrieve visual cues.
****

Semi-supervised learning by selective training with pseudo labels via confidence estimation
Masato Ishii
We propose a novel semi-supervised learning (SSL) method that adopts selective t raining with pseudo labels. In our method, we generate hard pseudo-labels and al so estimate their confidence, which represents how likely each pseudo-label is t o be correct. Then, we explicitly select which pseudo-labeled data should be use d to update the model. Specifically, assuming that loss on incorrectly pseudo-la beled data sensitively increase against data augmentation, we select the data co rresponding to relatively small loss after applying data augmentation. The confi dence is used not only for screening candidates of pseudo-labeled data to be sel ected but also for automatically deciding how many pseudo-labeled data should be selected within a mini-batch. Since accurate estimation of the confidence is cr ucial in our method, we also propose a new data augmentation method, called MixC onf, that enables us to obtain confidence-calibrated models even when the number of training data is small. Experimental results with several benchmark datasets validate the advantage of our SSL method as well as MixConf.
****

Fine-Tuning Offline Reinforcement Learning with Model-Based Policy Optimization
Adam Villaflor,John Dolan,Jeff Schneider
In offline reinforcement learning (RL), we attempt to learn a control policy fro m a fixed dataset of environment interactions. This setting has the potential be nefit of allowing us to learn effective policies without needing to collect addi tional interactive data, which can be expensive or dangerous in real-world syste ms. However, traditional off-policy RL methods tend to perform poorly in this se tting due to the distributional shift between the fixed data set and the learned policy. In particular, they tend to extrapolate optimistically and overestimate the action-values outside of the dataset distribution. Recently, two major aven ues have been explored to address this issue. First, behavior-regularized method s that penalize actions that deviate from the demonstrated action distribution. Second, uncertainty-aware model-based (MB) methods that discourage state-actions

where the dynamics are uncertain. In this work, we propose an algorithmic framework that consists of two stages. In the first stage, we train a policy using behavior-regularized model-free RL on the offline dataset. Then, a second stage where we fine-tune the policy using our novel Model-Based Behavior-Regularized Policy Optimization (MB2PO) algorithm. We demonstrate that for certain tasks and dataset distributions our conservative model-based fine-tuning can greatly increase performance and allow the agent to generalize and outperform the demonstrated behavior. We evaluate our method on a variety of the Gym-MuJoCo tasks in the D4RL benchmark and demonstrate that our method is competitive and in some cases superior to the state of the art for most of the evaluated tasks.
****

Representational correlates of hierarchical phrase structure in deep language models
Matteo Alleman,Jonathan Mamou,Miguel A Del Rio,Hanlin Tang,Yoon Kim,SueYeon Chung
While contextual representations from Transformer-based architectures have set a new standard for many NLP tasks, there is not yet a complete accounting of their inner workings. In particular, it is not entirely clear what aspects of sentence-level syntax are captured by these representations, nor how (if at all) they are built along the stacked layers of the network. In this paper, we aim to address such questions with a general class of input perturbation-based analyses of representations from Transformer networks pretrained on self-supervised objectives. Importing from computational and cognitive neuroscience the notion of representational invariance, we perform a series of probes designed to test the sensitivity of Transformer representations to several kinds of structure in sentences. Each probe involves swapping words in a sentence and comparing the representations from perturbed sentences against the original. We experiment with three different perturbations: (1) random permutations of n-grams of varying width, to test the scale at which a representation is sensitive to word position; (2) swapping of two spans which do or do not form a syntactic phrase, to test sensitivity to global phrase structure; and (3) swapping of two adjacent words which do or do not break apart a syntactic phrase, to test sensitivity to local phrase structure. We also connect our probe results to the Transformer architecture by relating the attention mechanism to syntactic distance between two words. Results from the three probes collectively suggest that Transformers build sensitivity to larger parts of the sentence along their layers, and that hierarchical phrase structure plays a role in this process. In particular, sensitivity to local phrase structure increases along deeper layers. Based on our analysis of attention, we show that this is at least partly explained by generally larger attention weights between syntactically distant words.
****

Extreme Memorization via Scale of Initialization
Harsh Mehta,Ashok Cutkosky,Behnam Neyshabur
We construct an experimental setup in which changing the scale of initialization strongly impacts the implicit regularization induced by SGD, interpolating from good generalization performance to completely memorizing the training set while making little progress on the test set. Moreover, we find that the extent and manner in which generalization ability is affected depends on the activation and loss function used, with sin activation being the most extreme. In the case of the homogeneous ReLU activation, we show that this behavior can be attributed to the loss function. Our empirical investigation reveals that increasing the scale of initialization correlates with misalignment of representations and gradients across examples in the same class. This insight allows us to device an alignment measure over gradients and representations which can capture this phenomenon. We demonstrate that our alignment measure correlates with generalization of deep models trained on image classification tasks.
****

Effective Regularization Through Loss-Function Metalearning
Santiago Gonzalez,Risto Miikkulainen
Loss-function metalearning can be used to discover novel, customized loss functi

ons for deep neural networks, resulting in improved performance, faster training , and improved data utilization. A likely explanation is that such functions di scourage overfitting, leading to effective regularization. This paper theoretica lly demonstrates that this is indeed the case: decomposition of learning rules m akes it possible to characterize the training dynamics and show that loss functi ons evolved through TaylorGLO regularize both in the beginning and end of learni ng, and maintain an invariant in between. The invariant can be utilized to make the metalearning process more efficient in practice, and the regularization can train networks that are robust against adversarial attacks. Loss-function optimi zation can thus be seen as a well-founded new aspect of metalearning in neural n etworks.

****

Energy-based View of Retrosynthesis

Ruoxi Sun,Hanjun Dai,Li Li,Steven Kearnes,Bo Dai

Retrosynthesis—the process of identifying a set of reactants to synthesize a tar get molecule—is of vital importance to material design and drug discovery. Exist ing machine learning approaches based on language models and graph neural networ ks have achieved encouraging results. However, the inner connections of these mo dels are rarely discussed, and rigorous evaluations of these models are largely in need. In this paper, we propose a framework that unifies sequence- and graph- based methods as energy-based models (EBMs) with different energy functions. Thi s unified point of view establishes connections between different models and ide ntifies the differences between them, thereby promoting the understanding of mod el design. We also provide a comprehensive assessment of performance to the comm unity. Moreover, we present a novel “dual” variant within the framework that per forms consistent training over Bayesian forward- and backward-prediction by cons training the agreement between the two directions. This model improves the state of the art for template-free approaches where the reaction type is unknown and known.

****

ProGAE: A Geometric Autoencoder-based Generative Model for Disentangling Protei n Conformational Space

Norman Joseph Tatro,Payel Das,Pin-Yu Chen,Vijil Chenthamarakshan,Rongjie Lai

Understanding the protein conformational landscape is critical, as protein funct ion, as well as modulations thereof due to ligand binding or changes in environm ent, are intimately connected with structural variations. This work focuses on l earning a generative neural network on a simulated ensemble of protein structure s obtained using molecular simulation to characterize the distinct structural fl uctuations of a protein bound to various drug molecules. Specifically, we use a geometric autoencoder framework to learn separate latent space encodings of th e intrinsic and extrinsic geometries of the system. For this purpose, the propos ed Protein Geometric AutoEncoder (ProGAE) model is trained on the length of the alpha-carbon pseudobonds and the orientation of the backbone bonds of the prot ein. Using ProGAE latent embeddings, we reconstruct and generate the conformatio nal ensemble of a protein at or near the experimental resolution. Empowered by t he disentangled latent space learning, the intrinsic latent embedding help in ge ometric error correction, whereas the extrinsic latent embedding is successfully used for classification or property prediction of different drugs bound to a s pecific protein. Additionally, ProGAE is able to be transferred to the structure s of a different state of the same protein or to a completely different protein of different size, where only the dense layer decoding from the latent represent ation needs to be retrained. Results show that our geometric learning-based meth od enjoys both accuracy and efficiency for generating complex structural variati ons, charting the path toward scalable and improved approaches for analyzing and enhancing molecular simulations.

****

Searching towards Class-Aware Generators for Conditional Generative Adversarial Networks

Peng Zhou,Lingxi Xie,XIAOPENG ZHANG,Bingbing Ni,Qi Tian

Conditional Generative Adversarial Networks (cGAN) were designed to generate images based on the provided conditions, e.g., class-level distributions. However, existing methods have used the same generating architecture for all classes. This paper presents a novel idea that adopts NAS to find a distinct architecture for each class. The search space contains regular and class-modulated convolutions, where the latter is designed to introduce class-specific information while avoiding the reduction of training data for each class generator. The search algorithm follows a weight-sharing pipeline with mixed-architecture optimization so that the search cost does not grow with the number of classes. To learn the sampling policy, a Markov decision process is embedded into the search algorithm and a moving average is applied for better stability. We evaluate our approach on CIFAR10 and CIFAR100. Besides achieving better image generation quality in terms of FID scores, we discover several insights that are helpful in designing cGAN models.

****

Teaching with Commentaries

Aniruddh Raghu,Maithra Raghu,Simon Kornblith,David Duvenaud,Geoffrey Hinton

Effective training of deep neural networks can be challenging, and there remain many open questions on how to best learn these models. Recently developed methods to improve neural network training examine teaching: providing learned information during the training process to improve downstream model performance. In this paper, we take steps towards extending the scope of teaching. We propose a flexible teaching framework using commentaries, learned meta-information helpful for training on a particular task. We present gradient-based methods to learn commentaries, leveraging recent work on implicit differentiation for scalability. We explore diverse applications of commentaries, from weighting training examples, to parameterising label-dependent data augmentation policies, to representing attention masks that highlight salient image regions. We find that commentaries can improve training speed and/or performance, and provide insights about the dataset and training process. We also observe that commentaries generalise: they can be reused when training new models to obtain performance benefits, suggesting a use-case where commentaries are stored with a dataset and leveraged in future for improved model training.

****

Gradient Vaccine: Investigating and Improving Multi-task Optimization in Massively Multilingual Models

Zirui Wang,Yulia Tsvetkov,Orhan Firat,Yuan Cao

Massively multilingual models subsuming tens or even hundreds of languages pose great challenges to multi-task optimization. While it is a common practice to apply a language-agnostic procedure optimizing a joint multilingual task objective, how to properly characterize and take advantage of its underlying problem structure for improving optimization efficiency remains under-explored. In this paper, we attempt to peek into the black-box of multilingual optimization through the lens of loss function geometry. We find that gradient similarity measured along the optimization trajectory is an important signal, which correlates well with not only language proximity but also the overall model performance. Such observation helps us to identify a critical limitation of existing gradient-based multi-task learning methods, and thus we derive a simple and scalable optimization procedure, named Gradient Vaccine, which encourages more geometrically aligned parameter updates for close tasks. Empirically, our method obtains significant model performance gains on multilingual machine translation and XTREME benchmark tasks for multilingual language models. Our work reveals the importance of properly measuring and utilizing language proximity in multilingual optimization, and has broader implications for multi-task learning beyond multilingual modeling.

****

Feature Integration and Group Transformers for Action Proposal Generation

He-Yen Hsieh,Ding-Jie Chen,Tung-Ying Lee,Tyng-Luh Liu

The task of temporal action proposal generation (TAPG) aims to provide high-quality video segments, i.e., proposals that potentially contain action events. The performance of tackling the TAPG task heavily depends on two key issues, feature

representation and scoring mechanism. To simultaneously take account of both aspects, we introduce an attention-based model, termed as FITS, to address the issues for retrieving high-quality proposals. We first propose a novel Feature-Integration (FI) module to seamlessly fuse two-stream features concerning their interaction to yield a robust video segment representation. We then design a group of Transformer-driven Scorers (TS) to gain the temporal contextual supports over the representations for estimating the starting or ending boundary of an action event. Unlike most previous work to estimate action boundaries without considering the long-range temporal neighborhood, the proposed action-boundary co-estimation mechanism in TS leverages the bi-directional contextual supports for such boundary estimation, which shows the advantage of removing several false-positive boundary predictions. We conduct experiments on two challenging datasets, ActivityNet-1.3 and THUMOS-14. The experimental results demonstrate that the proposed FITS model consistently outperforms state-of-the-art TAPG methods.
****

PlasticineLab: A Soft-Body Manipulation Benchmark with Differentiable Physics
Zhiao Huang,Yuanming Hu,Tao Du,Siyuan Zhou,Hao Su,Joshua B. Tenenbaum,Chuang Gan
Simulated virtual environments serve as one of the main driving forces behind developing and evaluating skill learning algorithms. However, existing environments typically only simulate rigid body physics. Additionally, the simulation process usually does not provide gradients that might be useful for planning and control optimizations. We introduce a new differentiable physics benchmark called PasticineLab, which includes a diverse collection of soft body manipulation tasks. In each task, the agent uses manipulators to deform the plasticine into a desired configuration. The underlying physics engine supports differentiable elastic and plastic deformation using the DiffTaichi system, posing many under-explored challenges to robotic agents. We evaluate several existing reinforcement learning (RL) methods and gradient-based methods on this benchmark. Experimental results suggest that 1) RL-based approaches struggle to solve most of the tasks efficiently; 2) gradient-based approaches, by optimizing open-loop control sequences with the built-in differentiable physics engine, can rapidly find a solution within tens of iterations, but still fall short on multi-stage tasks that require long-term planning. We expect that PlasticineLab will encourage the development of novel algorithms that combine differentiable physics and RL for more complex physics-based skill learning tasks. PlasticineLab will be made publicly available.
****

In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning
Mamshad Nayeem Rizve,Kevin Duarte,Yogesh S Rawat,Mubarak Shah
The recent research in semi-supervised learning (SSL) is mostly dominated by consistency regularization based methods which achieve strong performance. However, they heavily rely on domain-specific data augmentations, which are not easy to generate for all data modalities. Pseudo-labeling (PL) is a general SSL approach that does not have this constraint but performs relatively poorly in its original formulation. We argue that PL underperforms due to the erroneous high confidence predictions from poorly calibrated models; these predictions generate many incorrect pseudo-labels, leading to noisy training. We propose an uncertainty-aware pseudo-label selection (UPS) framework which improves pseudo labeling accuracy by drastically reducing the amount of noise encountered in the training process. Furthermore, UPS generalizes the pseudo-labeling process, allowing for the creation of negative pseudo-labels; these negative pseudo-labels can be used for multi-label classification as well as negative learning to improve the single-label classification. We achieve strong performance when compared to recent SSL methods on the CIFAR-10 and CIFAR-100 datasets. Also, we demonstrate the versatility of our method on the video dataset UCF-101 and the multi-label dataset Pascal VOC.
****

OpenCoS: Contrastive Semi-supervised Learning for Handling Open-set Unlabeled Data

Jongjin Park,Sukmin Yun,Jongheon Jeong,Jinwoo Shin
Modern semi-supervised learning methods conventionally assume both labeled and u
nlabeled data have the same class distribution. However, unlabeled data may incl
ude out-of-class samples in practice; those that cannot have one-hot encoded lab
els from a closed-set of classes in label data, i.e., unlabeled data is an open-
set. In this paper, we introduce OpenCoS, a method for handling this realistic s
emi-supervised learning scenario based on a recent framework of contrastive lear
ning. One of our key findings is that out-of-class samples in the unlabeled data
set can be identified effectively via (unsupervised) contrastive learning. OpenC
oS utilizes this information to overcome the failure modes in the existing state
-of-the-art semi-supervised methods, e.g., ReMixMatch or FixMatch. In particular
, we propose to assign soft-labels for out-of-class samples using the representa
tion learned from contrastive learning. Our extensive experimental results show
the effectiveness of OpenCoS, fixing the state-of-the-art semi-supervised method
s to be suitable for diverse scenarios involving open-set unlabeled data.
****

Tailoring: encoding inductive biases by optimizing unsupervised objectives at pr
ediction time
Ferran Alet,Kenji Kawaguchi,Maria Bauza Villalonga,Nurullah Giray Kuru,Tomas Per
ez,Leslie Pack Kaelbling
From CNNs to attention mechanisms, encoding inductive biases into neural network
s has been a fruitful source of improvement in machine learning. Auxiliary losse
s are a general way of encoding biases in order to help networks learn better re
presentations by adding extra terms to the loss function. However, since they ar
e minimized on the training data, they suffer from the same generalization gap a
s regular task losses. Moreover, by changing the loss function, the network is o
ptimizing a different objective than the one we care about. In this work we solv
e both problems: first, we take inspiration from transductive learning and note
that, after receiving an input but before making a prediction, we can fine-tune
our models on any unsupervised objective. We call this process tailoring, becaus
e we customize the model to each input. Second, we formulate a nested optimizati
on (similar to those in meta-learning) and train our models to perform well on t
he task loss after adapting to the tailoring loss. The advantages of tailoring a
nd meta-tailoring are discussed theoretically and demonstrated empirically on se
veral diverse examples: encoding inductive conservation laws from physics to imp
rove predictions, improving local smoothness to increase robustness to adversari
al examples, and using contrastive losses on the query image to improve generali
zation.
****

Learned Belief Search: Efficiently Improving Policies in Partially Observable Se
ttings
Hengyuan Hu,Adam Lerer,Noam Brown,Jakob Nicolaus Foerster
Search is an important tool for computing effective policies in single- and mult
i-agent environments, and has been crucial for achieving superhuman performance
in several benchmark fully and partially observable games. However, one major li
mitation of prior search approaches for partially observable environments is tha
t the computational cost scales poorly with the amount of hidden information. In
 this paper we present Learned Belief Search (LBS), a computationally efficient
search procedure for partially observable environments. Rather than maintaining
an exact belief distribution, LBS uses an approximate auto-regressive counterfac
tual belief that is learned as a supervised task. In multi-agent settings, LBS u
ses a novel public-private model architecture for underlying policies in order t
o efficiently evaluate these policies during rollouts. In the benchmark domain o
f Hanabi, LBS obtains more than 60% of the benefit of exact search while reducin
g compute requirements by 35×, allowing it to scale to larger settings that were
 inaccessible to previous search methods.
****

GeDi: Generative Discriminator Guided Sequence Generation
Ben Krause,Akhilesh Deepak Gotmare,Bryan McCann,Nitish Shirish Keskar,Shafiq Jot
y,richard socher,Nazneen Rajani

While large-scale language models (LMs) are able to imitate the distribution of natural language well enough to generate realistic text, it is difficult to control which regions of the distribution they generate. This is especially problematic because datasets used for training large LMs usually contain significant toxicity, hate, bias, and negativity. We propose GeDi as an efficient method for using smaller LMs as generative discriminators to guide generation from large LMs to make them safer and more controllable. GeDi guides generation at each step by computing classification probabilities for all possible next tokens via Bayes rule by normalizing over two class-conditional distributions; one conditioned on the desired attribute, or control code, and another conditioned on the undesired attribute, or anti control code. We find that GeDi gives controllability on par with or better than the state of the art method in a variety of settings, while also achieving generation speeds more than 30 times faster. Additionally, training GeDi on only three topics allows us to controllably generate new topics zero-shot from just a keyword. Lastly, we show that GeDi can make GPT-2 and GPT-3 significantly less toxic without sacrificing on linguistic fluency, making it by far the most practical existing method for detoxifying large language models while maintaining a fast generation speed.
****
Uniform Manifold Approximation with Two-phase Optimization
Hyung-Kwon Ko,Jaemin Jo,Yung-Kyun Noh,Jinwook Seo
We present a dimensionality reduction algorithm called Uniform Manifold Approximation with Two-phase Optimization (UMATO) which produces less biased global structures in the embedding results and is robust over diverse initialization methods than previous methods such as t-SNE and UMAP. We divide the optimization into two phases to alleviate the bias by establishing the global structure early using the representatives of the high-dimensional structures. The phases are 1) global optimization to obtain the overall skeleton of data and 2) local optimization to identify the regional characteristics of local areas. In our experiments with one synthetic and three real-world datasets, UMATO outperformed widely-used baseline algorithms, such as PCA, red, t-SNE, UMAP, topological autoencoders and Anchor t-SNE, in terms of quality metrics and 2D projection results.
****
Into the Wild with AudioScope: Unsupervised Audio-Visual Separation of On-Screen Sounds
Efthymios Tzinis,Scott Wisdom,Aren Jansen,Shawn Hershey,Tal Remez,Dan Ellis,John R. Hershey
Recent progress in deep learning has enabled many advances in sound separation and visual scene understanding. However, extracting sound sources which are apparent in natural videos remains an open problem. In this work, we present AudioScope, a novel audio-visual sound separation framework that can be trained without supervision to isolate on-screen sound sources from real in-the-wild videos. Prior audio-visual separation work assumed artificial limitations on the domain of sound classes (e.g., to speech or music), constrained the number of sources, and required strong sound separation or visual segmentation labels. AudioScope overcomes these limitations, operating on an open domain of sounds, with variable numbers of sources, and without labels or prior visual segmentation. The training procedure for AudioScope uses mixture invariant training (MixIT) to separate synthetic mixtures of mixtures (MoMs) into individual sources, where noisy labels for mixtures are provided by an unsupervised audio-visual coincidence model. Using the noisy labels, along with attention between video and audio features, AudioScope learns to identify audio-visual similarity and to suppress off-screen sounds. We demonstrate the effectiveness of our approach using a dataset of video clips extracted from open-domain YFCC100m video data. This dataset contains a wide diversity of sound classes recorded in unconstrained conditions, making the application of previous methods unsuitable. For evaluation and semi-supervised experiments, we collected human labels for presence of on-screen and off-screen sounds on a small subset of clips.
****
Smooth Activations and Reproducibility in Deep Networks

Gil I Shamir,Dong Lin,Lorenzo Coviello
Deep networks are gradually penetrating almost every domain in our lives due to their amazing success. However, with substantive performance accuracy improvements comes the price of irreproducibility. Two identical models, trained on the exact same training dataset may exhibit large differences in predictions on individual examples even when average accuracy is similar, especially when trained on highly distributed parallel systems. The popular Rectified Linear Unit (ReLU) activation has been key to recent success of deep networks. We demonstrate, however, that ReLU is also a catalyzer to irreproducibility in deep networks. We show that not only can activations smoother than ReLU provide better accuracy, but they can also provide better accuracy-reproducibility tradeoffs. We propose a new family of activations; Smooth ReLU (SmeLU), designed to give such better tradeoffs, while also keeping the mathematical expression simple, and thus implementation cheap. SmeLU is monotonic, mimics ReLU, while providing continuous gradients, yielding better reproducibility. We generalize SmeLU to give even more flexibility and then demonstrate that SmeLU and its generalized form are special cases of a more general methodology of REctified Smooth Continuous Unit (RESCU) activations.
Empirical results demonstrate the superior accuracy-reproducibility tradeoffs with smooth activations, SmeLU in particular.
****

DiP Benchmark Tests: Evaluation Benchmarks for Discourse Phenomena in MT
Prathyusha Jwalapuram,Barbara Rychalska,Shafiq Joty,Dominika Basaj
Despite increasing instances of machine translation (MT) systems including extrasentential context information, the evidence for translation quality improvement is sparse, especially for discourse phenomena. Popular metrics like BLEU are not expressive or sensitive enough to capture quality improvements or drops that are minor in size but significant in perception. We introduce the first of their kind MT benchmark testsets that aim to track and hail improvements across four main discourse phenomena: anaphora, lexical consistency, coherence and readability, and discourse connective translation. We also introduce evaluation methods for these tasks, and evaluate several competitive baseline MT systems on the curated datasets. Surprisingly, we find that the complex context-aware models that we test do not improve discourse-related translations consistently across languages and phenomena. Our evaluation benchmark is available as a leaderboard at <dipbenchmark1.github.io>.
****

Cut out the annotator, keep the cutout: better segmentation with weak supervision
Sarah Hooper,Michael Wornow,Ying Hang Seah,Peter Kellman,Hui Xue,Frederic Sala,Curtis Langlotz,Christopher Re
Constructing large, labeled training datasets for segmentation models is an expensive and labor-intensive process. This is a common challenge in machine learning, addressed by methods that require few or no labeled data points such as few-shot learning (FSL) and weakly-supervised learning (WS). Such techniques, however, have limitations when applied to image segmentation---FSL methods often produce noisy results and are strongly dependent on which few datapoints are labeled, while WS models struggle to fully exploit rich image information. We propose a framework that fuses FSL and WS for segmentation tasks, enabling users to train high-performing segmentation networks with very few hand-labeled training points. We use FSL models as weak sources in a WS framework, requiring a very small set of reference labeled images, and introduce a new WS model that focuses on key areas---areas with contention among noisy labels---of the image to fuse these weak sources. Empirically, we evaluate our proposed approach over seven well-motivated segmentation tasks. We show that our methods can achieve within 1.4 Dice points compared to fully supervised networks while only requiring five hand-labeled training points. Compared to existing FSL methods, our approach improves performance by a mean 3.6 Dice points over the next-best method.
****

Globetrotter: Unsupervised Multilingual Translation from Visual Alignment

Didac Suris Coll-Vinent,Dave Epstein,Carl Vondrick
Machine translation in a multi-language scenario requires large-scale parallel corpora for every language pair. Unsupervised translation is challenging because there is no explicit connection between languages, and the existing methods have to rely on topological properties of the language representations. We introduce a framework that leverages visual similarity to align multiple languages, using images as the bridge between them. We estimate the cross-modal alignment between language and images, and use this estimate to guide the learning of cross-lingual representations. Our language representations are trained jointly in one model with a single stage. Experiments with fifty-two languages show that our method outperforms prior work on unsupervised word-level and sentence-level translation using retrieval.
****
Measuring Visual Generalization in Continuous Control from Pixels
Jake Grigsby,Yanjun Jane Qi
Self-supervised learning and data augmentation have significantly reduced the performance gap between state and image-based reinforcement learning agents in continuous control tasks. However, it is still unclear whether current techniques can face the variety of visual conditions required by real-world environments. We propose a challenging benchmark that tests agents' visual generalization by adding graphical variety to existing continuous control domains. Our empirical analysis shows that current methods struggle to generalize across a diverse set of visual changes, and we examine the specific factors of variation that make these tasks difficult. We find that data augmentation techniques outperform self-supervised learning approaches and that more significant image transformations provide better visual generalization.
****
CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding
Yanru Qu,Dinghan Shen,Yelong Shen,Sandra Sajeev,Weizhu Chen,Jiawei Han
Data augmentation has been demonstrated as an effective strategy for improving model generalization and data efficiency. However, due to the discrete nature of natural language, designing label-preserving transformations for text data tends to be more challenging. In this paper, we propose a novel data augmentation frame-work dubbed CoDA, which synthesizes diverse and informative augmented examples by integrating multiple transformations organically. Moreover, a contrastive regularization is introduced to capture the global relationship among all the data samples. A momentum encoder along with a memory bank is further leveraged to better estimate the contrastive loss. To verify the effectiveness of the proposed framework, we apply CoDA to Transformer-based models on a wide range of natural language understanding tasks. On the GLUE benchmark, CoDA gives rise to an average improvement of 2.2%while applied to the Roberta-large model. More importantly, it consistently exhibits stronger results relative to several competitive data augmentation and adversarial training baselines (including the low-resource settings). Extensive experiments show that the proposed contrastive objective can be flexibly combined with various data augmentation approaches to further boost their performance, highlighting the wide applicability of the CoDA framework.
****
Global Convergence of Three-layer Neural Networks in the Mean Field Regime
Huy Tuan Pham,Phan-Minh Nguyen
In the mean field regime, neural networks are appropriately scaled so that as the width tends to infinity, the learning dynamics tends to a nonlinear and nontrivial dynamical limit, known as the mean field limit. This lends a way to study large-width neural networks via analyzing the mean field limit. Recent works have successfully applied such analysis to two-layer networks and provided global convergence guarantees. The extension to multilayer ones however has been a highly challenging puzzle, and little is known about the optimization efficiency in the mean field regime when there are more than two layers.

In this work, we prove a global convergence result for unregularized feedforward

three-layer networks in the mean field regime. We first develop a rigorous framework to establish the mean field limit of three-layer networks under stochastic gradient descent training. To that end, we propose the idea of a neuronal embedding, which comprises of a fixed probability space that encapsulates neural networks of arbitrary sizes. The identified mean field limit is then used to prove a global convergence guarantee under suitable regularity and convergence mode assumptions, which – unlike previous works on two-layer networks – does not rely critically on convexity. Underlying the result is a universal approximation property, natural of neural networks, which importantly is shown to hold at any finite training time (not necessarily at convergence) via an algebraic topology argument.

****

Provable Memorization via Deep Neural Networks using Sub-linear Parameters

Sejun Park,Jaeho Lee,Chulhee Yun,Jinwoo Shin

It is known that $\Theta(N)$ parameters are sufficient for neural networks to memorize arbitrary N input-label pairs. By exploiting depth, we show that $\Theta(N^{2/3})$ parameters suffice to memorize N pairs, under a mild condition on the separation of input points. In particular, deeper networks (even with width 3) are shown to memorize more pairs than shallow networks, which also agrees with the recent line of works on the benefits of depth for function approximation. We also provide empirical results that support our theoretical findings.

****

TOMA: Topological Map Abstraction for Reinforcement Learning

Zhao Heng Yin,Wu-Jun Li

Animals are able to discover the topological map (graph) of surrounding environment, which will be used for navigation. Inspired by this biological phenomenon, researchers have recently proposed to learn a graph representation for Markov decision process (MDP) and use such graphs for planning in reinforcement learning (RL). However, existing learning-based graph generation methods suffer from many drawbacks. One drawback is that existing methods do not learn an abstraction for graphs, which results in high memory and computation cost. This drawback also makes generated graph non-robust, which degrades the planning performance. Another drawback is that existing methods cannot be used for facilitating exploration which is important in RL. In this paper, we propose a new method, called topological map abstraction (TOMA), for graph generation. TOMA can learn an abstract graph representation for MDP, which costs much less memory and computation cost than existing methods. Furthermore, TOMA can be used for facilitating exploration. In particular, we propose planning to explore, in which TOMA is used to accelerate exploration by guiding the agent towards unexplored states. A novel experience replay module called vertex memory is also proposed to improve exploration performance. Experimental results show that TOMA can outperform existing methods to achieve the state-of-the-art performance.

****

Ensembles of Generative Adversarial Networks for Disconnected Data

Lorenzo Luzi,Randall Balestriero,Richard Baraniuk

Most computer vision datasets are composed of disconnected sets, such as images of different objects. We prove that distributions of this type of data cannot be represented with a continuous generative network without error, independent of the learning algorithm used. Disconnected datasets can be represented in two ways: with an ensemble of networks or with a single network using a truncated latent space. We show that ensembles are more desirable than truncated distributions for several theoretical and computational reasons. We construct a regularized optimization problem that rigorously establishes the relationships between a single continuous GAN, an ensemble of GANs, conditional GANs, and Gaussian Mixture GANs. The regularization can be computed efficiently, and we show empirically that our framework has a performance sweet spot that can be found via hyperparameter tuning. The ensemble framework provides better performance than a single continuous GAN or cGAN while maintaining fewer total parameters.

****

Deep Learning meets Projective Clustering

Alaa Maalouf,Harry Lang,Daniela Rus,Dan Feldman
A common approach for compressing Natural Language Processing (NLP) networks is to encode the embedding layer as a matrix $A\in\mathbb{R}^{n\times d}$, compute its rank-$j$ approximation A_j via SVD (Singular Value Decomposition), and then factor A_j into a pair of matrices that correspond to smaller fully-connected layers to replace the original embedding layer. Geometrically, the rows of A represent points in $\mathbb{R}^d$, and the rows of A_j represent their projections onto the j-dimensional subspace that minimizes the sum of squared distances (``errors'') to the points.
In practice, these rows of A may be spread around $k>1$ subspaces, so factoring A based on a single subspace may lead to large errors that turn into large drops in accuracy.

Inspired by \emph{projective clustering} from computational geometry, we suggest replacing this subspace by a set of k subspaces, each of dimension j, that minimizes the sum of squared distances over every point (row in A) to its \emph{closest} subspace. Based on this approach, we provide a novel architecture that replaces the original embedding layer by a set of k small layers that operate in parallel and are then recombined with a single fully-connected layer.

Extensive experimental results on the GLUE benchmark yield networks that are both more accurate and smaller compared to the standard matrix factorization (SVD). For example, we further compress DistilBERT by reducing the size of the embedding layer by 40% while incurring only a 0.5% average drop in accuracy over all nine GLUE tasks, compared to a 2.8% drop using the existing SVD approach. On RoBERTa we achieve 43% compression of the embedding layer with less than a 0.8% average drop in accuracy as compared to a 3% drop previously.
****
Learning to Deceive Knowledge Graph Augmented Models via Targeted Perturbation
Mrigank Raman,Aaron Chan,Siddhant Agarwal,PeiFeng Wang,Hansen Wang,Sungchul Kim,Ryan Rossi,Handong Zhao,Nedim Lipka,Xiang Ren
Knowledge graphs (KGs) have helped neural models improve performance on various knowledge-intensive tasks, like question answering and item recommendation. By using attention over the KG, such KG-augmented models can also "explain" which KG information was most relevant for making a given prediction. In this paper, we question whether these models are really behaving as we expect. We show that, through a reinforcement learning policy (or even simple heuristics), one can produce deceptively perturbed KGs, which maintain the downstream performance of the original KG while significantly deviating from the original KG's semantics and structure. Our findings raise doubts about KG-augmented models' ability to reason about KG information and give sensible explanations.
****
Knowledge Distillation as Semiparametric Inference
Tri Dao,Govinda M Kamath,Vasilis Syrgkanis,Lester Mackey
A popular approach to model compression is to train an inexpensive student model to mimic the class probabilities of a highly accurate but cumbersome teacher model. Surprisingly, this two-step knowledge distillation process often leads to higher accuracy than training the student directly on labeled data. To explain and enhance this phenomenon, we cast knowledge distillation as a semiparametric inference problem with the optimal student model as the target, the unknown Bayes class probabilities as nuisance, and the teacher probabilities as a plug-in nuisance estimate. By adapting modern semiparametric tools, we derive new guarantees for the prediction error of standard distillation and develop two enhancements—cross-fitting and loss correction—to mitigate the impact of teacher overfitting and underfitting on student performance. We validate our findings empirically on both tabular and image data and observe consistent improvements from our knowledge distillation enhancements.
****
It's Hard for Neural Networks to Learn the Game of Life
Jacob M. Springer,Garrett T. Kenyon

Efforts to improve the learning abilities of neural networks have focused mostly on the role of optimization methods rather than on weight initializations. Recent findings, however, suggest that neural networks rely on lucky random initial weights of subnetworks called "lottery tickets" that converge quickly to a solution. To investigate how weight initializations affect performance, we examine small convolutional networks that are trained to predict n steps of the two-dimensional cellular automaton Conway’s Game of Life, the update rules of which can be implemented efficiently in a small CNN. We find that networks of this architecture trained on this task rarely converge. Rather, networks require substantially more parameters to consistently converge. Furthermore, we find that the initialization parameters that gradient descent converges to a solution are sensitive to small perturbations, such as a single sign change. Finally, we observe a critical value d_0 such that training minimal networks with examples in which cells are alive with probability d_0 dramatically increases the chance of convergence to a solution. Our results are consistent with the lottery ticket hypothesis.

****

Meta-Learning with Neural Tangent Kernels

Yufan Zhou,Zhenyi Wang,Jiayi Xian,Changyou Chen,Jinhui Xu

Model Agnostic Meta-Learning (MAML) has emerged as a standard framework for meta-learning, where a meta-model is learned with the ability of fast adapting to new tasks. However, as a double-looped optimization problem, MAML needs to differentiate through the whole inner-loop optimization path for every outer-loop training step, which may lead to both computational inefficiency and sub-optimal solutions. In this paper, we generalize MAML to allow meta-learning to be defined in function spaces, and propose the first meta-learning paradigm in the Reproducing Kernel Hilbert Space (RKHS) induced by the meta-model's Neural Tangent Kernel (NTK). Within this paradigm, we introduce two meta-learning algorithms in the RKHS, which no longer need a sub-optimal iterative inner-loop adaptation as in the MAML framework. We achieve this goal by 1) replacing the adaptation with a fast-adaptive regularizer in the RKHS; and 2) solving the adaptation analytically based on the NTK theory. Extensive experimental studies demonstrate advantages of our paradigm in both efficiency and quality of solutions compared to related meta-learning algorithms. Another interesting feature of our proposed methods is that they are demonstrated to be more robust to adversarial attacks and out-of-distribution adaptation than popular baselines, as demonstrated in our experiments.

****

Vulnerability-Aware Poisoning Mechanism for Online RL with Unknown Dynamics

Yanchao Sun,Da Huo,Furong Huang

Poisoning attacks on Reinforcement Learning (RL) systems could take advantage of RL algorithm’s vulnerabilities and cause failure of the learning. However, prior works on poisoning RL usually either unrealistically assume the attacker knows the underlying Markov Decision Process (MDP), or directly apply the poisoning methods in supervised learning to RL. In this work, we build a generic poisoning framework for online RL via a comprehensive investigation of heterogeneous poisoning models in RL. Without any prior knowledge of the MDP, we propose a strategic poisoning algorithm called Vulnerability-Aware Adversarial Critic Poison (VA2C-P), which works for on-policy deep RL agents, closing the gap that no poisoning method exists for policy-based RL agents. VA2C-P uses a novel metric, stability radius in RL, that measures the vulnerability of RL algorithms. Experiments on multiple deep RL agents and multiple environments show that our poisoning algorithm successfully prevents agents from learning a good policy or teaches the agents to converge to a target policy, with a limited attacking budget.

****

Understanding and Improving Lexical Choice in Non-Autoregressive Translation

Liang Ding,Longyue Wang,Xuebo Liu,Derek F. Wong,Dacheng Tao,Zhaopeng Tu

Knowledge distillation (KD) is essential for training non-autoregressive translation (NAT) models by reducing the complexity of the raw data with an autoregressive teacher model. In this study, we empirically show that as a side effect of this training, the lexical choice errors on low-frequency words are propagated to

the NAT model from the teacher model. To alleviate this problem, we propose to expose the raw data to NAT models to restore the useful information of low-frequency words, which are missed in the distilled data. To this end, we introduce an extra Kullback-Leibler divergence term derived by comparing the lexical choice of NAT model and that embedded in the raw data. Experimental results across language pairs and model architectures demonstrate the effectiveness and universality of the proposed approach. Extensive analyses confirm our claim that our approach improves performance by reducing the lexical choice errors on low-frequency words. Encouragingly, our approach pushes the SOTA NAT performance on the WMT14 English-German and WMT16 Romanian-English datasets up to 27.8 and 33.8 BLEU points, respectively.

****

One Vertex Attack on Graph Neural Networks-based Spatiotemporal Forecasting

Fuqiang Liu,Luis Miranda Moreno,Lijun Sun

Spatiotemporal forecasting plays an essential role in intelligent transportation systems (ITS) and numerous applications, such as route planning, navigation, and automatic driving. Deep Spatiotemporal Graph Neural Networks, which capture both spatial and temporal patterns, have achieved great success in traffic forecasting applications. Though Deep Neural Networks (DNNs) have been proven to be vulnerable to carefully designed perturbations in multiple domains like objection classification and graph classification, these adversarial works cannot be directly applied to spatiotemporal GNNs because of their causality and spatiotemporal mechanism. There is still a lack of studies on the vulnerability and robustness of spatiotemporal GNNs. Particularly, if spatiotemporal GNNs are vulnerable in real-world traffic applications, a hacker can easily cause serious traffic congestion and even a city-scale breakdown. To fill this gap, we design One Vertex Attack to break deep spatiotemporal GNNs by attacking a single one vertex. To achieve this, we apply the genetic algorithm with a universal attack method as the evaluation function to locate the weakest vertex; then perturbations are generated by solving an optimization problem with the inverse estimation. Empirical studies prove that perturbations in one vertex can be diffused into most of the graph when spatiotemporal GNNs are under One Vertex Attack.

****

Rethinking Architecture Selection in Differentiable NAS

Ruochen Wang,Minhao Cheng,Xiangning Chen,Xiaocheng Tang,Cho-Jui Hsieh

Differentiable Neural Architecture Search is one of the most popular Neural Architecture Search (NAS) methods for its search efficiency and simplicity, accomplished by jointly optimizing the model weight and architecture parameters in a weight-sharing supernet via gradient-based algorithms. At the end of the search phase, the operations with the largest architecture parameters will be selected to form the final architecture, with the implicit assumption that the values of architecture parameters reflect the operation strength. While much has been discussed about the supernet's optimization, the architecture selection process has received little attention. We provide empirical and theoretical analysis to show that the magnitude of architecture parameters does not necessarily indicate how much the operation contributes to the supernet's performance. We propose an alternative perturbation-based architecture selection that directly measures each operation's influence on the supernet. We re-evaluate several differentiable NAS methods with the proposed architecture selection and find that it is able to extract significantly improved architectures from the underlying supernets consistently. Furthermore, we find that several failure modes of DARTS can be greatly alleviated with the proposed selection method, indicating that much of the poor generalization observed in DARTS can be attributed to the failure of magnitude-based architecture selection rather than entirely the optimization of its supernet.

****

Anti-Distillation: Improving Reproducibility of Deep Networks

Gil I Shamir,Lorenzo Coviello

Deep networks have been revolutionary in improving performance of machine learning and artificial intelligence systems.

Their high prediction accuracy, however, comes at a price of model irreproducibi

lity in very high levels that do not occur with classical linear models. Two models, even if they are supposedly identical, with identical architecture and identical trained parameter sets, and that are trained on the same set of training examples, while possibly providing identical average prediction accuracies, may predict very differently on individual, previously unseen, examples. Prediction differences may be as large as the order of magnitude of the predictions themselves. Ensembles have been shown to somewhat mitigate this behavior, but without an extra push, may not be utilizing their full potential. In this work, a novel approach, Anti-Distillation, is proposed to address irreproducibility in deep networks, where ensemble models are used to generate predictions. Anti-Distillation forces ensemble components away from one another by techniques like de-correlating their outputs over mini-batches of examples, forcing them to become even more different and more diverse. Doing so enhances the benefit of ensembles, making the final predictions more reproducible. Empirical results demonstrate substantial prediction difference reductions achieved by Anti-Distillation on benchmark and real datasets.
****

Distributional Sliced-Wasserstein and Applications to Generative Modeling
Khai Nguyen,Nhat Ho,Tung Pham,Hung Bui
Sliced-Wasserstein distance (SW) and its variant, Max Sliced-Wasserstein distance (Max-SW), have been used widely in the recent years due to their fast computation and scalability even when the probability measures lie in a very high dimensional space. However, SW requires many unnecessary projection samples to approximate its value while Max-SW only uses the most important projection, which ignores the information of other useful directions. In order to account for these weaknesses, we propose a novel distance, named Distributional Sliced-Wasserstein distance (DSW), that finds an optimal distribution over projections that can balance between exploring distinctive projecting directions and the informativeness of projections themselves. We show that the DSW is a generalization of Max-SW, and it can be computed efficiently by searching for the optimal push-forward measure over a set of probability measures over the unit sphere satisfying certain regularizing constraints that favor distinct directions. Finally, we conduct extensive experiments with large-scale datasets to demonstrate the favorable performances of the proposed distances over the previous sliced-based distances in generative modeling applications.
****

Maximum Categorical Cross Entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in Convolutional Neural Networks (CNNs) by reducing overfitting
Nidhi Gowdra,Roopak Sinha,Stephen MacDonell,WeiQi Yan
Categorical Cross Entropy (CCE) is the most commonly used loss function in deep neural networks such as Convolutional Neural Networks (CNNs) for multi-class classification problems. In spite of the fact that CCE is highly susceptible to noise; CNN models trained without accounting for the unique noise characteristics of the input data, or noise introduced during model training, invariably suffer from overfitting affecting model generalizability. The lack of generalizability becomes especially apparent in the context of ethnicity/racial image classification problems encountered in the domain of computer vision. One such problem is the unintended discriminatory racial bias that CNN models trained using CCE fail to adequately address. In other words, CNN models trained using CCE offer a skewed representation of classification performance favoring lighter skin tones.

In this paper, we propose and empirically validate a novel noise-robust extension to the existing CCE loss function called Maximum Categorical Cross-Entropy (MCCE), which utilizes CCE loss and a novel reconstruction loss, calculated using the Maximum Entropy (ME) measures of the convolutional kernel weights and input training dataset. We compare the use of MCCE with CCE-trained models on two benchmarking datasets, colorFERET and UTKFace, using a Residual Network (ResNet) CNN architecture. MCCE-trained models reduce overfitting by 5.85% and 4.3% on colorFERET and UTKFace datasets respectively. In cross-validation testing, MCCE-traine

d models outperform CCE-trained models by 8.8% and 25.16% on the colorFERET and UTKFace datasets respectively. MCCE addresses and mitigates the persistent problem of inadvertent racial bias for facial recognition problems in the domain of computer vision.

****

The act of remembering: A study in partially observable reinforcement learning

Rodrigo Toro Icarte,Richard Valenzano,Toryn Q. Klassen,Phillip Christoffersen,Amir-massoud Farahmand,Sheila A. McIlraith

Partial observability remains a major challenge for reinforcement learning (RL). In fully observable environments it is sufficient for RL agents to learn memoryless policies. However, some form of memory is necessary when RL agents are faced with partial observability. In this paper we study a lightweight approach: we augment the environment with an external memory and additional actions to control what, if anything, is written to the memory. At every step, the current memory state is part of the agent's observation, and the agent selects a tuple of actions: one action that modifies the environment and another that modifies the memory. When the external memory is sufficiently expressive, optimal memoryless policies yield globally optimal solutions. We develop the theory for memory-augmented environments and formalize the RL problem. Previous attempts to use external memory in the form of binary memory have produced poor results in practice. We propose and experimentally evaluate alternative forms of k-size buffer memory where the agent can decide to remember observations by pushing (or not) them into the buffer. Our memories are simple to implement and outperform binary and LSTM-based memories in well-established partially observable domains.

****

Evolving Reinforcement Learning Algorithms

John D Co-Reyes,Yingjie Miao,Daiyi Peng,Esteban Real,Quoc V Le,Sergey Levine,Honglak Lee,Aleksandra Faust

We propose a method for meta-learning reinforcement learning algorithms by searching over the space of computational graphs which compute the loss function for a value-based model-free RL agent to optimize. The learned algorithms are domain-agnostic and can generalize to new environments not seen during training. Our method can both learn from scratch and bootstrap off known existing algorithms, like DQN, enabling interpretable modifications which improve performance. Learning from scratch on simple classical control and gridworld tasks, our method rediscovers the temporal-difference (TD) algorithm. Bootstrapped from DQN, we highlight two learned algorithms which obtain good generalization performance over other classical control tasks, gridworld type tasks, and Atari games. The analysis of the learned algorithm behavior shows resemblance to recently proposed RL algorithms that address overestimation in value-based methods.

****

Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms

Chao Yu,Akash Velu,Eugene Vinitsky,Yu Wang,Alexandre Bayen,Yi Wu

We benchmark commonly used multi-agent deep reinforcement learning (MARL) algorithms on a variety of cooperative multi-agent games. While there has been significant innovation in MARL algorithms, algorithms tend to be tested and tuned on a single domain and their average performance across multiple domains is less characterized. Furthermore, since the hyperparameters of the algorithms are carefully tuned to the task of interest, it is unclear whether hyperparameters can easily be found that allow the algorithm to be repurposed for other cooperative tasks with different reward structure and environment dynamics. To investigate the consistency of the performance of MARL algorithms, we build an open-source library of multi-agent algorithms including DDPG/TD3/SAC with centralized Q functions, PPO with centralized value functions, and QMix and test them across a range of tasks that vary in coordination difficulty and agent number. The domains include the particle-world environments, starcraft micromanagement challenges, the Hanabi challenge, and the hide-and-seek environments. Finally, we investigate the ease of hyper-parameter tuning for each of the algorithms by tuning hyper-parameters in one environment per domain and re-using them in the other environments within the domain.

****
Systematic generalisation with group invariant predictions
Faruk Ahmed,Yoshua Bengio,Harm van Seijen,Aaron Courville
We consider situations where the presence of dominant simpler correlations with the target variable in a training set can cause an SGD-trained neural network to be less reliant on more persistently correlating complex features. When the non-persistent, simpler correlations correspond to non-semantic background factors, a neural network trained on this data can exhibit dramatic failure upon encountering systematic distributional shift, where the correlating background features are recombined with different objects. We perform an empirical study on three synthetic datasets, showing that group invariance methods across inferred partitionings of the training set can lead to significant improvements at such test-time situations. We also suggest a simple invariance penalty, showing with experiments on our setups that it can perform better than alternatives. We find that even without assuming access to any systematically shifted validation sets, one can still find improvements over an ERM-trained reference model.
****
Counterfactual Fairness through Data Preprocessing
Haoyu Chen,Wenbin Lu,Rui Song,Pulak Ghosh
Machine learning has become more important in real-life decision-making but people are concerned about the ethical problems it may bring when used improperly. Recent work brings the discussion of machine learning fairness into the causal framework and elaborates on the concept of Counterfactual Fairness. In this paper, we develop the Fair Learning through dAta Preprocessing (FLAP) algorithm to learn counterfactually fair decisions from biased training data and formalize the conditions where different data preprocessing procedures should be used to guarantee counterfactual fairness. We also show that Counterfactual Fairness is equivalent to the conditional independence of the decisions and the sensitive attributes given the processed non-sensitive attributes, which enables us to detect discrimination in the original decision using the processed data. The performance of our algorithm is illustrated using simulated data and real-world applications.
****
Achieving Explainability in a Visual Hard Attention Model through Content Prediction
Samrudhdhi B. Rangrej,James J. Clark
A visual hard attention model actively selects and observes a sequence of subregions in an image to make a prediction. Unlike in the deep convolution network, in hard attention it is explainable which regions of the image contributed to the prediction. However, the attention policy used by the model to select these regions is not explainable. The majority of hard attention models determine the attention-worthy regions by first analyzing a complete image. However, it may be the case that the entire image is not available in the beginning but instead sensed gradually through a series of partial observations. In this paper, we design an efficient hard attention model for classifying partially observable scenes. The attention policy used by our model is explainable and non-parametric. The model estimates expected information gain (EIG) obtained from attending various regions by predicting their content ahead of time. It compares EIG using Bayesian Optimal Experiment Design and attends to the region with maximum EIG. We train our model with a differentiable objective, optimized using gradient descent, and test it on several datasets. The performance of our model is comparable to or better than the baseline models.
****
Layer-adaptive Sparsity for the Magnitude-based Pruning
Jaeho Lee,Sejun Park,Sangwoo Mo,Sungsoo Ahn,Jinwoo Shin
Recent discoveries on neural network pruning reveal that, with a carefully chosen layerwise sparsity, a simple magnitude-based pruning achieves state-of-the-art tradeoff between sparsity and performance. However, without a clear consensus on ``how to choose,'' the layerwise sparsities are mostly selected algorithm-by-algorithm, often resorting to handcrafted heuristics or an extensive hyperparameter search. To fill this gap, we propose a novel importance score for global prun

ing, coined layer-adaptive magnitude-based pruning (LAMP) score; the score is a rescaled version of weight magnitude that incorporates the model-level ℓ_2 distortion incurred by pruning, and does not require any hyperparameter tuning or heavy computation.
Under various image classification setups, LAMP consistently outperforms popular existing schemes for layerwise sparsity selection.
Furthermore, we observe that LAMP continues to outperform baselines even in weight-rewinding setups, while the connectivity-oriented layerwise sparsity (the strongest baseline overall) performs worse than a simple global magnitude-based pruning in this case. Code: https://github.com/jaeho-lee/layer-adaptive-sparsity
****

TimeAutoML: Autonomous Representation Learning for Multivariate Irregularly Sampled Time Series
Yang Jiao,Kai Yang,shaoyu dou,pan luo,Sijia Liu,Dongjin Song
Multivariate time series (MTS) data are becoming increasingly ubiquitous in diverse domains, e.g., IoT systems, health informatics, and 5G networks. To obtain an effective representation of MTS data, it is not only essential to consider unpredictable dynamics and highly variable lengths of these data but also important to address the irregularities in the sampling rates of MTS. Existing parametric approaches rely on manual hyperparameter tuning and may cost a huge amount of labor effort. Therefore, it is desirable to learn the representation automatically and efficiently. To this end, we propose an autonomous representation learning approach for multivariate time series (TimeAutoML) with irregular sampling rates and variable lengths. As opposed to previous works, we first present a representation learning pipeline in which the configuration and hyperparameter optimization
are fully automatic and can be tailored for various tasks, e.g., anomaly detection, clustering, etc. Next, a negative sample generation approach and an auxiliary classification task are developed and integrated within TimeAutoML to enhance its representation capability. Extensive empirical studies on real-world datasets demonstrate that the proposed TimeAutoML outperforms competing approaches on various tasks by a large margin. In fact, it achieves the best anomaly detection performance among all comparison algorithms on 78 out of all 85 UCR datasets, acquiring up to 20% performance improvement in terms of AUC score.
****

Evaluating representations by the complexity of learning low-loss predictors
William F Whitney,Min Jae Song,David Brandfonbrener,Jaan Altosaar,Kyunghyun Cho
We consider the problem of evaluating representations of data for use in solving a downstream task. We propose to measure the quality of a representation by the complexity of learning a predictor on top of the representation that achieves low loss on a task of interest. To this end, we introduce two measures: surplus description length (SDL) and ε sample complexity (εSC). To compare our methods to prior work, we also present a framework based on plotting the validation loss versus dataset size (the "loss-data" curve). Existing measures, such as mutual information and minimum description length, correspond to slices and integrals along the data-axis of the loss-data curve, while ours correspond to slices and integrals along the loss-axis. This analysis shows that prior methods measure properties of an evaluation dataset of a specified size, whereas our methods measure properties of a predictor with a specified loss. We conclude with experiments on real data to compare the behavior of these methods over datasets of varying size.
****

Contrastive Code Representation Learning
Paras Jain,Ajay Jain,Tianjun Zhang,Pieter Abbeel,Joseph E. Gonzalez,Ion Stoica
Machine-aided programming tools such as automated type predictors and autocomplete are increasingly learning-based. However, current approaches predominantly rely on supervised learning with task-specific datasets. We propose Contrastive Code Representation Learning (ContraCode), a self-supervised algorithm for learning task-agnostic semantic representations of programs via contrastive learning. Our approach uses no human-provided labels, only the raw text of programs. Contra

Code optimizes for a representation that is invariant to semantic-preserving cod
e transformations. We develop an automated source-to-source compiler that genera
tes textually divergent variants of source programs. We then train a neural netw
ork to identify variants of anchor programs within a large batch of non-equivale
nt negatives. To solve this task, the network must extract features representing
 the functionality, not form, of the program. In experiments, we pre-train Contr
aCode with 1.8M unannotated JavaScript methods mined from GitHub, then transfer
 to downstream tasks by fine-tuning. Pre-training with ContraCode consistently im
proves the F1 score of code summarization baselines and top-1 accuracy of type i
nference baselines by 2% to 13%. ContraCode achieves 9% higher top-1 accuracy th
an the current state-of-the-art static type analyzer for TypeScript. Finally, re
presentations learned through a hybrid contrastive and reconstruction objective
 transfer in zero-shot to code clone detection with +10% AUROC over a static text
 similarity measure and +5% over reconstruction alone.
****

Understanding and Improving Encoder Layer Fusion in Sequence-to-Sequence Learnin
g
Xuebo Liu,Longyue Wang,Derek F. Wong,Liang Ding,Lidia S. Chao,Zhaopeng Tu
Encoder layer fusion (EncoderFusion) is a technique to fuse all the encoder laye
rs (instead of the uppermost layer) for sequence-to-sequence (Seq2Seq) models, w
hich has proven effective on various NLP tasks. However, it is still not entirel
y clear why and when EncoderFusion should work. In this paper, our main contribu
tion is to take a step further in understanding EncoderFusion. Many of previous
 studies believe that the success of EncoderFusion comes from exploiting surface
 and syntactic information embedded in lower encoder layers. Unlike them, we find
 that the encoder embedding layer is more important than other intermediate enco
der layers. In addition, the uppermost decoder layer consistently pays more atte
ntion to the encoder embedding layer across NLP tasks. Based on this observation
, we propose a simple fusion method, SurfaceFusion, by fusing only the encoder e
mbedding layer for the softmax layer. Experimental results show that SurfaceFusi
on outperforms EncoderFusion on several NLP benchmarks, including machine transl
ation, text summarization, and grammatical error correction. It obtains the stat
e-of-the-art performance on WMT16 Romanian-English and WMT14 English-French tran
slation tasks. Extensive analyses reveal that SurfaceFusion learns more expressi
ve bilingual word embeddings by building a closer relationship between relevant
 source and target embeddings. Source code is freely available at https://github.
com/SunbowLiu/SurfaceFusion.
****

Dynamic Graph Representation Learning with Fourier Temporal State Embedding
Yihan He,Wei Cao,Shun Zheng,Zhifeng Gao,Jiang Bian
Static graph representation learning has been applied in many tasks over the yea
rs thanks to the invention of unsupervised graph embedding methods and more rece
ntly, graph neural networks (GNNs). However, in many cases, we are to handle dyn
amic graphs where the structures of graphs and labels of the nodes are evolving
 steadily with time. This has posed a great challenge to existing methods in time
 and memory efficiency. In this work, we present a new method named Fourier Temp
oral State Embedding (FTSE) to address the temporal information in dynamic graph
 representation learning. FTSE offered time and memory-efficient solution throug
h applying signal processing techniques to the temporal graph signals. We paired
 the Fourier Transform with an efficient edge network and provided a new prototy
pe of modeling dynamic graph evolution with high precision. FTSE can also preven
t the 'history explosion' that exists in sequential models. The empirical study
 shows that our proposed approach achieves significantly better performance than
 previous approaches on public datasets across multiple tasks.
****

SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularizat
ion
A F M Shahab Uddin,Mst. Sirazam Monira,Wheemyung Shin,TaeChoong Chung,Sung-Ho Ba
e
Advanced data augmentation strategies have widely been studied to improve the ge

neralization ability of deep learning models. Regional dropout is one of the pop
ular solutions that guides the model to focus on less discriminative parts by ra
ndomly removing image regions, resulting in improved regularization. However, su
ch information removal is undesirable. On the other hand, recent strategies sugg
est to randomly cut and mix patches and their labels among training images, to e
njoy the advantages of regional dropout without having any pointless pixel in th
e augmented images. We argue that such random selection strategies of the patche
s may not necessarily represent sufficient information about the corresponding o
bject and thereby mixing the labels according to that uninformative patch enable
s the model to learn unexpected feature representation. Therefore, we propose Sa
liencyMix that carefully selects a representative image patch with the help of a
 saliency map and mixes this indicative patch with the target image, thus leadin
g the model to learn more appropriate feature representation. SaliencyMix achiev
es the best known top-1 error of 21.26% and 20.09% for ResNet-50 and ResNe
t-101 architectures on ImageNet classification, respectively, and also improves
the model robustness against adversarial perturbations. Furthermore, models that
 are trained with SaliencyMix, help to improve the object detection performance.
  Source code is available at \url{https://github.com/SaliencyMix/SaliencyMix}.
****

ReaPER: Improving Sample Efficiency in Model-Based Latent Imagination
Martin A Bertran,Guillermo Sapiro,mariano phielipp
Deep Reinforcement Learning (DRL) can distill behavioural policies from sensory
input that solve complex tasks, however, the policies tend to be task-specific a
nd sample inefficient, requiring a large number of interactions with the environ
ment that may be costly or impractical for many real world applications. Model-b
ased DRL (MBRL) can allow learned behaviours and dynamics from one task to be tr
anslated to a new task in a related environment, but still suffer from low sampl
e efficiency. In this work we introduce ReaPER, an algorithm that addresses the
sample efficiency challenge in model-based DRL, we illustrate the power of the p
roposed solution on the DeepMind Control benchmark. Our improvements are driven
by sparse , self-supervised, contrastive model representations and efficient use
 of past experience. We empirically analyze each novel component of ReaPER and
analyze how they contribute to sample efficiency. We also illustrate how other s
tandard alternatives fail to improve upon previous methods. Code will be made av
ailable.
****

Active Learning in CNNs via Expected Improvement Maximization
Udai G. Nagpal,David A. Knowles
Deep learning models such as Convolutional Neural Networks (CNNs) have demonstra
ted high levels of effectiveness in a variety of domains, including computer vis
ion and more recently, computational biology. However, training effective models
 often requires assembling and/or labeling large datasets, which may be prohibit
ively time-consuming or costly. Pool-based active learning techniques have the p
otential to mitigate these issues, leveraging models trained on limited data to
selectively query unlabeled data points from a pool in an attempt to expedite th
e learning process. Here we present "Dropout-based Expected IMprOvementS" (DEIMO
S), a flexible and computationally-efficient approach to active learning that qu
eries points that are expected to maximize the model's improvement across a repr
esentative sample of points. The proposed framework enables us to maintain a pre
diction covariance matrix capturing model uncertainty, and to dynamically update
 this matrix in order to generate diverse batches of points in the batch-mode se
tting. Our active learning results demonstrate that DEIMOS outperforms several e
xisting baselines across multiple regression and classification tasks taken from
 computer vision and genomics.
****

Efficient Estimators for Heavy-Tailed Machine Learning
Vishwak Srinivasan,Adarsh Prasad,Sivaraman Balakrishnan,Pradeep Kumar Ravikumar
A dramatic improvement in data collection technologies has aided in procuring ma
ssive amounts of unstructured and heterogeneous datasets. This has consequently
led to a prevalence of heavy-tailed distributions across a broad range of tasks

in machine learning. In this work, we perform thorough empirical studies to show that modern machine learning models such as generative adversarial networks and invertible flow models are plagued with such ill-behaved distributions during the phase of training them. To alleviate this problem, we develop a computationally-efficient estimator for mean estimation with provable guarantees which can handle such ill-behaved distributions. We provide specific consequences of our theory for supervised learning tasks such as linear regression and generalized linear models. Furthermore, we study the performance of our algorithm on synthetic tasks and real-world experiments and show that our methods convincingly outperform a variety of practical baselines.
****

Image GANs meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering

Yuxuan Zhang,Wenzheng Chen,Huan Ling,Jun Gao,Yinan Zhang,Antonio Torralba,Sanja Fidler

Differentiable rendering has paved the way to training neural networks to perform “inverse graphics” tasks such as predicting 3D geometry from monocular photographs. To train high performing models, most of the current approaches rely on multi-view imagery which are not readily available in practice. Recent Generative Adversarial Networks (GANs) that synthesize images, in contrast, seem to acquire 3D knowledge implicitly during training: object viewpoints can be manipulated by simply manipulating the latent codes. However, these latent codes often lack further physical interpretation and thus GANs cannot easily be inverted to perform explicit 3D reasoning. In this paper, we aim to extract and disentangle 3D knowledge learned by generative models by utilizing differentiable renderers. Key to our approach is to exploit GANs as a multi-view data generator to train an inverse graphics network using an off-the-shelf differentiable renderer, and the trained inverse graphics network as a teacher to disentangle the GAN's latent code into interpretable 3D properties. The entire architecture is trained iteratively using cycle consistency losses. We show that our approach significantly outperforms state-of-the-art inverse graphics networks trained on existing datasets, both quantitatively and via user studies. We further showcase the disentangled GAN as a controllable 3D “neural renderer", complementing traditional graphics renderers.
****

Meta-Learning with Implicit Processes

YIZHOU CHEN,DONG LI,NA LI,TONG LIANG,SHIZHUO ZHANG,Bryan Kian Hsiang Low

This paper presents a novel implicit process-based meta-learning (IPML) algorithm that, in contrast to existing works, explicitly represents each task as a continuous latent vector and models its probabilistic belief within the highly expressive IP framework. Unfortunately, meta-training in IPML is computationally challenging due to its need to perform intractable exact IP inference in task adaptation. To resolve this, we propose a novel expectation-maximization algorithm based on the stochastic gradient Hamiltonian Monte Carlo sampling method to perform meta-training. Our delicate design of the neural network architecture for meta-training in IPML allows competitive meta-learning performance to be achieved. Unlike existing works, IPML offers the benefits of being amenable to the characterization of a principled distance measure between tasks using the maximum mean discrepancy, active task selection without needing the assumption of known task contexts, and synthetic task generation by modeling task-dependent input distributions. Empirical evaluation on benchmark datasets shows that IPML outperforms existing Bayesian meta-learning algorithms. We have also empirically demonstrated on an e-commerce company's real-world dataset that IPML outperforms the baselines and identifies outliers tasks which can potentially degrade meta-testing performance.
****

EMaQ: Expected-Max Q-Learning Operator for Simple Yet Effective Offline and Online RL

Seyed Kamyar Seyed Ghasemipour,Dale Schuurmans,Shixiang Gu

Off-policy reinforcement learning (RL) holds the promise of sample-efficient lea

rning of decision-making policies by leveraging past experience. However, in the offline RL setting -- where a fixed collection of interactions are provided and no further interactions are allowed -- it has been shown that standard off-policy RL methods can significantly underperform. Recently proposed methods often aim to address this shortcoming by constraining learned policies to remain close to the given dataset of interactions. In this work, we closely investigate an important simplification of BCQ~\citep{fujimoto2018off} -- a prior approach for offline RL -- which removes a heuristic design choice and naturally restrict extracted policies to remain \emph{exactly} within the support of a given behavior policy. Importantly, in contrast to their original theoretical considerations, we derive this simplified algorithm through the introduction of a novel backup operator, Expected-Max Q-Learning (EMaQ), which is more closely related to the resulting practical algorithm. Specifically, in addition to the distribution support, EMaQ explicitly considers the number of samples and the proposal distribution, allowing us to derive new sub-optimality bounds which can serve as a novel measure of complexity for offline RL problems. In the offline RL setting -- the main focus of this work -- EMaQ matches and outperforms prior state-of-the-art in the D4RL benchmarks~\citep{fu2020d4rl}. In the online RL setting, we demonstrate that EMaQ is competitive with Soft Actor Critic (SAC). The key contributions of our empirical findings are demonstrating the importance of careful generative model design for estimating behavior policies, and an intuitive notion of complexity for offline RL problems. With its simple interpretation and fewer moving parts, such as no explicit function approximator representing the policy, EMaQ serves as a strong yet easy to implement baseline for future work.
****

Geometry of Program Synthesis
James Clift,Daniel Murfet,James Wallbridge
We present a new perspective on program synthesis in which programs may be identified with singularities of analytic functions. As an example, Turing machines are synthesised from input-output examples by propagating uncertainty through a smooth relaxation of a universal Turing machine. The posterior distribution over weights is approximated using Markov chain Monte Carlo and bounds on the generalisation error of these models is estimated using the real log canonical threshold, a geometric invariant from singular learning theory.

****

Federated Learning With Quantized Global Model Updates
Mohammad Mohammadi Amiri,Deniz Gunduz,Sanjeev Kulkarni,H. Vincent Poor
We study federated learning (FL), which enables mobile devices to utilize their local datasets to collaboratively train a global model with the help of a central server, while keeping data localized. At each iteration, the server broadcasts the current global model to the devices for local training, and aggregates the local model updates from the devices to update the global model. Previous work on the communication efficiency of FL has mainly focused on the aggregation of model updates from the devices, assuming perfect broadcasting of the global model. In this paper, we instead consider broadcasting a compressed version of the global model. This is to further reduce the communication cost of FL, which can be particularly limited when the global model is to be transmitted over a wireless medium. We introduce a lossy FL (LFL) algorithm, in which both the global model and the local model updates are quantized before being transmitted. We analyze the convergence behavior of the proposed LFL algorithm assuming the availability of accurate local model updates at the server. Numerical experiments show that the proposed LFL scheme, which quantizes the global model update (with respect to the global model estimate at the devices) rather than the global model itself, significantly outperforms other existing schemes studying quantization of the global model at the PS-to-device direction. Also, the performance loss of the proposed scheme is marginal compared to the fully lossless approach, where the PS and the devices transmit their messages entirely without any quantization.
****

A Half-Space Stochastic Projected Gradient Method for Group Sparsity Regularizat

ion
Tianyi Chen,Guanyi Wang,Tianyu DING,Bo Ji,Sheng Yi,Zhihui Zhu
Optimizing with group sparsity is significant in enhancing model interpretabilit y in machining learning applications, e.g., feature selection, compressed sensin g and model compression. However, for large-scale stochastic training problems, effective group-sparsity exploration are typically hard to achieve. Particularl y, the state-of-the-art stochastic optimization algorithms usually generate mere ly dense solutions. To overcome this shortage, we propose a stochastic method—Ha lf-space Stochastic Projected Gradient method (HSPG) to search solutions of high group sparsity while maintain the convergence. Initialized by a simple Prox-SG Step, the HSPG method relies on a novel Half-Space Step to substantially boosts the sparsity level. Numerically, HSPG demonstrates its superiority in deep neur al networks, e.g., VGG16, ResNet18 and MobileNetV1, by computing solutions of hi gher group sparsity, competitive objective values and generalization accuracy.
****
Augmentation-Interpolative AutoEncoders for Unsupervised Few-Shot Image Generati on
Davis Wertheimer,Omid Poursaeed,Bharath Hariharan
We aim to build image generation models that generalize to new domains from few examples. To this end, we first investigate the generalization properties of cla ssic image generators, and discover that autoencoders generalize extremely well to new domains, even when trained on highly constrained data. We leverage this i nsight to produce a robust, unsupervised few-shot image generation algorithm, an d introduce a novel training procedure based on recovering an image from data au gmentations. Our Augmentation-Interpolative AutoEncoders synthesize realistic im ages of novel objects from only a few reference images, and outperform both prio r interpolative models and supervised few-shot image generators. Our procedure i s simple and lightweight, generalizes broadly, and requires no category labels o r other supervision during training.
****
Are wider nets better given the same number of parameters?
Anna Golubeva,Guy Gur-Ari,Behnam Neyshabur
Empirical studies demonstrate that the performance of neural networks improves w ith increasing number of parameters. In most of these studies, the number of par ameters is increased by increasing the network width. This begs the question: Is the observed improvement due to the larger number of parameters, or is it due t o the larger width itself? We compare different ways of increasing model width w hile keeping the number of parameters constant. We show that for models initiali zed with a random, static sparsity pattern in the weight tensors, network width is the determining factor for good performance, while the number of weights is s econdary, as long as the model achieves high training accuarcy. As a step toward s understanding this effect, we analyze these models in the framework of Gaussia n Process kernels. We find that the distance between the sparse finite-width mod el kernel and the infinite-width kernel at initialization is indicative of model performance.
****
TextTN: Probabilistic Encoding of Language on Tensor Network
Peng Zhang,Jing Zhang,Xindian Ma,Siwei Rao,Guangjian Tian,Jun Wang
As a novel model that bridges machine learning and quantum theory, tensor networ k (TN) has recently gained increasing attention and successful applications for processing natural images. However, for natural languages, it is unclear how to design a probabilistic encoding architecture to efficiently and accurately lear n and classify texts based on TN. This paper proposes a general two-step scheme of text classification based on Tensor Network, which is named as TextTN. TextTN first encodes the word vectors in a probabilistic space by a generative TN (wor d-GTN), and then classifies a text sentence using a discriminative TN (sentence-DTN). Moreover, in sentence-DTN, its hyper-parameter (i.e., bond-dimension) can be analyzed and selected by the theoretical property of TextTN's expressive pow er. In experiments, our TextTN also obtains the state-of-the-art result on SST-5 sentiment classification task.

****
A Truly Constant-time Distribution-aware Negative Sampling
Shabnam Daghaghi,Tharun Medini,Beidi Chen,Mengnan Zhao,Anshumali Shrivastava
Softmax classifiers with a very large number of classes naturally occur in many applications such as natural language processing and information retrieval. The calculation of full-softmax is very expensive from the computational and energy perspective. There have been a variety of sampling approaches to overcome this challenge, popularly known as negative sampling (NS). Ideally, NS should sample negative classes from a distribution that is dependent on the input data, the current parameters, and the correct positive class. Unfortunately, due to the dynamically updated parameters and data samples, there does not exist any sampling scheme that is truly adaptive and also samples the negative classes in constant time every iteration. Therefore, alternative heuristics like random sampling, static frequency-based sampling, or learning-based biased sampling; which primarily trade either the sampling cost or the adaptivity of samples per iteration, are adopted. In this paper, we show a class of distribution where the sampling scheme is truly adaptive and provably generates negative samples in constant time. We demonstrate a negative sampling implementation that is significantly faster, in terms of wall clock time, compared to the most optimized TensorFlow implementations of standard softmax or other sampling approaches on the best available GPUs (V100s).
****
Adaptive Learning Rates for Multi-Agent Reinforcement Learning
Jiechuan Jiang,Zongqing Lu
In multi-agent reinforcement learning (MARL), the learning rates of actors and critic are mostly hand-tuned and fixed. This not only requires heavy tuning but more importantly limits the learning. With adaptive learning rates according to gradient patterns, some optimizers have been proposed for general optimizations, which however do not take into consideration the characteristics of MARL. In this paper, we propose AdaMa to bring adaptive learning rates to cooperative MARL. AdaMa evaluates the contribution of actors' updates to the improvement of Q-value and adaptively updates the learning rates of actors to the direction of maximally improving the Q-value. AdaMa could also dynamically balance the learning rates between the critic and actors according to their varying effects on the learning. Moreover, AdaMa can incorporate the second-order approximation to capture the contribution of pairwise actors' updates and thus more accurately updates the learning rates of actors. Empirically, we show that AdaMa could accelerate the learning and improve the performance in a variety of multi-agent scenarios, and the visualizations of learning rates during training clearly explain how and why AdaMa works.
****
Discovering Non-monotonic Autoregressive Orderings with Variational Inference
Xuanlin Li,Brandon Trabucco,Dong Huk Park,Michael Luo,Sheng Shen,Trevor Darrell,Yang Gao
The predominant approach for language modeling is to encode a sequence of tokens from left to right, but this eliminates a source of information: the order by which the sequence was naturally generated. One strategy to recover this information is to decode both the content and ordering of tokens. Some prior work supervises content and ordering with hand-designed loss functions to encourage specific orders or bootstraps from a predefined ordering. These approaches require domain-specific insight. Other prior work searches over valid insertion operations that lead to ground truth sequences during training, which has high time complexity and cannot be efficiently parallelized. We address these limitations with an unsupervised learner that can be trained in a fully-parallelizable manner to discover high-quality autoregressive orders in a data driven way without a domain-specific prior. The learner is a neural network that performs variational inference with the autoregressive ordering as a latent variable. Since the corresponding variational lower bound is not differentiable, we develop a practical algorithm for end-to-end optimization using policy gradients. Strong empirical results with our solution on sequence modeling tasks suggest that our algorithm is capabl

e of discovering various autoregressive orders for different sequences that are competitive with or even better than fixed orders.
****

Adaptive Spatial-Temporal Inception Graph Convolutional Networks for Multi-step Spatial-Temporal Network Data Forecasting
Xing Wang,Lin Zhu,Juan Zhao,Zhou Xu,Zhao Li,Junlan Feng,Chao Deng
Spatial-temporal data forecasting is of great importance for industries such as telecom network operation and transportation management. However, spatial-temporal data is inherent with complex spatial-temporal correlations and behaves heterogeneities among the spatial and temporal aspects, which makes the forecasting remain as a very challenging task though recently great work has been done. In this paper, we propose a novel model, Adaptive Spatial-Temporal Inception Graph Convolution Networks (ASTI-GCN), to solve the multi-step spatial-temporal data forecasting problem. The model proposes multi-scale spatial-temporal joint graph convolution block to directly model the spatial-temporal joint correlations without introducing elaborately constructed mechanisms. Moreover inception mechanism combined with the graph node-level attention is introduced to make the model capture the heterogeneous nature of the graph adaptively. Our experiments on three real-world datasets from two different fields consistently show ASTI-GCN outperforms the state-of-the-art performance. In addition, ASTI-GCN is proved to generalize well.
****

Efficient Robust Training via Backward Smoothing
Jinghui Chen,Yu Cheng,Zhe Gan,Quanquan Gu,Jingjing Liu
Adversarial training is so far the most effective strategy in defending against adversarial examples. However, it suffers from high computational cost due to the iterative adversarial attacks in each training step. Recent studies show that it is possible to achieve Fast Adversarial Training by performing a single-step attack with random initialization. Yet, it remains a mystery why random initialization helps. Besides, such an approach still lags behind state-of-the-art adversarial training algorithms on both stability and model robustness. In this work, we develop a new understanding towards Fast Adversarial Training, by viewing random initialization as performing randomized smoothing for better optimization of the inner maximization problem. From this perspective, we show that the smoothing effect by random initialization is not sufficient under the adversarial perturbation constraint. A new initialization strategy, \emph{backward smoothing}, is proposed to address this issue and significantly improves both stability and model robustness over single-step robust training methods. Experiments on multiple benchmarks demonstrate that our method achieves similar model robustness as the original TRADES method, while using much less training time (~3x improvement with the same training schedule).
****

Deep Graph Neural Networks with Shallow Subgraph Samplers
Hanqing Zeng,Muhan Zhang,Yinglong Xia,Ajitesh Srivastava,Rajgopal Kannan,Viktor Prasanna,Long Jin,Andrey Malevich,Ren Chen
While Graph Neural Networks (GNNs) are powerful models for learning representations on graphs, most state-of-the-art models do not have significant accuracy gain beyond two to three layers. Deep GNNs fundamentally need to address: 1). expressivity challenge due to oversmoothing, and 2). computation challenge due to neighborhood explosion. We propose a simple "deep GNN, shallow sampler" design principle to improve both the GNN accuracy and efficiency --- to generate representation of a target node, we use a deep GNN to pass messages only within a shallow, localized subgraph. A properly sampled subgraph may exclude irrelevant or even noisy nodes, and still preserve the critical neighbor features and graph structures. The deep GNN then smooths the informative local signals to enhance feature learning, rather than oversmoothing the global graph signals into just "white noise". We theoretically justify why the combination of deep GNNs with shallow samplers yields the best learning performance. We then propose various sampling algorithms and neural architecture extensions to achieve good empirical results. Experiments on five large graphs show that our models achieve significantly higher

accuracy and efficiency, compared with state-of-the-art.
****
CompOFA – Compound Once-For-All Networks for Faster Multi-Platform Deployment
Manas Sahni,Shreya Varshini,Alind Khare,Alexey Tumanov
The emergence of CNNs in mainstream deployment has necessitated methods to design and train efficient architectures tailored to maximize the accuracy under diverse hardware and latency constraints. To scale these resource-intensive tasks with an increasing number of deployment targets, Once-For-All (OFA) proposed an approach to jointly train several models at once with a constant training cost. However, this cost remains as high as 40-50 GPU days and also suffers from a combinatorial explosion of sub-optimal model configurations. We seek to reduce this search space -- and hence the training budget -- by constraining search to models close to the accuracy-latency Pareto frontier. We incorporate insights of compound relationships between model dimensions to build CompOFA, a design space smaller by several orders of magnitude. Through experiments on ImageNet, we demonstrate that even with simple heuristics we can achieve a 2x reduction in training time and 216x speedup in model search/extraction time compared to the state of the art, without loss of Pareto optimality! We also show that this smaller design space is dense enough to support equally accurate models for a similar diversity of hardware and latency targets, while also reducing the complexity of the training and subsequent extraction algorithms. Our source code is available at https://github.com/gatech-sysml/CompOFA
****
Implicit Regularization of SGD via Thermophoresis
Mingwei Wei,David J. Schwab
A central ingredient in the impressive predictive performance of deep neural networks is optimization via stochastic gradient descent (SGD). While some theoretical progress has been made, the effect of SGD in neural networks is still unclear, especially during the early phase of training. Here we generalize the theory of thermophoresis from statistical mechanics and show that there exists an effective entropic force from SGD that pushes to reduce the gradient variance. We study this effect in detail in a simple two-layer model, where the thermophoretic force functions to decreases the weight norm and activation rate of the units. The strength of this effect is proportional to squared learning rate and inverse batch size, and is more effective during the early phase of training when the model's predictions are poor. Lastly we test our quantitative predictions with experiments on various models and datasets.

****
Max-sliced Bures Distance for Interpreting Discrepancies
Austin J. Brockmeier,Claudio Cesar Claros,Carlos H. Mendoza-Cardenas,Yüksel Karahan,Matthew S. Emigh,Luis Gonzalo Sanchez Giraldo
We propose the max-sliced Bures distance, a lower bound on the max-sliced Wasserstein-2 distance, to identify the instances associated with the maximum discrepancy between two samples. The max-slicing can be decomposed into two asymmetric divergences each expressed in terms of an optimal slice or equivalently a witness function that has large magnitude evaluations on a localized subset of instances in one distribution versus the other. We show how witness functions can be used to detect and correct for covariate shift through reweighting and to evaluate generative adversarial networks. Unlike heuristic algorithms for the max-sliced Wasserstein-2 distance that may fail to find the optimal slice, we detail a tractable algorithm that finds the global optimal slice and scales to large sample sizes. As the Bures distance quantifies differences in covariance, we generalize the max-sliced Bures distance by using non-linear mappings, enabling it to capture changes in higher-order statistics. We explore two types of non-linear mappings: positive semidefinite kernels where the witness functions belong to a reproducing kernel Hilbert space, and task-relevant mappings corresponding to a neural network. In the context of samples of natural images, our approach provides an interpretation of the Fréchet Inception distance by identifying the synthetic and natural instances that are either over-represented or under-represented with

respect to the other sample. We apply the proposed measure to detect imbalances in class distributions in various data sets and to critique generative models.
****
Blind Pareto Fairness and Subgroup Robustness
Natalia Martinez,Martin Bertran,Afroditi Papadaki,Miguel R. D. Rodrigues,Guillermo Sapiro
With the wide adoption of machine learning algorithms across various application domains, there is a growing interest in the fairness properties of such algorithms. The vast majority of the activity in the field of group fairness addresses disparities between predeⓝned groups based on protected features such as gender, age, and race, which need to be available at train, and often also at test, time. These approaches are static and retrospective, since algorithms designed to protect groups identified a priori cannot anticipate and protect the needs of different at-risk groups in the future. In this work we analyze the space of solutions for worst-case fairness beyond demographics, and propose Blind Pareto Fairness (BPF), a method that leverages no-regret dynamics to recover a fair minimax classiⓕer that reduces worst-case risk of any potential subgroup of sufⓒient size, and guarantees that the remaining population receives the best possible level of service. BPF addresses fairness beyond demographics, that is, it does not rely on predeⓝned notions of at-risk groups, neither at train nor at test time. Our experimental results show that the proposed framework improves worst-case risk in multiple standard datasets, while simultaneously providing better levels of service for the remaining population, in comparison to competing methods.
****
Meta Gradient Boosting Neural Networks
Manqing Dong,Lina Yao,Xianzhi Wang,Xiwei Xu,Liming Zhu
Meta-optimization is an effective approach that learns a shared set of parameters across tasks for parameter initialization in meta-learning.
A key challenge for meta-optimization based approaches is to determine whether an initialization condition can be generalized to tasks with diverse distributions to accelerate learning.
To address this issue, we design a meta-gradient boosting framework that uses a base learner to learn shared information across tasks and a series of gradient-boosted modules to capture task-specific information to fit diverse distributions.
We evaluate the proposed model on both regression and classification tasks with multi-mode distributions.
The results demonstrate both the effectiveness of our model in modulating task-specific meta-learned priors and its advantages on multi-mode distributions.
****
Representing Partial Programs with Blended Abstract Semantics
Maxwell Nye,Yewen Pu,Matthew Bowers,Jacob Andreas,Joshua B. Tenenbaum,Armando Solar-Lezama
Synthesizing programs from examples requires searching over a vast, combinatorial space of possible programs. In this search process, a key challenge is representing the behavior of a partially written program before it can be executed, to judge if it is on the right track and predict where to search next. We introduce a general technique for representing partially written programs in a program synthesis engine. We take inspiration from the technique of abstract interpretation, in which an approximate execution model is used to determine if an unfinished program will eventually satisfy a goal specification. Here we learn an approximate execution model implemented as a modular neural network. By constructing compositional program representations that implicitly encode the interpretation semantics of the underlying programming language, we can represent partial programs using a flexible combination of concrete execution state and learned neural representations, using the learned approximate semantics when concrete semantics are not known (in unfinished parts of the program). We show that these hybrid neuro-symbolic representations enable execution-guided synthesizers to use more powerful language constructs, such as loops and higher-order functions, and can be used to synthesize programs more accurately for a given search budget than pure n

eural approaches in several domains.
****
When Are Neural Pruning Approximation Bounds Useful?
Mitchell A Gordon
Approximation bounds for neural network pruning attempt to predict the trade-off between sparsity and fidelity while shrinking neural networks. In the first half of this paper, we empirically evaluate the predictive power of two recently proposed methods based on coreset algorithms. We identify several circumstances in which the bounds are loose or impractical to use and provide a brief analysis of the components of the bounds that contribute to those short-comings. In the second half, we examine the role of fine-tuning in prunability and observe that even tight approximation bounds would be poor predictors of accuracy after fine-tuning. This is because fine-tuning can recover large amounts of accuracy while simultaneously maintaining or increasing approximation error. We discuss the implications of these finding on the application of coreset-based pruning methods in practice and the role of approximation in the pruning community. Our code is available in the attached supplementary material.
****
MONGOOSE: A Learnable LSH Framework for Efficient Neural Network Training
Beidi Chen,Zichang Liu,Binghui Peng,Zhaozhuo Xu,Jonathan Lingjie Li,Tri Dao,Zhao Song,Anshumali Shrivastava,Christopher Re
Recent advances by practitioners in the deep learning community have breathed new life into Locality Sensitive Hashing (LSH), using it to reduce memory and time bottlenecks in neural network (NN) training. However, while LSH has sub-linear guarantees for approximate near-neighbor search in theory, it is known to have inefficient query time in practice due to its use of random hash functions. Moreover, when model parameters are changing, LSH suffers from update overhead. This work is motivated by an observation that model parameters evolve slowly, such that the changes do not always require an LSH update to maintain performance. This phenomenon points to the potential for a reduction in update time and allows for a modified learnable version of data-dependent LSH to improve query time at a low cost. We use the above insights to build MONGOOSE, an end-to-end LSH framework for efficient NN training. In particular, MONGOOSE is equipped with a scheduling algorithm to adaptively perform LSH updates with provable guarantees and learnable hash functions to improve query efficiency. Empirically, we validate MONGOOSE on large-scale deep learning models for recommendation systems and language modeling. We find that it achieves up to 8% better accuracy compared to previous LSH approaches, with $6.5 \times$ speed-up and $6\times$ reduction in memory usage.
****
GenQu: A Hybrid System for Learning Classical Data in Quantum States
Samuel A. Stein,Ray Marie Tischio,Betis Baheri,Yiwen Chen,Ying Mao,Qiang Guan,Ang Li,Bo Fang
Deep neural network-powered artificial intelligence has rapidly changed our daily life with various applications. However, as one of the essential steps of deep neural networks, training a heavily-weighted network requires a tremendous amount of computing resources. Especially in the post Moore's Law era, the limit of semiconductor fabrication technology has restricted the development of learning algorithms to cope with the increasing high-intensity training data. Meanwhile, quantum computing has exhibited its significant potential in terms of speeding up the traditionally compute-intensive workloads. For example, Google illustrates quantum supremacy by completing a sampling calculation task in 200 seconds, which is otherwise impracticable on the world's largest supercomputers. To this end, quantum-based learning becomes an area of interest, with the promising of a quantum speedup. In this paper, we propose GenQu, a hybrid and general-purpose quantum framework for learning classical data through quantum states. We evaluate GenQu with real datasets and conduct experiments on both simulations and real quantum computer IBM-Q. Our evaluation demonstrates that, comparing with classical solutions, the proposed models running on GenQu framework achieve similar accuracy with a much smaller number of qubits, while significantly reducing the parame

ter size by up to 95.8\% and converging speedup by 66.67% faster.
****
PolarNet: Learning to Optimize Polar Keypoints for Keypoint Based Object Detecti
on
Wu Xiongwei,Doyen Sahoo,Steven HOI
A variety of anchor-free object detectors have been actively proposed as possibl
e alternatives to the mainstream anchor-based detectors that often rely on compl
icated design of anchor boxes. Despite achieving promising performance on par wi
th anchor-based detectors, the existing anchor-free detectors such as FCOS or Ce
nterNet predict objects based on standard Cartesian coordinates, which often yie
ld poor quality keypoints. Further, the feature representation is also scale-sen
sitive. In this paper, we propose a new anchor-free keypoint based detector ``Po
larNet", where keypoints are represented as a set of Polar coordinates instead o
f Cartesian coordinates. The ``PolarNet" detector learns offsets pointing to the
 corners of objects in order to learn high quality keypoints. Additionally, Pola
rNet uses features of corner points to localize objects, making the localization
 scale-insensitive. Finally in our experiments, we show that PolarNet, an anchor
-free detector, outperforms the existing anchor-free detectors, and it is able t
o achieve highly competitive result on COCO test-dev benchmark (47.8% and $50
.3\%$ AP under the single-model single-scale and multi-scale testing) which is o
n par with the state-of-the-art two-stage anchor-based object detectors. The cod
e and the models are available at https://github.com/XiongweiWu/PolarNetV1
****
Multi-agent Policy Optimization with Approximatively Synchronous Advantage Estim
ation
Lipeng Wan,Xuwei Song,Xuguang Lan,Nanning Zheng
Cooperative multi-agent tasks require agents to deduce their own contributions w
ith shared global rewards, known as the challenge of credit assignment. General
methods for policy based multi-agent reinforcement learning to solve the challen
ge introduce differentiate value functions or advantage functions for individual
 agents. In multi-agent system, polices of different agents need to be evaluated
 jointly. In order to update polices synchronously, such value functions or adva
ntage functions also need synchronous evaluation. However, in current methods, v
alue functions or advantage functions use counter-factual joint actions which ar
e evaluated asynchronously, thus suffer from natural estimation bias. In this wo
rk, we propose the approximatively synchronous advantage estimation. We first de
rive the marginal advantage function, an expansion from single-agent advantage f
unction to multi-agent system. Further more, we introduce a policy approximation
 for synchronous advantage estimation, and break down the multi-agent policy opt
imization problem into multiple sub-problems of single-agent policy optimization
. Our method is compared with baseline algorithms on StarCraft multi-agent chall
enges, and shows the best performance on most of the tasks.
****
Random Coordinate Langevin Monte Carlo
Zhiyan Ding,Qin Li,Jianfeng Lu,Stephen Wright
Langevin Monte Carlo (LMC) is a popular Markov chain Monte Carlo sampling method
. One drawback is that it requires the computation of the full gradient at each
iteration, an expensive operation if the dimension of the problem is high. We pr
opose a new sampling method: Random Coordinate LMC (RC-LMC). At each iteration,
a single coordinate is randomly selected to be updated by a multiple of the part
ial derivative along this direction plus noise, and all other coordinates remain
 untouched. We investigate the total complexity of RC-LMC and compare it with th
e classical LMC for log-concave probability distributions. When the gradient of
the log-density is Lipschitz, RC-LMC is less expensive than the classical LMC if
 the log-density is highly skewed for high dimensional problems, and when both t
he gradient and the Hessian of the log-density are Lipschitz, RC-LMC is always c
heaper than the classical LMC, by a factor proportional to the square root of th
e problem dimension. In the latter case, our estimate of complexity is sharp wit
h respect to the dimension.
****

Multiscale Invertible Generative Networks for High-Dimensional Bayesian Inferenc
e
Shumao Zhang,Thomas Hou,Pengchuan Zhang
High-dimensional Bayesian inference problems cast a long-standing challenge in g
enerating samples, especially when the posterior has multiple modes. For a wide
class of Bayesian inference problems equipped with the multiscale structure that
 low-dimensional (coarse-scale) surrogate can approximate the original high-dime
nsional (fine-scale) problem well, we propose to train a Multiscale Invertible G
enerative Network (MsIGN) for sample generation. A novel prior conditioning laye
r is designed to bridge networks at different resolutions, enabling coarse-to-fi
ne multi-stage training. Jeffreys divergence is adopted as the training objectiv
e to avoid mode dropping. On two high-dimensional Bayesian inverse problems, MsI
GN approximates the posterior accurately and clearly captures multiple modes, sh
owing superior performance compared with previous deep generative network approa
ches. On the natural image synthesis task, MsIGN achieves the superior performan
ce in bits-per-dimension compared with our baseline models and yields great inte
rpret-ability of its neurons in intermediate layers.
****

Mitigating Deep Double Descent by Concatenating Inputs
John Chen,Qihan Wang,Anastasios Kyrillidis
The double descent curve is one of the most intriguing properties of deep neural
 networks. It contrasts the classical bias-variance curve with the behavior of m
odern neural networks, occurring where the number of samples nears the number of
 parameters. In this work, we explore the connection between the double descent
phenomena and the number of samples in the deep neural network setting. In parti
cular, we propose a construction which augments the existing dataset by artifici
ally increasing the number of samples. This construction empirically mitigates t
he double descent curve in this setting. We reproduce existing work on deep doub
le descent, and observe a smooth descent into the overparameterized region for o
ur construction. This occurs both with respect to the model size, and with respe
ct to the number epochs.
****

Neural Architecture Search on ImageNet in Four GPU Hours: A Theoretically Inspir
ed Perspective
Wuyang Chen,Xinyu Gong,Zhangyang Wang
Neural Architecture Search (NAS) has been explosively studied to automate the di
scovery of top-performer neural networks. Current works require heavy training o
f supernet or intensive architecture evaluations, thus suffering from heavy reso
urce consumption and often incurring search bias due to truncated training or ap
proximations. Can we select the best neural architectures without involving any
training and eliminate a drastic portion of the search cost?

We provide an affirmative answer, by proposing a novel framework called \textit{
training-free neural architecture search} ($\textbf{TE-NAS}$). TE-NAS ranks arch
itectures by analyzing the spectrum of the neural tangent kernel (NTK), and the
number of linear regions in the input space. Both are motivated by recent theory
 advances in deep networks, and can be computed without any training. We show th
at: (1) these two measurements imply the $\textit{trainability}$ and $\textit{ex
pressivity}$ of a neural network; and (2) they strongly correlate with the netwo
rk's actual test accuracy. Further on, we design a pruning-based NAS mechanism t
o achieve a more flexible and superior trade-off between the trainability and ex
pressivity during the search. In NAS-Bench-201 and DARTS search spaces, TE-NAS c
ompletes high-quality search but only costs $\textbf{0.5}$ and $\textbf{4}$ GPU
hours with one 1080Ti on CIFAR-10 and ImageNet, respectively. We hope our work t
o inspire more attempts in bridging between the theoretic findings of deep netwo
rks and practical impacts in real NAS applications.
****

Unsupervised Representation Learning for Time Series with Temporal Neighborhood
Coding
Sana Tonekaboni,Danny Eytan,Anna Goldenberg

Time series are often complex and rich in information but sparsely labeled and t
herefore challenging to model. In this paper, we propose a self-supervised frame
work for learning robust and generalizable representations for time series. Our
approach, called Temporal Neighborhood Coding (TNC), takes advantage of the loca
l smoothness of a signal's generative process to define neighborhoods in time wi
th stationary properties. Using a debiased contrastive objective, our framework
learns time series representations by ensuring that in the encoding space, the d
istribution of signals from within a neighborhood is distinguishable from the di
stribution of non-neighboring signals. Our motivation stems from the medical fie
ld, where the ability to model the dynamic nature of time series data is especia
lly valuable for identifying, tracking, and predicting the underlying patients'
latent states in settings where labeling data is practically impossible. We comp
are our method to recently developed unsupervised representation learning approa
ches and demonstrate superior performance on clustering and classification tasks
for multiple datasets.
****
Certified Robustness of Nearest Neighbors against Data Poisoning Attacks
Jinyuan Jia,Xiaoyu Cao,Neil Zhenqiang Gong
Data poisoning attacks aim to corrupt a machine learning model via modifying, ad
ding, and/or removing some carefully selected training examples, such that the c
orrupted model predicts any or attacker-chosen incorrect labels for testing exam
ples. The key idea of state-of-the-art certified defenses against data poisoning
attacks is to create a \emph{majority vote} mechanism to predict the label of a
testing example. Moreover, each voter is a base classifier trained on a subset
of the training dataset. Nearest neighbor algorithms such as k nearest neighbo
rs (kNN) and radius nearest neighbors (rNN) have intrinsic majority vote mechan
isms. In this work, we show that the intrinsic majority vote mechanisms in kNN a
nd rNN already provide certified robustness guarantees against general data pois
oning attacks. Moreover, our empirical evaluation results on MNIST and CIFAR10 s
how that the intrinsic certified robustness guarantees of kNN and rNN outperform
those provided by state-of-the-art certified defenses.
****
Representational aspects of depth and conditioning in normalizing flows
Frederic Koehler,Viraj Mehta,Andrej Risteski
Normalizing flows are among the most popular paradigms in generative modeling, e
specially for images, primarily because we can efficiently evaluate the likeliho
od of a data point. This is desirable both for evaluating the fit of a model, an
d for ease of training, as maximizing the likelihood can be done by gradient des
cent. However, training normalizing flows comes with difficulties as well: model
s which produce good samples typically need to be extremely deep -- which comes
with accompanying vanishing/exploding gradient problems. A very related problem
is that they are often poorly \emph{conditioned}: since they are parametrized as
invertible maps from $\mathbb{R}^d \to \mathbb{R}^d$, and typical training data
like images intuitively is lower-dimensional, the learned maps often have Jacob
ians that are close to being singular.

In our paper, we tackle representational aspects around depth and conditioning o
f normalizing flows---both for general invertible architectures, and for a parti
cular common architecture---affine couplings.

For general invertible architectures, we prove that invertibility comes at a cos
t in terms of depth: we show examples where a much deeper normalizing flow model
may need to be used to match the performance of a non-invertible generator.

For affine couplings, we first show that the choice of partitions isn't a likely
bottleneck for depth: we show that any invertible linear map (and hence a permu
tation) can be simulated by a constant number of affine coupling layers, using a
fixed partition. This shows that the extra flexibility conferred by 1x1 convolu
tion layers, as in GLOW, can in principle be simulated by increasing the size by
a constant factor. Next, in terms of conditioning, we show that affine coupling

s are universal approximators -- provided the Jacobian of the model is allowed t
o be close to singular. We furthermore empirically explore the benefit of differ
ent kinds of padding -- a common strategy for improving conditioning.
****

Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network R
epresentations Vary with Width and Depth
Thao Nguyen,Maithra Raghu,Simon Kornblith
A key factor in the success of deep neural networks is the ability to scale mode
ls to improve performance by varying the architecture depth and width. This simp
le property of neural network design has resulted in highly effective architectu
res for a variety of tasks. Nevertheless, there is limited understanding of effe
cts of depth and width on the learned representations. In this paper, we study t
his fundamental question. We begin by investigating how varying depth and width
affects model hidden representations, finding a characteristic block structure i
n the hidden representations of larger capacity (wider or deeper) models. We dem
onstrate that this block structure arises when model capacity is large relative
to the size of the training set, and is indicative of the underlying layers pres
erving and propagating the dominant principal component of their representations
. This discovery has important ramifications for features learned by different m
odels, namely, representations outside the block structure are often similar acr
oss architectures with varying widths and depths, but the block structure is uni
que to each model. We analyze the output predictions of different model architec
tures, finding that even when the overall accuracy is similar, wide and deep mod
els exhibit distinctive error patterns and variations across classes.
****

Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning
Beliz Gunel,Jingfei Du,Alexis Conneau,Veselin Stoyanov
State-of-the-art natural language understanding classification models follow two
-stages: pre-training a large language model on an auxiliary task, and then fine
-tuning the model on a task-specific labeled dataset using cross-entropy loss. H
owever, the cross-entropy loss has several shortcomings that can lead to sub-opt
imal generalization and instability. Driven by the intuition that good generaliz
ation requires capturing the similarity between examples in one class and contra
sting them with examples in other classes, we propose a supervised contrastive l
earning (SCL) objective for the fine-tuning stage. Combined with cross-entropy,
our proposed SCL loss obtains significant improvements over a strong RoBERTa-Lar
ge baseline on multiple datasets of the GLUE benchmark in few-shot learning sett
ings, without requiring specialized architecture, data augmentations, memory ban
ks, or additional unsupervised data. Our proposed fine-tuning objective leads to
 models that are more robust to different levels of noise in the fine-tuning tra
ining data, and can generalize better to related tasks with limited labeled data
.
****

Differentially Private Generative Models Through Optimal Transport
Tianshi Cao,Alex Bie,Karsten Kreis,Sanja Fidler
Although machine learning models trained on massive data have led to breakthroug
hs in several areas, their deployment in privacy-sensitive domains remains limit
ed due to restricted access to data. Generative models trained with privacy cons
traints on private data can sidestep this challenge and provide indirect access
to the private data instead. We propose DP-Sinkhorn, a novel optimal transport-b
ased generative method for learning data distributions from private data with di
fferential privacy. DP-Sinkhorn relies on minimizing the Sinkhorn divergence---a
 computationally efficient approximation to the exact optimal transport distance
---between the model and the data in a differentially private manner and also us
es a novel technique for conditional generation in the Sinkhorn framework. Unlik
e existing approaches for training differentially private generative models, whi
ch are mostly based on generative adversarial networks, we do not rely on advers
arial objectives, which are notoriously difficult to optimize, especially in the
 presence of noise imposed by the privacy constraints. Hence, DP-Sinkhorn is eas
y to train and deploy. Experimentally, despite our method's simplicity we improv

e upon the state-of-the-art on multiple image modeling benchmarks. We also show differentially private synthesis of informative RGB images, which has not been demonstrated before by differentially private generative models without the use of auxiliary public data.

****

R-LAtte: Attention Module for Visual Control via Reinforcement Learning

Mandi Zhao,Qiyang Li,Aravind Srinivas,Ignasi Clavera,Kimin Lee,Pieter Abbeel

Attention mechanisms are generic inductive biases that have played a critical role in improving the state-of-the-art in supervised learning, unsupervised pre-training and generative modeling for multiple domains including vision, language and speech. However, they remain relatively under-explored for neural network architectures typically used in reinforcement learning (RL) from high dimensional inputs such as pixels. In this paper, we propose and study the effectiveness of augmenting a simple attention module in the convolutional encoder of an RL agent. Through experiments on the widely benchmarked DeepMind Control Suite environments, we demonstrate that our proposed module can (i) extract interpretable task-relevant information such as agent locations and movements without the need for data augmentations or contrastive losses; (ii) significantly improve the sample-efficiency and final performance of the agents. We hope our simple and effective approach will serve as a strong baseline for future research incorporating attention mechanisms in reinforcement learning and control.

****

Zero-Shot Recognition through Image-Guided Semantic Classification

Mei-Chen Yeh,Fang Li,Bo-Heng Li

We present a new visual-semantic embedding method for generalized zero-shot learning. Existing embedding-based methods aim to learn the correspondence between an image classifier (visual representation) and its class prototype (semantic representation) for each class. Inspired by the binary relevance method for multi-label classification, we learn the mapping between an image and its semantic classifier. Given an input image, the proposed Image-Guided Semantic Classification (IGSC) method creates a label classifier, being applied to all label embeddings to determine whether a label belongs to the input image. Therefore, a semantic classifier is image conditioned and is generated during inference. We also show that IGSC is a unifying framework for two state-of-the-art deep-embedding methods. We validate our approach with four standard benchmark datasets.

****

A Unified Spectral Sparsification Framework for Directed Graphs

ying zhang,Zhiqiang Zhao,Zhuo Feng

Recent spectral graph sparsification research allows constructing nearly-linear-sized subgraphs that can well preserve the spectral (structural) properties of the original graph, such as the first few eigenvalues and eigenvectors of the graph Laplacian, leading to the development of a variety of nearly-linear time numerical and graph algorithms. However, there is not a unified approach that allows for truly scalable spectral sparsification of both directed and undirected graphs. For the first time, we prove the existence of linear-sized spectral sparsifiers for general directed
graphs and introduce a practically-efficient and unified spectral graph sparsification approach that allows sparsifying real-world, large-scale directed and undirected graphs with guaranteed preservation of the original graph spectra. By exploiting a highly-scalable (nearly-linear complexity) spectral matrix perturbation analysis framework for constructing nearly-linear sized (directed) subgraphs, it enables us to well preserve the key eigenvalues and eigenvectors of the original (directed) graph
Laplacians. The proposed method has been validated using various kinds of directed graphs obtained from public domain sparse matrix collections, showing promising results for solving directed graph Laplacians, spectral embedding, and partitioning of general directed graphs, as well as approximately computing (personalized) PageRank vectors.

****
Joint Perception and Control as Inference with an Object-based Implementation
Minne Li,Zheng Tian,Pranav Nashikkar,Ian Davies,Ying Wen,Jun Wang
Existing model-based reinforcement learning methods often study perception modeling and decision making separately. We introduce joint Perception and Control as Inference (PCI), a general framework to combine perception and control for partially observable environments through Bayesian inference. Based on the fact that object-level inductive biases are critical in human perceptual learning and reasoning, we propose Object-based Perception Control (OPC), an instantiation of PCI which manages to facilitate control using automatic discovered object-based representations. We develop an unsupervised end-to-end solution and analyze the convergence of the perception model update. Experiments in a high-dimensional pixel environment demonstrate the learning effectiveness of our object-based perception control approach. Specifically, we show that OPC achieves good perceptual grouping quality and outperforms several strong baselines in accumulated rewards.
****
Memory Optimization for Deep Networks
Aashaka Shah,Chao-Yuan Wu,Jayashree Mohan,Vijay Chidambaram,Philipp Kraehenbuehl
Deep learning is slowly, but steadily, hitting a memory bottleneck. While the tensor computation in top-of-the-line GPUs increased by $32\times$ over the last five years, the total available memory only grew by $2.5\times$. This prevents researchers from exploring larger architectures, as training large networks requires more memory for storing intermediate outputs. In this paper, we present MONeT, an automatic framework that minimizes both the memory footprint and computational overhead of deep networks. MONeT jointly optimizes the checkpointing schedule and the implementation of various operators. MONeT is able to outperform all prior hand-tuned operations as well as automated checkpointing. MONeT reduces the overall memory requirement by $3\times$ for various PyTorch models, with a 9-16$\%$ overhead in computation. For the same computation cost, MONeT requires 1.2-1.8$\times$ less memory than current state-of-the-art automated checkpointing frameworks. Our code will be made publicly available upon acceptance.
****
Parrot: Data-Driven Behavioral Priors for Reinforcement Learning
Avi Singh,Huihan Liu,Gaoyue Zhou,Albert Yu,Nicholas Rhinehart,Sergey Levine
Reinforcement learning provides a general framework for flexible decision making and control, but requires extensive data collection for each new task that an agent needs to learn. In other machine learning fields, such as natural language processing or computer vision, pre-training on large, previously collected datasets to bootstrap learning for new tasks has emerged as a powerful paradigm to reduce data requirements when learning a new task. In this paper, we ask the following question: how can we enable similarly useful pre-training for RL agents? We propose a method for pre-training behavioral priors that can capture complex input-output relationships observed in successful trials from a wide range of previously seen tasks, and we show how this learned prior can be used for rapidly learning new tasks without impeding the RL agent's ability to try out novel behaviors. We demonstrate the effectiveness of our approach in challenging robotic manipulation domains involving image observations and sparse reward functions, where our method outperforms prior works by a substantial margin. Additional materials can be found on our project website: https://sites.google.com/view/parrot-rl
****
Early Stopping in Deep Networks: Double Descent and How to Eliminate it
Reinhard Heckel,Fatih Furkan Yilmaz
Over-parameterized models, such as large deep networks, often exhibit a double descent phenomenon, whereas a function of model size, error first decreases, increases, and decreases at last. This intriguing double descent behavior also occurs as a function of training epochs and has been conjectured to arise because training epochs control the model complexity. In this paper, we show that such epoch-wise double descent occurs for a different reason: It is caused by a superposition of two or more bias-variance tradeoffs that arise because different parts of the network are learned at different epochs, and mitigating this by proper sca

ling of stepsizes can significantly improve the early stopping performance. We s
how this analytically for i) linear regression, where differently scaled feature
s give rise to a superposition of bias-variance tradeoffs, and for ii) a wide tw
o-layer neural network, where the first and second layers govern bias-variance t
radeoffs. Inspired by this theory, we study two standard convolutional networks
empirically and show that eliminating epoch-wise double descent through adjustin
g stepsizes of different layers improves the early stopping performance.
****

TOWARDS NATURAL ROBUSTNESS AGAINST ADVERSARIAL EXAMPLES

Haoyu Chu,Shikui Wei,Yao Zhao

Recent studies have shown that deep neural networks are vulnerable to adversaria
l examples, but most of the methods proposed to defense adversarial examples can
not solve this problem fundamentally. In this paper, we theoretically prove that
 there is an upper bound for neural networks with identity mappings to constrain
 the error caused by adversarial noises. However, in actual computations, this k
ind of neural network no longer holds any upper bound and is therefore susceptib
le to adversarial examples. Following similar procedures, we explain why adversa
rial examples can fool other deep neural networks with skip connections. Further
more, we demonstrate that a new family of deep neural networks called Neural ODE
s (Chen et al., 2018) holds a weaker upper bound. This weaker upper bound preven
ts the amount of change in the result from being too large. Thus, Neural ODEs ha
ve natural robustness against adversarial examples. We evaluate the performance
of Neural ODEs compared with ResNet under three white-box adversarial attacks (F
GSM, PGD, DI2-FGSM) and one black-box adversarial attack (Boundary Attack). Fina
lly, we show that the natural robustness of Neural ODEs is even better than the
robustness of neural networks that are trained with adversarial training methods
, such as TRADES and YOPO.

****

R-MONet: Region-Based Unsupervised Scene Decomposition and Representation via Co
nsistency of Object Representations

Shengxin Qian

Decomposing a complex scene into multiple objects is a natural instinct of an in
telligent vision system. Recently, the interest in unsupervised scene representa
tion learning emerged and many previous works tackle this by decomposing scenes
into object representations either in the form of segmentation masks or position
 and scale latent variables (i.e. bounding boxes). We observe that these two typ
es of representation both contain object geometric information and should be con
sistent with each other. Inspired by this observation, we provide an unsupervise
d generative framework called R-MONet that can generate objects geometric repres
entation in the form of bounding boxes and segmentation masks simultaneously. Wh
ile bounding boxes can represent the region of interest (ROI) for generating for
eground segmentation masks, the foreground segmentation masks can also be used t
o supervise bounding boxes learning with the Multi-Otsu Thresholding method. Thr
ough the experiments on CLEVR and Multi-dSprites datasets, we show that ensuring
 the consistency of two types of representation can help the model to decompose
the scene and learn better object geometric representations.
****

Contrastive Syn-to-Real Generalization

Wuyang Chen,Zhiding Yu,Shalini De Mello,Sifei Liu,Jose M. Alvarez,Zhangyang Wang
,Anima Anandkumar

Training on synthetic data can be beneficial for label or data-scarce scenarios.
 However, synthetically trained models often suffer from poor generalization in
real domains due to domain gaps. In this work, we make a key observation that th
e diversity of the learned feature embeddings plays an important role in the gen
eralization performance. To this end, we propose contrastive synthetic-to-real g
eneralization (CSG), a novel framework that leverage the pre-trained ImageNet kn
owledge to prevent overfitting to the synthetic domain, while promoting the dive
rsity of feature embeddings as an inductive bias to improve generalization. In a
ddition, we enhance the proposed CSG framework with attentional pooling (A-pool)

to let the model focus on semantically important regions and further improve it s generalization. We demonstrate the effectiveness of CSG on various synthetic t raining tasks, exhibiting state-of-the-art performance on zero-shot domain gener alization.
****

Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairn ess Metrics

Charan Reddy,Soroush Mehri,Deepak Sharma,Samira Shabanian,Sina Honari

With the recent expanding attention of machine learning researchers and practiti oners to fairness, there is a void of a common framework to analyze and compare the capabilities of proposed models in deep representation learning. In this pap er, we evaluate different fairness methods trained with deep neural networks on a common synthetic dataset to obtain a better insight into the working of these methods. In particular, we train about 2000 different models in various setups, including unbalanced and correlated data configurations, to verify the limits of the current models and better understand in which setups they are subject to fa ilure. In doing so we present a dataset, a large subset of proposed fairness met rics in the literature, and rigorously evaluate recent promising debiasing algor ithms in a common framework hoping the research community would take this benchm ark as a common entry point for fair deep learning.
****

Benchmarks for Deep Off-Policy Evaluation

Justin Fu,Mohammad Norouzi,Ofir Nachum,George Tucker,ziyu wang,Alexander Novikov ,Mengjiao Yang,Michael R Zhang,Yutian Chen,Aviral Kumar,Cosmin Paduraru,Sergey L evine,Thomas Paine

Off-policy evaluation (OPE) holds the promise of being able to leverage large, o ffline datasets for both evaluating and selecting complex policies for decision making. The ability to learn offline is particularly important in many real-worl d domains, such as in healthcare, recommender systems, or robotics, where online data collection is an expensive and potentially dangerous process. Being able t o accurately evaluate and select high-performing policies without requiring onli ne interaction could yield significant benefits in safety, time, and cost for th ese applications. While many OPE methods have been proposed in recent years, com paring results between papers is difficult because currently there is a lack of a comprehensive and unified benchmark, and measuring algorithmic progress has be en challenging due to the lack of difficult evaluation tasks. In order to addres s this gap, we present a collection of policies that in conjunction with existin g offline datasets can be used for benchmarking off-policy evaluation. Our tasks include a range of challenging high-dimensional continuous control problems, wi th wide selections of datasets and policies for performing policy selection. The goal of our benchmark is to provide a standardized measure of progress that is motivated from a set of principles designed to challenge and test the limits of existing OPE methods. We perform an evaluation of state-of-the-art algorithms an d provide open-source access to our data and code to foster future research in t his area.
****

Multi-agent Deep FBSDE Representation For Large Scale Stochastic Differential G ames

Tianrong Chen,Ziyi Wang,Ioannis Exarchos,Evangelos Theodorou

In this paper we present a deep learning framework for solving large-scale mu lti-agent non-cooperative stochastic games using fictitious play. The Hamilton-J acobi-Bellman (HJB) PDE associated with each agent is reformulated into a set o f Forward-Backward Stochastic Differential Equations (FBSDEs) and solved via for ward sampling on a suitably defined neural network architecture. Decision-making in multi-agent systems suffers from the curse of dimensionality and strategy de generation as the number of agents and time horizon increase. We propose a novel Deep FBSDE controller framework which is shown to outperform the current state- of-the-art deep fictitious play algorithm on a high dimensional inter-bank lendi ng/borrowing problem. More importantly, our approach mitigates the curse of many agents and reduces computational and memory complexity, allowing us to scale up

to 1,000 agents in simulation, a scale which, to the best of our knowledge, represents a new state of the art. Finally, we showcase the framework's applicability in robotics on a belief-space autonomous racing problem.
****
Improving Sampling Accuracy of Stochastic Gradient MCMC Methods via Non-uniform Subsampling of Gradients
Ruilin Li,Xin Wang,Hongyuan Zha,Molei Tao
Common Stochastic Gradient MCMC methods approximate gradients by stochastic ones via uniformly subsampled data points. A non-uniform subsampling scheme, however, can reduce the variance introduced by the stochastic approximation and make the sampling of a target distribution more accurate. For this purpose, an exponentially weighted stochastic gradient approach (EWSG) is developed to match the transition kernel of a non-uniform-SG-MCMC method with that of a batch-gradient-MCMC method. If needed to be put in the importance sampling (IS) category, EWSG can be viewed as a way to extend the IS+SG approach successful for optimization to the sampling setup. EWSG works for a range of MCMC methods, and a demonstration on Stochastic-Gradient 2nd-order Langevin is provided. In our practical implementation of EWSG, the non-uniform subsampling is performed efficiently via a Metropolis-Hasting chain on the data index, which is coupled to the sampling algorithm. The fact that our method has reduced local variance with high probability is theoretically analyzed. A non-asymptotic global error analysis is also presented. As a practical implementation contains hyperparameters, numerical experiments based on both synthetic and real world data sets are provided, to both demonstrate the empirical performances and recommend hyperparameter choices. Notably, while statistical accuracy has improved, the speed of convergence, with appropriately chosen hyper-parameters, was empirically observed to be at least comparable to the uniform version, which renders EWSG a practically useful alternative to common variance reduction treatments.
****
Pre-training Text-to-Text Transformers for Concept-centric Common Sense
Wangchunshu Zhou,Dong-Ho Lee,Ravi Kiran Selvam,Seyeon Lee,Xiang Ren
Pretrained language models (PTLM) have achieved impressive results in a range of natural language understanding (NLU) and generation (NLG) tasks that require a syntactic and semantic understanding of the text. However, current pre-training objectives such as masked token prediction (for BERT-style PTLMs) and masked span infilling (for T5-style PTLMs) do not explicitly model the relational and compositional commonsense knowledge about everyday concepts, which is crucial to many downstream tasks requiring commonsense reasoning. To augment PTLMs with common sense, we propose generative and contrastive objectives as intermediate self-supervised pre-training tasks between general pre-training and downstream task-specific fine-tuning. We also propose a joint training framework to unify generative and contrastive objectives so that these objectives can be more effective.
Our proposed objectives can pack more commonsense knowledge into the parameters of a pre-trained text-to-text transformer without relying on external knowledge bases, yielding better performance on both NLU and NLG tasks. We apply our method on a pre-trained T5 model in an intermediate task transfer learning fashion to train a concept-aware language model (CALM) and experiment with five commonsense benchmarks (four NLU tasks and one NLG task). Experimental results show that CALM outperforms baseline methods by a consistent margin.
****
Combining Label Propagation and Simple Models out-performs Graph Neural Networks
Qian Huang,Horace He,Abhay Singh,Ser-Nam Lim,Austin Benson
Graph Neural Networks (GNNs) are a predominant technique for learning over graphs. However, there is relatively little understanding of why GNNs are successful in practice and whether they are necessary for good performance. Here, we show that for many standard transductive node classification benchmarks, we can exceed or match the performance of state-of-the-art GNNs by combining shallow models that ignore the graph structure with two simple post-processing steps that exploit correlation in the label structure: (i) an “error correlation” that spreads residual errors in training data to correct errors in test data and (ii) a “predic

tion correlation” that smooths the predictions on the test data. We call this overall procedure Correct and Smooth (C&S), and the post-processing steps are implemented via simple modifications to standard label propagation techniques that have long been used in graph-based semi-supervised learning. Our approach exceeds or nearly matches the performance of state-of-the-art GNNs on a wide variety of benchmarks, with just a small fraction of the parameters and orders of magnitude faster runtime. For instance, we exceed the best-known GNN performance on the OGB-Products dataset with 137 times fewer parameters and greater than 100 times less training time. The performance of our methods highlights how directly incorporating label information into the learning algorithm (as is common in traditional methods) yields easy and substantial performance gains. We can also incorporate our techniques into big GNN models, providing modest gains in some cases.
****

Success-Rate Targeted Reinforcement Learning by Disorientation Penalty
Haichuan Gao,Zhile Yang,Tian Tan,Feng Chen
Current reinforcement learning generally uses discounted return as its learning objective. However, real-world tasks may often demand a high success rate, which can be quite different from optimizing rewards. In this paper, we explicitly formulate the success rate as an undiscounted form of return with {0, 1}-binary reward function. Unfortunately, applying traditional Bellman updates to value function learning can be problematic for learning undiscounted return, and thus not suitable for optimizing success rate. From our theoretical analysis, we discover that values across different states tend to converge to the same value, resulting in the agent wandering around those states without making any actual progress. This further leads to reduced learning efficiency and inability to complete a task in time. To combat the aforementioned issue, we propose a new method, which introduces Loop Penalty (LP) into value function learning, to penalize disoriented cycling behaviors in the agent's decision-making. We demonstrate the effectiveness of our proposed LP on three environments, including grid-world cliff-walking, Doom first-person navigation and robot arm control, and compare our method with Q-learning, Monte-Carlo and Proximal Policy Optimization (PPO). Empirically, LP improves the convergence of training and achieves a higher success rate.
****

Decorrelated Double Q-learning
GANG CHEN
Q-learning with value function approximation may have the poor performance because of overestimation bias and imprecise estimate. Specifically, overestimation bias is from the maximum operator over noise estimate, which is exaggerated using the estimate of a subsequent state. Inspired by the recent advance of deep reinforcement learning and Double Q-learning, we introduce the decorrelated double Q-learning (D2Q). Specifically, we introduce the decorrelated regularization item to reduce the correlation between value function approximators, which can lead to less biased estimation and low variance. The experimental results on a suite of MuJoCo continuous control tasks demonstrate that our decorrelated double Q-learning can effectively improve the performance.
****

Graph Learning via Spectral Densification
Zhuo Feng,Yongyu Wang,Zhiqiang Zhao
Graph learning plays important role in many data mining and machine learning tasks, such as manifold learning, data representation and analysis, dimensionality reduction, data clustering, and visualization, etc. For the first time, we present a highly-scalable spectral graph densification approach (GRASPEL) for graph learning from data. By limiting the precision matrix to be a graph-Laplacian-like matrix in graphical Lasso, our approach aims to learn ultra-sparse undirected graphs from potentially high-dimensional input data. A very unique property of the graphs learned by GRASPEL is that the spectral embedding (or approximate effective-resistance) distances on the graph will encode the similarities between the original input data points. By interleaving the latest high-performance nearly-linear
time spectral methods, ultrasparse yet spectrally-robust graphs can be learned b

y identifying and including the most spectrally-critical edges into the graph. C
ompared with prior state-of-the-art graph learning approaches, GRASPEL is more s
calable and allows substantially improving computing efficiency and solution qua
lity of a variety of data mining and
machine learning applications, such as manifold learning, spectral clustering (S
C), and dimensionality reduction.
****

Learning Long-term Visual Dynamics with Region Proposal Interaction Networks
Haozhi Qi,Xiaolong Wang,Deepak Pathak,Yi Ma,Jitendra Malik
Learning long-term dynamics models is the key to understanding physical common s
ense. Most existing approaches on learning dynamics from visual input sidestep l
ong-term predictions by resorting to rapid re-planning with short-term models. T
his not only requires such models to be super accurate but also limits them only
 to tasks where an agent can continuously obtain feedback and take action at eac
h step until completion. In this paper, we aim to leverage the ideas from succes
s stories in visual recognition tasks to build object representations that can c
apture inter-object and object-environment interactions over a long range. To th
is end, we propose Region Proposal Interaction Networks (RPIN), which reason abo
ut each object's trajectory in a latent region-proposal feature space. Thanks to
 the simple yet effective object representation, our approach outperforms prior
methods by a significant margin both in terms of prediction quality and their ab
ility to plan for downstream tasks, and also generalize well to novel environmen
ts. Code, pre-trained models, and more visualization results are available at ht
tps://haozhi.io/RPIN.
****

Chaos of Learning Beyond Zero-sum and Coordination via Game Decompositions
Yun Kuen Cheung,Yixin Tao
It is of primary interest for ML to understand how agents learn and interact dyn
amically in competitive environments and games (e.g. GANs). But this has been a
difficult task, as irregular behaviors are commonly observed in such systems. Th
is can be explained theoretically, for instance, by the works of Cheung and Pili
ouras (COLT 2019; NeurIPS 2020), which showed that in two-person zero-sum games,
 if agents employ one of the most well-known learning algorithms, Multiplicative
 Weights Update (MWU), then Lyapunov chaos occurs everywhere in the payoff space
. In this paper, we study how persistent chaos can occur in the more general nor
mal game settings, where the agents might have the motivation to coordinate (whi
ch is not true for zero-sum games) and the number of agents can be arbitrary.

We characterize bimatrix games where MWU, its optimistic variant (OMWU) or Follo
w-the-Regularized-Leader (FTRL) algorithms are Lyapunov chaotic almost everywher
e in the payoff space. Technically, our characterization is derived by extending
 the volume-expansion argument of Cheung and Piliouras via the canonical game de
composition into zero-sum and coordination components. Interestingly, the two co
mponents induce opposite volume-changing behaviors, so the overall behavior can
be analyzed by comparing the strengths of the components against each other. The
 comparison is done via our new notion of "matrix domination" or via a linear pr
ogram. For multi-player games, we present a local equivalence of volume change b
etween general games and graphical games, which is used to perform volume and ch
aos analyses of MWU and OMWU in potential games.
****

DIET-SNN: A Low-Latency Spiking Neural Network with Direct Input Encoding & Leak
age and Threshold Optimization
Nitin Rathi,Kaushik Roy
Bio-inspired spiking neural networks (SNNs), operating with asynchronous binary
signals (or spikes) distributed over time, can potentially lead to greater compu
tational efficiency on event-driven hardware. The state-of-the-art SNNs suffer f
rom high inference latency, resulting from inefficient input encoding, and sub-o
ptimal settings of the neuron parameters (firing threshold, and membrane leak).
We propose DIET-SNN, a low latency deep spiking network that is trained with gra
dient descent to optimize the membrane leak and the firing threshold along with

other network parameters (weights). The membrane leak and threshold for each lay er of the SNN are optimized with end-to-end backpropagation to achieve competiti ve accuracy at reduced latency. The analog pixel values of an image are directly applied to the input layer of DIET-SNN without the need to convert to spike-tra in. The first convolutional layer is trained to convert inputs into spikes where leaky-integrate-and-fire (LIF) neurons integrate the weighted inputs and genera te an output spike when the membrane potential crosses the trained firing thresh old. The trained membrane leak controls the flow of input information and attenu ates irrelevant inputs to increase the activation sparsity in the convolutional and linear layers of the network. The reduced latency combined with high activat ion sparsity provides large improvements in computational efficiency. We evaluat e DIET-SNN on image classification tasks from CIFAR and ImageNet datasets on VGG and ResNet architectures. We achieve top-1 accuracy of 69% with 5 timesteps (in ference latency) on the ImageNet dataset with 12x less compute energy than an eq uivalent standard ANN. Additionally, DIET-SNN performs 20-500x faster inference compared to other state-of-the-art SNN models.

****

Transformers with Competitive Ensembles of Independent Mechanisms

Alex Lamb,Di He,Anirudh Goyal,Guolin Ke,Chien-Feng Liao,Mirco Ravanelli,Yoshua B engio

An important development in deep learning from the earliest MLPs has been a move towards architectures with structural inductive biases which enable the model t o keep distinct sources of information and routes of processing well-separated. This structure is linked to the notion of independent mechanisms from the causa lity literature, in which a mechanism is able to retain the same processing as i rrelevant aspects of the world are changed. For example, convnets enable separa tion over positions, while attention-based architectures (especially Transformer s) learn which combination of positions to process dynamically. In this work we explore a way in which the Transformer architecture is deficient: it represents each position with a large monolithic hidden representation and a single set of parameters which are applied over the entire hidden representation. This poten tially throws unrelated sources of information together, and limits the Transfor mer's ability to capture independent mechanisms. To address this, we propose Tr ansformers with Independent Mechanisms (TIM), a new Transformer layer which divi des the hidden representation and parameters into multiple mechanisms, which onl y exchange information through attention. Additionally, we propose a competitio n mechanism which encourages these mechanisms to specialize over time steps, and thus be more independent. We study TIM on a large scale BERT model, on the Ima ge Transformer, and on speech enhancement and find evidence for semantically mea ningful specialization as well as improved performance.

****

The Quenching-Activation Behavior of the Gradient Descent Dynamics for Two-layer Neural Network Models

Chao Ma,Lei Wu,Weinan E

A numerical and phenomenological study of the gradient descent (GD) algorithm f or training two-layer neural network models is carried out for different paramet er regimes. It is found that there are two distinctive phases in the GD dynamics in the under-parameterized regime: An early phase in which the GD dynamics foll ows closely that of the corresponding random feature model, followed by a late p hase in which the neurons are divided into two groups: a group of a few (maybe n one) “activated” neurons that dominate the dynamics and a group of ``quenched” n eurons that support the continued activation and deactivation process. In parti cular, when the target function can be accurately approximated by a relatively s mall number of neurons, this quenching-activation process biases GD to picking s parse solutions. This neural network-like behavior is continued into the mildly over-parameterized regime, in which it undergoes a transition to a random featu re-like behavior where the inner-layer parameters are effectively frozen during the training process. The quenching process seems to provide a clear mechanism for ``implicit regularization''. This is qualitatively different from the GD dy namics associated with the ``mean-field'' scaling where all neurons participate

equally.
****
Differentially Private Synthetic Data: Applied Evaluations and Enhancements
Lucas Rosenblatt,Xiaoyan Liu,Samira Pouyanfar,Eduardo de Leon,Anuj Desai,Joshua Allen
Machine learning practitioners frequently seek to leverage the most informative available data, without violating the data owner's privacy, when building predictive models. Differentially private data synthesis protects personal details from exposure, and allows for the training of differentially private machine learning models on privately generated datasets. But how can we effectively assess the efficacy of differentially private synthetic data? In this paper, we survey four differentially private generative adversarial networks for data synthesis. We evaluate each of them at scale on five standard tabular datasets, and in two applied industry scenarios. We benchmark with novel metrics from recent literature and other standard machine learning tools. Our results suggest some synthesizers are more applicable for different privacy budgets, and we further demonstrate complicating domain-based tradeoffs in selecting an approach. We offer experimental learning on applied machine learning scenarios with private internal data to researchers and practitioners alike. In addition, we propose QUAIL, a two model hybrid approach to generating synthetic data. We examine QUAIL's tradeoffs, and note circumstances in which it outperforms baseline differentially private supervised learning models under the same budget constraint.
****
Control-Aware Representations for Model-based Reinforcement Learning
Brandon Cui,Yinlam Chow,Mohammad Ghavamzadeh
A major challenge in modern reinforcement learning (RL) is efficient control of dynamical systems from high-dimensional sensory observations. Learning controllable embedding (LCE) is a promising approach that addresses this challenge by embedding the observations into a lower-dimensional latent space, estimating the latent dynamics, and utilizing it to perform control in the latent space. Two important questions in this area are how to learn a representation that is amenable to the control problem at hand, and how to achieve an end-to-end framework for representation learning and control. In this paper, we take a few steps towards addressing these questions. We first formulate a LCE model to learn representations that are suitable to be used by a policy iteration style algorithm in the latent space.We call this model control-aware representation learning(CARL). We derive a loss function and three implementations for CARL. In the offline implementation, we replace the locally-linear control algorithm (e.g., iLQR) used by the existing LCE methods with a RL algorithm, namely model-based soft actor-critic, and show that it results in significant improvement. In online CARL, we interleave representation learning and control, and demonstrate further gain in performance. Finally, we propose value-guided CARL, a variation in which we optimize a weighted version of the CARL loss function, where the weights depend on the TD-error of the current policy. We evaluate the proposed algorithms by extensive experiments on benchmark tasks and compare them with several LCE baselines.
****
Provably robust classification of adversarial examples with detection
Fatemeh Sheikholeslami,Ali Lotfi,J Zico Kolter
Adversarial attacks against deep networks can be defended against either by building robust classifiers or, by creating classifiers that can \emph{detect} the presence of adversarial perturbations. Although it may intuitively seem easier to simply detect attacks rather than build a robust classifier, this has not bourne out in practice even empirically, as most detection methods have subsequently been broken by adaptive attacks, thus necessitating \emph{verifiable} performance for detection mechanisms. In this paper, we propose a new method for jointly training a provably robust classifier and detector. Specifically, we show that by introducing an additional "abstain/detection" into a classifier, we can modify existing certified defense mechanisms to allow the classifier to either robustly classify \emph{or} detect adversarial attacks. We extend the common interval bound propagation (IBP) method for certified robustness under ℓ_∞ pe

rturbations to account for our new robust objective, and show that the method ou tperforms traditional IBP used in isolation, especially for large perturbation s izes. Specifically, tests on MNIST and CIFAR-10 datasets exhibit promising resu lts, for example with provable robust error less than 63.63% and 67.92%, f or 55.6% and 66.37% natural error, for $\epsilon=8/255$ and $16/255$ on th e CIFAR-10 dataset, respectively.

****
Deep Ensemble Kernel Learning
Devanshu Agrawal,Jacob D Hinkle
Gaussian processes (GPs) are nonparametric Bayesian models that are both flexibl e and robust to overfitting. One of the main challenges of GP methods is selecti ng the kernel. In the deep kernel learning (DKL) paradigm, a deep neural network or ``feature network'' is used to map inputs into a latent feature space, where  a GP with a ``base kernel'' acts; the resulting model is then trained in an end -to-end fashion. In this work, we introduce the ``deep ensemble kernel learning' ' (DEKL) model, which is a special case of DKL. In DEKL, a linear base kernel is used, enabling exact optimization of the base kernel hyperparameters and a scal able inference method that does not require approximation by inducing points. We also represent the feature network as a concatenation of an ensemble of learner networks with a common architecture, allowing for easy model parallelism. We sh ow that DEKL is able to approximate any kernel if the number of learners in the ensemble is arbitrarily large. Comparing the DEKL model to DKL and deep ensemble (DE) baselines on both synthetic and real-world regression tasks, we find that DEKL often outperforms both baselines in terms of predictive performance and tha t the DEKL learners tend to be more diverse (i.e., less correlated with one anot her) compared to the DE learners.
****
Return-Based Contrastive Representation Learning for Reinforcement Learning
Guoqing Liu,Chuheng Zhang,Li Zhao,Tao Qin,Jinhua Zhu,Li Jian,Nenghai Yu,Tie-Yan Liu
Recently, various auxiliary tasks have been proposed to accelerate representatio n learning and improve sample efficiency in deep reinforcement learning (RL). Ho wever, existing auxiliary tasks do not take the characteristics of RL problems i nto consideration and are unsupervised. By leveraging returns, the most importan t feedback signals in RL, we propose a novel auxiliary task that forces the lear nt representations to discriminate state-action pairs with different returns. Ou r auxiliary loss is theoretically justified to learn representations that captur e the structure of a new form of state-action abstraction, under which state-act ion pairs with similar return distributions are aggregated together. Empirically , our algorithm outperforms strong baselines on complex tasks in Atari games and DeepMind Control suite, and achieves even better performance when combined with existing auxiliary tasks.
****
Meta-learning Transferable Representations with a Single Target Domain
Hong Liu,Jeff Z. HaoChen,Colin Wei,Tengyu Ma
Recent works found that fine-tuning and joint training---two popular approaches for transfer learning---do not always improve accuracy on downstream tasks. Firs t, we aim to understand more about when and why fine-tuning and joint training c an be suboptimal or even harmful for transfer learning. We design semi-synthetic datasets where the source task can be solved by either source-specific features or transferable features. We observe that (1) pre-training may not have incenti ve to learn transferable features and (2) joint training may simultaneously lear n source-specific features and overfit to the target. Second, to improve over f ine-tuning and joint training, we propose Meta Representation Learning MeRLin to learn transferable features. MeRLin meta-learns representations by ensuring tha t a head fit on top of the representations with target training data also perfor ms well on target validation data. We also prove that MeRLin recovers the target ground-truth model with a quadratic neural net parameterization and a source di stribution that contains both transferable and source-specific features. On the

same distribution, pre-training and joint training provably fail to learn trans ferable features. MeRLin empirically outperforms previous state-of-the-art trans fer learning algorithms on various real-world vision and NLP transfer learning benchmarks.
****
Double Q-learning: New Analysis and Sharper Finite-time Bound
Lin Zhao,Huaqing Xiong,Yingbin Liang,Wei Zhang
Double Q-learning \citep{hasselt2010double} has gained significant success in pr actice due to its effectiveness in overcoming the overestimation issue of Q-lear ning. However, theoretical understanding of double Q-learning is rather limited and the only existing finite-time analysis was recently established in \citet{xi ong2020double} under a polynomial learning rate. This paper analyzes the more ch allenging case with a rescaled linear/constant learning rate for which the previ ous method does not appear to be applicable. We develop new analytical tools tha t achieve an order-level better finite-time convergence rate than the previously established result. Specifically, we show that synchronous double Q-learning at tains an ϵ-accurate global optimum with a time complexity of $\Omega\le ft(\frac{\ln D}{(1-\gamma)^7\epsilon^2} \right)$, and the asynchronous algorith m attains a time complexity of $\tilde{\Omega}\left(\frac{L}{(1-\gamma)^7\epsilo n^2} \right)$, where D is the cardinality of the state-action space, γ is the discount factor, and L is a parameter related to the sampling strategy for asynchronous double Q-learning. These results improve the order-level depend ence of the convergence rate on all major parameters $(\epsilon,1-\gamma, D, L)$ provided in \citet{xiong2020double}. The new analysis in this paper presents a more direct and succinct approach for characterizing the finite-time convergence rate of double Q-learning.
****
Information Theoretic Meta Learning with Gaussian Processes
Michalis Titsias,Sotirios Nikoloutsopoulos,Alexandre Galashov
We formulate meta learning using information theoretic concepts such as mutual i nformation and the information bottleneck. The idea is to learn a stochastic rep resentation or encoding of the task description, given by a training or support set, that is highly informative about predicting the validation set. By making u se of variational approximations to the mutual information, we derive a general and tractable framework for meta learning. We particularly develop new memory-b ased meta learning algorithms based on Gaussian processes and derive extensions that combine memory and gradient-based meta learning. We demonstrate our method on few-shot regression and classification by using standard benchmarks such as O mniglot, mini-Imagenet and Augmented Omniglot.

****
Contrastive Learning of Medical Visual Representations from Paired Images and Te xt
Yuhao Zhang,Hang Jiang,Yasuhide Miura,Christopher D Manning,Curtis Langlotz
Learning visual representations of medical images is core to medical image under standing but its progress has been held back by the small size of hand-labeled d atasets. Existing work commonly relies on transferring weights from ImageNet pre training, which is suboptimal due to drastically different image characteristics , or rule-based label extraction from the textual report data paired with medica l images, which is inaccurate and hard to generalize. We propose an alternative unsupervised strategy to learn medical visual representations directly from the naturally occurring pairing of images and textual data. Our method of pretrainin g medical image encoders with the paired text data via a bidirectional contrasti ve objective between the two modalities is domain-agnostic, and requires no addi tional expert input. We test our method by transferring our pretrained weights t o 4 medical image classification tasks and 2 zero-shot retrieval tasks, and show that our method leads to image representations that considerably outperform str ong baselines in most settings. Notably, in all 4 classification tasks, our meth od requires only 10% as much labeled training data as an ImageNet initialized co unterpart to achieve better or comparable performance, demonstrating superior da

ta efficiency.
****
Self-Supervised Time Series Representation Learning by Inter-Intra Relational Re
asoning
Haoyi Fan,Fengbin Zhang,Yue Gao
Self-supervised learning achieves superior performance in many domains by extrac
ting useful representations from the unlabeled data. However, most of traditiona
l self-supervised methods mainly focus on exploring the inter-sample structure w
hile less efforts have been concentrated on the underlying intra-temporal struct
ure, which is important for time series data. In this paper, we present SelfTime
: a general self-supervised time series representation learning framework, by ex
ploring the inter-sample relation and intra-temporal relation of time series to
learn the underlying structure feature on the unlabeled time series. Specificall
y, we first generate the inter-sample relation by sampling positive and negative
 samples of a given anchor sample, and intra-temporal relation by sampling time
pieces from this anchor. Then, based on the sampled relation, a shared feature e
xtraction backbone combined with two separate relation reasoning heads are emplo
yed to quantify the relationships of the sample pairs for inter-sample relation
reasoning, and the relationships of the time piece pairs for intra-temporal rela
tion reasoning, respectively. Finally, the useful representations of time series
 are extracted from the backbone under the supervision of relation reasoning hea
ds. Experimental results on multiple real-world time series datasets for time se
ries classification task demonstrate the effectiveness of the proposed method. C
ode and data are publicly available.
****
Unified Principles For Multi-Source Transfer Learning Under Label Shifts
changjian shui,Zijian Li,jiaqi li,Christian Gagné,Charles Ling,Boyu Wang
We study the label shift problem in multi-source transfer learning and derive ne
w generic principles. Our proposed framework unifies the principles of condition
al feature alignment, label distribution ratio estimation, and domain relation w
eights estimation. Based on inspired practical principles, we provide a unified
practical framework for three multi-source label shift transfer scenarios: learn
ing with limited target data, unsupervised domain adaptation, and label partial
unsupervised domain adaptation. We evaluate the proposed method on these scenari
os by extensive experiments and show that our proposed algorithm can significant
ly outperform the baselines.
****
Generalisation Guarantees For Continual Learning With Orthogonal Gradient Descen
t
Mehdi Abbana Bennani,Thang Doan,Masashi Sugiyama
In Continual Learning settings, deep neural networks are prone to Catastrophic F
orgetting. Orthogonal Gradient Descent (Farajtabar et al., 2019) was proposed to
 tackle the challenge. However, no theoretical guarantees have been proven yet.
We present a theoretical framework to study Continual Learning algorithms in the
 NTK regime. This framework comprises closed form expression of the model throug
h tasks and proxies for transfer learning, generalisation and tasks similarity.
In this framework, we prove that OGD is robust to Catastrophic Forgetting then d
erive the first generalisation bound for SGD and OGD for Continual Learning. Fin
ally, we study the limits of this framework in practice for OGD and highlight th
e importance of the NTK variation for Continual Learning.
****
Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classific
ation
Francisco Utrera,Evan Kravitz,N. Benjamin Erichson,Rajiv Khanna,Michael W. Mahon
ey
Transfer learning has emerged as a powerful methodology for adapting pre-trained
 deep neural networks on image recognition tasks to new domains. This process co
nsists of taking a neural network pre-trained on a large feature-rich source dat
aset, freezing the early layers that encode essential generic image properties,
and then fine-tuning the last few layers in order to capture specific informatio

n related to the target situation. This approach is particularly useful when onl y limited or weakly labeled data are available for the new task. In this work, w e demonstrate that adversarially-trained models transfer better than non-adversa rially-trained models, especially if only limited data are available for the new domain task. Further, we observe that adversarial training biases the learnt re presentations to retaining shapes, as opposed to textures, which impacts the tra nsferability of the source models. Finally, through the lens of influence functi ons, we discover that transferred adversarially-trained models contain more huma n-identifiable semantic information, which explains -- at least partly -- why ad versarially-trained models transfer better.
****

Multi-Representation Ensemble in Few-Shot Learning
Qing Chen,Jian Zhang
Deep neural networks (DNNs) compute representations in a layer by layer fashion, producing a final representation at the top layer of the pipeline, and classifi cation or regression is made using the final representation. A number of DNNs (e .g., ResNet, DenseNet) have shown that representations from the earlier layers c an be beneficial. They improved performance by aggregating representations from different layers. In this work, we asked the question, besides forming an aggreg ation, whether these representations can be utilized directly with the classific ation layer(s) to obtain better performance. We started our quest to the answer by investigating the classifiers based on the representations from different lay ers and observed that these classifiers were diverse and many of their decisions were complementary to each other, hence having the potential to generate a bett er overall decision when combined. Following this observation, we propose an ens emble method that creates an ensemble of classifiers, each taking a representati on from a different depth of a base DNN as the input. We tested this ensemble me thod in the setting of few-shot learning. Experiments were conducted on the mini -ImageNet and tieredImageNet datasets which are commonly used in the evaluation of few-shot learning methods. Our ensemble achieves the new state-of-the-art res ults for both datasets, comparing to previous regular and ensemble approaches.
****

Learning Structural Edits via Incremental Tree Transformations
Ziyu Yao,Frank F. Xu,Pengcheng Yin,Huan Sun,Graham Neubig
While most neural generative models generate outputs in a single pass, the human creative process is usually one of iterative building and refinement. Recent wo rk has proposed models of editing processes, but these mostly focus on editing s equential data and/or only model a single editing pass. In this paper, we presen t a generic model for incremental editing of structured data (i.e. ''structural edits''). Particularly, we focus on tree-structured data, taking abstract syntax trees of computer programs as our canonical example. Our editor learns to itera tively generate tree edits (e.g. deleting or adding a subtree) and applies them to the partially edited data, thereby the entire editing process can be formulat ed as consecutive, incremental tree transformations. To show the unique benefits of modeling tree edits directly, we further propose a novel edit encoder for le arning to represent edits, as well as an imitation learning method that allows t he editor to be more robust. We evaluate our proposed editor on two source code edit datasets, where results show that, with the proposed edit encoder, our edit or significantly improves accuracy over previous approaches that generate the ed ited program directly in one pass. Finally, we demonstrate that training our edi tor to imitate experts and correct its mistakes dynamically can further improve its performance.
****

Cross-Attentional Audio-Visual Fusion for Weakly-Supervised Action Localization
Jun-Tae Lee,Mihir Jain,Hyoungwoo Park,Sungrack Yun
Temporally localizing actions in videos is one of the key components for video u nderstanding. Learning from weakly-labeled data is seen as a potential solution towards avoiding expensive frame-level annotations. Different from other works w hich only depend on visual-modality, we propose to learn richer audiovisual repr esentation for weakly-supervised action localization. First, we propose a multi-

stage cross-attention mechanism to collaboratively fuse audio and visual features, which preserves the intra-modal characteristics. Second, to model both foreground and background frames, we construct an open-max classifier that treats the background class as an open-set. Third, for precise action localization, we design consistency losses to enforce temporal continuity for the action class prediction, and also help with foreground-prediction reliability. Extensive experiments on two publicly available video-datasets (AVE and ActivityNet1.2) show that the proposed method effectively fuses audio and visual modalities, and achieves the state-of-the-art results for weakly-supervised action localization.
****

Active Feature Acquisition with Generative Surrogate Models
Yang Li,Junier Oliva
Many real-world situations allow for the acquisition of additional relevant information when making an assessment with limited or uncertain data. However, traditional ML approaches either require all features to be acquired beforehand or regard part of them as missing data that cannot be acquired. In this work, we propose models that perform active feature acquisition (AFA) to improve the prediction assessments at evaluation time. We formulate the AFA problem as a Markov decision process (MDP) and resolve it using reinforcement learning (RL). The AFA problem yields sparse rewards and contains a high-dimensional complicated action space. Thus, we propose learning a generative surrogate model that captures the complicated dependencies among input features to assess potential information gain from acquisitions. We also leverage the generative surrogate model to provide intermediate rewards and auxiliary information to the agent. Furthermore, we extend AFA in a task we coin active instance recognition (AIR) for the unsupervised case where the target variables are the unobserved features themselves and the goal is to collect information for a particular instance in a cost-efficient way. Empirical results demonstrate that our approach achieves considerably better performance than previous state of the art methods on both supervised and unsupervised tasks.
****

Decoy-enhanced Saliency Maps
Yang Young Lu,Wenbo Guo,Xinyu Xing,William Noble
Saliency methods can make deep neural network predictions more interpretable by identifying a set of critical features in an input sample, such as pixels that contribute most strongly to a prediction made by an image classifier. Unfortunately, recent evidence suggests that many saliency methods poorly perform, especially in situations where gradients are saturated, inputs contain adversarial perturbations, or predictions rely upon inter-feature dependence. To address these issues, we propose a framework that improves the robustness of saliency methods by following a two-step procedure. First, we introduce a perturbation mechanism that subtly varies the input sample without changing its intermediate representations. Using this approach, we can gather a corpus of perturbed data samples while ensuring that the perturbed and original input samples follow the same distribution. Second, we compute saliency maps for the perturbed samples and propose a new method to aggregate saliency maps. With this design, we offset the gradient saturation influence upon interpretation. From a theoretical perspective, we show that the aggregated saliency map not only captures inter-feature dependence but, more importantly, is robust against previously described adversarial perturbation methods. Following our theoretical analysis, we present experimental results suggesting that, both qualitatively and quantitatively, our saliency method outperforms existing methods, in a variety of applications.
****

Exploiting structured data for learning contagious diseases under incomplete testing
Maggie Makar,Lauren West,David Hooper,Eric Horvitz,Erica Shenoy,John Guttag
One of the ways that machine learning algorithms can help control the spread of an infectious disease is by building models that predict who is likely to get infected whether or not they display any symptoms, making them good candidates for preemptive isolation. In this work we ask: can we build reliable infection pred

iction models when the observed data is collected under limited, and biased testing that prioritizes testing symptomatic individuals? Our analysis suggests that under favorable conditions, incomplete testing might be sufficient to achieve relatively good out-of-sample prediction error. Favorable conditions occur when untested-infected individuals have sufficiently different characteristics from untested-healthy, and when the infected individuals are "potent", meaning they infect a large majority of their neighbors. We develop an algorithm that predicts infections, and show that it outperforms benchmarks on simulated data. We apply our model to data from a large hospital to predict Clostridioides difficile infections; a communicable disease that is characterized by asymptomatic (i.e., untested) carriers. Using a proxy instead of the unobserved untested-infected state, we show that our model outperforms benchmarks in predicting infections.
****

Are all outliers alike? On Understanding the Diversity of Outliers for Detecting OODs
Ramneet Kaur,Susmit Jha,Anirban Roy
Deep neural networks (DNNs) are known to produce incorrect predictions with very high confidence on out-of-distribution (OOD) inputs. This limitation is one of the key challenges in the adoption of deep learning models in high-assurance systems such as autonomous driving, air traffic management, and medical diagnosis. This challenge has received significant attention recently, and several techniques have been developed to detect inputs where the model's prediction cannot be trusted. These techniques use different statistical, geometric, or topological signatures. This paper presents a taxonomy of OOD outlier inputs based on their source and nature of uncertainty. We demonstrate how different existing detection approaches fail to detect certain types of outliers. We utilize these insights to develop a novel integrated detection approach that uses multiple attributes corresponding to different types of outliers. Our results include experiments on CIFAR10, SVHN and MNIST as in-distribution data and Imagenet, LSUN, SVHN (for CIFAR10), CIFAR10 (for SVHN), KMNIST, and F-MNIST as OOD data across different DNN architectures such as ResNet34, WideResNet, DenseNet, and LeNet5.
****

Improved Estimation of Concentration Under ℓ_p-Norm Distance Metrics Using Half Spaces
Jack Prescott,Xiao Zhang,David Evans
Concentration of measure has been argued to be the fundamental cause of adversarial vulnerability. Mahloujifar et al. (2019) presented an empirical way to measure the concentration of a data distribution using samples, and employed it to find lower bounds on intrinsic robustness for several benchmark datasets. However, it remains unclear whether these lower bounds are tight enough to provide a useful approximation for the intrinsic robustness of a dataset. To gain a deeper understanding of the concentration of measure phenomenon, we first extend the Gaussian Isoperimetric Inequality to non-spherical Gaussian measures and arbitrary ℓ_p-norms ($p \geq 2$). We leverage these theoretical insights to design a method that uses half-spaces to estimate the concentration of any empirical dataset under ℓ_p-norm distance metrics. Our proposed algorithm is more efficient than Mahloujifar et al. (2019)'s, and experiments on synthetic datasets and image benchmarks demonstrate that it is able to find much tighter intrinsic robustness bounds. These tighter estimates provide further evidence that rules out intrinsic dataset concentration as a possible explanation for the adversarial vulnerability of state-of-the-art classifiers.
****

Compositional Models: Multi-Task Learning and Knowledge Transfer with Modular Networks
Andrey Zhmoginov,Dina Bashkirova,Mark Sandler
Conditional computation and modular networks have been recently proposed for multitask learning and other problems as a way to decompose problem solving into multiple reusable computational blocks. We propose a novel fully-differentiable approach for learning modular networks.In our method, the modules can be invoked repeatedly and allow knowledge transfer to novel tasks by adjusting the order of

computation. This allows soft weight sharing between tasks with only a small inc
rease in the number of parameters. We show that our method leads to interpretabl
e self-organization of modules in case of multi-task learning, transfer learning
 and domain adaptation while achieving competitive results on those tasks. From
practical perspective, our approach allows to: (a) reuse existing modules for le
arning new task by adjusting the computation order, (b) use it for unsupervised
multi-source domain adaptation to illustrate that adaptation to unseen data can
be achieved by only manipulating the order of pretrained modules, (c) show how o
ur approach can be used to increase accuracy of existing architectures for image
 classification tasks such as ImageNet, without any parameter increase, by reusi
ng the same block multiple times.
****

Concentric Spherical GNN for 3D Representation Learning
James S Fox,Bo Zhao,Sivasankaran Rajamanickam,Rampi Ramprasad,Le Song
Learning 3D representations that generalize well to arbitrarily oriented inputs
is a challenge of practical importance in applications varying from computer vis
ion to physics and chemistry.
We propose a novel multi-resolution convolutional architecture for learning over
 concentric spherical feature maps, of which the single sphere representation is
 a special case.
Our hierarchical architecture is based on alternatively learning to incorporate
both intra-sphere and inter-sphere information.
We show the applicability of our method for two different types of 3D inputs, me
sh objects, which can be regularly sampled, and point clouds, which are irregula
rly distributed.
We also propose an efficient mapping of point clouds to concentric spherical ima
ges using radial basis functions, thereby bridging spherical convolutions on gri
ds with general point clouds.
We demonstrate the effectiveness of our approach in achieving state-of-the-art p
erformance on 3D classification tasks with rotated data.
****

Adversarial Meta-Learning
Chengxiang Yin,Jian Tang,Zhiyuan Xu,Yanzhi Wang
Meta-learning enables a model to learn from very limited data to undertake a new
 task. In this paper, we study the general meta-learning with adversarial sample
s. We present a meta-learning algorithm, ADML (ADversarial Meta-Learner), which
leverages clean and adversarial samples to optimize the initialization of a lear
ning model in an adversarial manner. ADML leads to the following desirable prope
rties: 1) it turns out to be very effective even in the cases with only clean sa
mples; 2) it is robust to adversarial samples, i.e., unlike other meta-learning
algorithms, it only leads to a minor performance degradation when there are adve
rsarial samples; 3) it sheds light on tackling the cases with limited and even c
ontaminated samples. It has been shown by extensive experimental results that AD
ML outperforms several representative meta-learning algorithms in the cases invo
lving adversarial samples generated by different attack mechanisms, on two widel
y-used image datasets, MiniImageNet and CIFAR100, in terms of both accuracy and
robustness.
****

Understanding Self-supervised Learning with Dual Deep Networks
Yuandong Tian,Lantao Yu,Xinlei Chen,Surya Ganguli
We propose a novel theoretical framework to understand self-supervised learning
methods that employ dual pairs of deep ReLU networks (e.g., SimCLR, BYOL). First
, we prove that in each SGD update of SimCLR, the weights at each layer are upda
ted by a \emph{covariance operator} that specifically amplifies initial random s
electivities that vary across data samples but survive averages over data augmen
tations. We show this leads to the emergence of hierarchical features, if the in
put data are generated from a hierarchical latent tree model. With the same fram
ework, we also show analytically that in BYOL, the combination of BatchNorm and
a predictor network creates an implicit contrastive term, acting as an approxima
te covariance operator. Additionally, for linear architectures we derive exact s

olutions for BYOL that provide conceptual insights into how BYOL can learn usefu
l non-collapsed representations without any contrastive terms that separate nega
tive pairs. Extensive ablation studies justify our theoretical findings.
****

PDE-regularized Neural Networks for Image Classification
Jungeun Kim,Seunghyun Hwang,Jihyun Hwang,Kookjin Lee,Dongeun Lee,Noseong Park
Neural ordinary differential equations (neural ODEs) introduced an approach to a
pproximate a neural network as a system of ODEs after considering its layer as a
 continuous variable and discretizing its hidden dimension. While having several
 good characteristics, neural ODEs are known to be numerically unstable and slow
 in solving their integral problems, resulting in errors and/or much computation
 of the forward-pass inference. In this work, we present a novel partial differe
ntial equation (PDE)-based approach that removes the necessity of solving integr
al problems and considers both the layer and the hidden dimension as continuous
variables. Owing to the recent advancement of learning PDEs, the presented novel
 concept, called PR-Net, can be implemented. Our method shows comparable (or bet
ter) accuracy and robustness in much shorter forward-pass inference time for var
ious datasets and tasks in comparison with neural ODEs and Isometric MobileNet V
3. For the efficient nature of PR-Net, it is suitable to be deployed in resource
-scarce environments, e.g., deploying instead of MobileNet.
****

Towards a Reliable and Robust Dialogue System for Medical Automatic Diagnosis
Junfan Lin,Lin Xu,Ziliang Chen,Liang Lin
Dialogue system for medical automatic diagnosis (DSMAD) aims to learn an agent t
hat mimics the behavior of a human doctor, i.e. inquiring symptoms and informing
 diseases. Since DSMAD has been formulated as a Markov decision-making process,
many studies apply reinforcement learning methods to solve it. Unfortunately, ex
isting works solely rely on simple diagnostic accuracy to justify the effectiven
ess of their DSMAD agents while ignoring the medical rationality of the inquirin
g process. From the perspective of medical application, it's critical to develop
 an agent that is able to produce reliable and convincing diagnosing processes a
nd also is robust in making diagnosis facing noisy interaction with patients. To
 this end, we propose a novel DSMAD agent, INS-DS (Introspective Diagnosis Syste
m) comprising of two separate yet cooperative modules, i.e., an inquiry module f
or proposing symptom-inquiries and an introspective module for deciding when to
inform a disease. INS-DS is inspired by the introspective decision-making proces
s of human, where the inquiry module first proposes the most valuable symptom in
quiry, and then the introspective module intervenes the potential responses of t
his inquiry and decides to inquire only if the diagnoses of these interventions
vary.
****

EarlyBERT: Efficient BERT Training via Early-bird Lottery Tickets
Xiaohan Chen,Yu Cheng,Shuohang Wang,Zhe Gan,Zhangyang Wang,Jingjing Liu
Deep, heavily overparameterized language models such as BERT, XLNet and T5 have
achieved impressive success in many NLP tasks. However, their high model complex
ity requires enormous computation resources and extremely long training time for
 both pre-training and fine-tuning. Many works have studied model compression on
 large NLP models, but only focus on reducing inference cost/time, while still r
equiring expensive training process. Other works use extremely large batch sizes
 to shorten the pre-training time at the expense of high demand for computation
resources. In this paper, inspired by the Early-Bird Lottery Tickets studied for
 computer vision tasks, we propose EarlyBERT, a general computationally-efficien
t training algorithm applicable to both pre-training and fine-tuning of large-sc
ale language models. We are the first to identify structured winning tickets in
the early stage of BERT training, and use them for efficient training. Comprehen
sive pre-training and fine-tuning experiments on GLUE and SQuAD downstream tasks
 show that EarlyBERT easily achieves comparable performance to standard BERT wit
h 35~45% less training time.
****

Catastrophic Fisher Explosion: Early Phase Fisher Matrix Impacts Generalization

Stanislaw Kamil Jastrzebski,Devansh Arpit,Oliver Åstrand,Giancarlo Kerg,Huan Wang,Caiming Xiong,richard socher,Kyunghyun Cho,Krzysztof J. Geras
The early phase of training has been shown to be important in two ways for deep neural networks. First, the degree of regularization in this phase significantly impacts the final generalization. Second, it is accompanied by a rapid change in the local loss curvature influenced by regularization choices. Connecting these two findings, we show that stochastic gradient descent (SGD) implicitly penalizes the trace of the Fisher Information Matrix (FIM) from the beginning of training. We argue it is an implicit regularizer in SGD by showing that explicitly penalizing the trace of the FIM can significantly improve generalization. We further show that the early value of the trace of the FIM correlates strongly with the final generalization. We highlight that in the absence of implicit or explicit regularization, the trace of the FIM can increase to a large value early in training, to which we refer as catastrophic Fisher explosion. Finally, to gain insight into the regularization effect of penalizing the trace of the FIM, we show that it limits memorization by reducing the learning speed of examples with noisy labels more than that of the clean examples, and 2) trajectories with a low initial trace of the FIM end in flat minima, which are commonly associated with good generalization.
****
LATENT OPTIMIZATION VARIATIONAL AUTOENCODER FOR CONDITIONAL MOLECULAR GENERATION
Kisoo Kwon,Jung-Hyun Park,Kuhwan Jeong,Sunjae Lee,Hoshik Lee
Variational autoencoder (VAE) is a generation algorithm, consisting of an encoder and a decoder, and the latent variable from the encoder is used as the input of the decoder.
VAE is widely used for image, audio and text generation tasks. In general, the training of VAE is at risk of posterior collapsing especially for long sequential data. To alleviate this, modified evidence lower bounds (ELBOs) were proposed.
However, these approaches heuristically control training loss using a hyper-parameter, and it is not way to solve the fundamental problem of vanilla VAE.
In this paper, we propose a method to insert an optimization step of the latent variable and alternately update the encoder and decoder of conditional VAE for maximizing ELBOs.
In experiments, we applied the latent optimization VAE (LOVAE) on ZINC database, consisting of string representation of molecules, for the inverse molecular design.
We showed that the proposed LOVAE achieves better performance than vanilla VAE in terms of ELBOs and molecular generation performance. In addition, the proposed method showed better performance in property satisfaction and property maximization tasks compared to existing works.
****
Unsupervised Progressive Learning and the STAM Architecture
James Smith,Cameron Ethan Taylor,Seth Baer,Constantine Dovrolis
We first pose the Unsupervised Progressive Learning (UPL) problem: an online representation learning problem in which the learner observes a non-stationary and unlabeled data stream, and identifies a growing number of features that persist over time even though the data is not stored or replayed. To solve the UPL problem we propose the Self-Taught Associative Memory (STAM) architecture. Layered hierarchies of STAM modules learn based on a combination of online clustering, novelty detection, forgetting outliers, and storing only prototypical features rather than specific examples. We evaluate STAM representations using classification and clustering tasks. While there are no existing learning scenarios which are directly comparable to UPL, we compare the STAM architecture with two recent continual learning works; Memory Aware Synapses (MAS), and Gradient Episodic Memories (GEM), which have been modified to be suitable for the UPL setting.
****
Beyond Categorical Label Representations for Image Classification
Boyuan Chen,Yu Li,Sunand Raghupathi,Hod Lipson
We find that the way we choose to represent data labels can have a profound effect on the quality of trained models. For example, training an image classifier t

o regress audio labels rather than traditional categorical probabilities produce
s a more reliable classification. This result is surprising, considering that au
dio labels are more complex than simpler numerical probabilities or text. We hyp
othesize that high dimensional, high entropy label representations are generally
 more useful because they provide a stronger error signal. We support this hypot
hesis with evidence from various label representations including constant matric
es, spectrograms, shuffled spectrograms, Gaussian mixtures, and uniform random m
atrices of various dimensionalities. Our experiments reveal that high dimensiona
l, high entropy labels achieve comparable accuracy to text (categorical) labels
on standard image classification tasks, but features learned through our label r
epresentations exhibit more robustness under various adversarial attacks and bet
ter effectiveness with a limited amount of training data. These results suggest
that label representation may play a more important role than previously thought
.
****

Fantastic Four: Differentiable and Efficient Bounds on Singular Values of Convol
ution Layers

Sahil Singla,Soheil Feizi

In deep neural networks, the spectral norm of the Jacobian of a layer bounds the
 factor by which the norm of a signal changes during forward/backward propagatio
n. Spectral norm regularizations have been shown to improve generalization, robu
stness and optimization of deep learning methods. Existing methods to compute th
e spectral norm of convolution layers either rely on heuristics that are efficie
nt in computation but lack guarantees or are theoretically-sound but computation
ally expensive. In this work, we obtain the best of both worlds by deriving {\it
 four} provable upper bounds on the spectral norm of a standard 2D multi-channel
 convolution layer. These bounds are differentiable and can be computed efficien
tly during training with negligible overhead. One of these bounds is in fact the
 popular heuristic method of Miyato et al. (multiplied by a constant factor depe
nding on filter sizes). Each of these four bounds can achieve the tightest gap d
epending on convolution filters. Thus, we propose to use the minimum of these fo
ur bounds as a tight, differentiable and efficient upper bound on the spectral n
orm of convolution layers. Moreover, our spectral bound is an effective regulari
zer and can be used to bound either the lipschitz constant or curvature values (
eigenvalues of the Hessian) of neural networks. Through experiments on MNIST and
 CIFAR-10, we demonstrate the effectiveness of our spectral bound in improving g
eneralization and robustness of deep networks.
****

Pretrain Knowledge-Aware Language Models

Corbin L Rosset,Chenyan Xiong,Minh Phan,Xia Song,Paul N. Bennett,saurabh tiwary

How much knowledge do pretrained language models hold? Recent research observed
that pretrained transformers are adept at modeling semantics but it is unclear t
o what degree they grasp human knowledge, or how to ensure they do so. In this p
aper we incorporate knowledge-awareness in language model pretraining without ch
anging the transformer architecture, inserting explicit knowledge layers, or add
ing external storage of semantic information. Rather, we simply signal the exist
ence of entities to the input of the transformer in pretraining, with an entity-
extended tokenizer; and at the output, with an additional entity prediction task
. Our experiments show that solely by adding these entity signals in pretraining
, significantly more knowledge is packed into the transformer parameters: we obs
erve improved language modeling accuracy, factual correctness in LAMA knowledge
probing tasks, and semantics in the hidden representations through edge probing.
 We also show that our knowledge-aware language model (\kalm{}) can serve as a d
rop-in replacement for GPT-2 models, significantly improving downstream tasks li
ke zero-shot question-answering with no task-related training.
****

Accelerating Convergence of Replica Exchange Stochastic Gradient MCMC via Varian
ce Reduction

Wei Deng,Qi Feng,Georgios P. Karagiannis,Guang Lin,Faming Liang

Replica exchange stochastic gradient Langevin dynamics (reSGLD) has shown promis

e in accelerating the convergence in non-convex learning; however, an excessivel
y large correction for avoiding biases from noisy energy estimators has limited
the potential of the acceleration. To address this issue, we study the variance
reduction for noisy energy estimators, which promotes much more effective swaps.
 Theoretically, we provide a non-asymptotic analysis on the exponential converge
nce for the underlying continuous-time Markov jump process; moreover, we conside
r a generalized Girsanov theorem which includes the change of Poisson measure to
 overcome the crude discretization based on the Gr\"{o}wall's inequality and yie
lds a much tighter error in the 2-Wasserstein ($\mathcal{W}_2$) distance. Numeri
cally, we conduct extensive experiments and obtain state-of-the-art results in o
ptimization and uncertainty estimates for synthetic experiments and image data.
****

A Technical and Normative Investigation of Social Bias Amplification
Angelina Wang,Olga Russakovsky
The conversation around the fairness of machine learning models is growing and e
volving. In this work, we focus on the issue of bias amplification: the tendency
 of models trained from data containing social biases to further amplify these b
iases. This problem is brought about by the algorithm, on top of the level of bi
as already present in the data. We make two main contributions regarding its mea
surement. First, building off of Zhao et al. (2017), we introduce and analyze a
new, decoupled metric for measuring bias amplification, $\text{BiasAmp}_{\righta
rrow}, which possesses a number of attractive properties, including the ability
to pinpoint the cause of bias amplification. Second, we thoroughly analyze and d
iscuss the normative implications of this metric. We provide suggestions about i
ts measurement by cautioning against predicting sensitive attributes, encouragin
g the use of confidence intervals due to fluctuations in the fairness of models
across runs, and discussing what bias amplification means in the context of doma
ins where labels either don't exist at test time or correspond to uncertain futu
re events. Throughout this paper, we work to provide a deeply interrogative look
 at the technical measurement of bias amplification, guided by our normative ide
as of what we want it to encompass.
****

Normalizing Flows for Calibration and Recalibration
Achintya Gopal,Aaron Key
In machine learning, due to model misspecification and overfitting, estimates of
 the aleatoric uncertainty are often inaccurate.
One approach to fix this is isotonic regression, in which a monotonic function i
s fit on a validation set to map the model's CDF to an optimally calibrated CDF.
 However, this makes it infeasible to compute additional statistics of interest
on the model distribution (such as the mean). In this paper, through a reframing
 of recalibration as MLE, we replace isotonic regression with normalizing flows.
 This allows us to retain the ability to compute the statistical properties of t
he model (such as closed-form likelihoods, mean, correlation, etc.) and provides
 an opportunity for additional capacity at the cost of possible overfitting. Mo
st importantly, the fundamental properties of normalizing flows allow us to gene
ralize recalibration to conditional and multivariate distributions. To aid in de
tecting miscalibration and measuring our success at fixing it, we use a simple e
xtension of the calibration Q-Q plot.
****

Testing Robustness Against Unforeseen Adversaries
Daniel Kang,Yi Sun,Dan Hendrycks,Tom B Brown,Jacob Steinhardt
Most existing adversarial defenses only measure robustness to L_p adversarial
attacks. Not only are adversaries unlikely to exclusively create small L_p per
turbations, adversaries are unlikely to remain fixed. Adversaries adapt and evol
ve their attacks; hence adversarial defenses must be robust to a broad range of
unforeseen attacks. We address this discrepancy between research and reality by
proposing a new evaluation framework called ImageNet-UA. Our framework enables t
he research community to test ImageNet model robustness against attacks not enco
untered during training. To create ImageNet-UA's diverse attack suite, we introd
uce a total of four novel adversarial attacks. We also demonstrate that, in comp

arison to ImageNet-UA, prevailing L_∞ robustness assessments give a narro
w account of adversarial robustness. By evaluating current defenses with ImageNe
t-UA, we find they provide little robustness to unforeseen attacks. We hope the
greater variety and realism of ImageNet-UA enables development of more robust de
fenses which can generalize beyond attacks seen during training.
****

A Bayesian-Symbolic Approach to Learning and Reasoning for Intuitive Physics
Kai Xu,Akash Srivastava,Dan Gutfreund,Felix Sosa,Tomer Ullman,Joshua B. Tenenbau
m,Charles Sutton
Humans are capable of reasoning about physical phenomena by inferring laws of ph
ysics from a very limited set of observations. The inferred laws can potentially
 depend on unobserved properties, such as mass, texture, charge, etc. This sampl
e-efficient physical reasoning is considered a core domain of human common-sense
 knowledge and hints at the existence of a physics engine in the head. In this p
aper, we propose a Bayesian symbolic framework for learning sample-efficient mod
els of physical reasoning and prediction, which are of special interests in the
field of intuitive physics. In our framework, the environment is represented by
a top-down generative model with a collection of entities with some known and un
known properties as latent variables to capture uncertainty. The physics engine
depends on physical laws which are modeled as interpretable symbolic expressions
 and are assumed to be functions of the latent properties of the entities intera
cting under simple Newtonian physics. As such, learning the laws is then reduced
 to symbolic regression and Bayesian inference methods are used to obtain the di
stribution of unobserved properties. These inference and regression steps are pe
rformed in an iterative manner following the expectation–maximization algorithm
to infer the unknown properties and use them to learn the laws from a very small
 set of observations. We demonstrate that on three physics learning tasks that c
ompared to the existing methods of learning physics, our proposed framework is m
ore data-efficient, accurate and makes joint reasoning and learning possible.
****

Sharing Less is More: Lifelong Learning in Deep Networks with Selective Layer Tr
ansfer
Seungwon Lee,Sima Behpour,ERIC EATON
Effective lifelong learning across diverse tasks requires diverse knowledge, yet
 transferring irrelevant knowledge may lead to interference and catastrophic for
getting. In deep networks, transferring the appropriate granularity of knowledge
 is as important as the transfer mechanism, and must be driven by the relationsh
ips among tasks. We first show that the lifelong learning performance of several
 current deep learning architectures can be significantly improved by transfer a
t the appropriate layers. We then develop an expectation-maximization (EM) metho
d to automatically select the appropriate transfer configuration and optimize th
e task network weights. This EM-based selective transfer is highly effective, as
 demonstrated on three algorithms in several lifelong object classification scen
arios.
****

Improved Contrastive Divergence Training of Energy Based Models
Yilun Du,Shuang Li,Joshua B. Tenenbaum,Igor Mordatch
We propose several different techniques to improve contrastive divergence traini
ng of energy-based models (EBMs). We first show that a gradient term neglected i
n the popular contrastive divergence formulation is both tractable to estimate a
nd is important to avoid training instabilities in previous models. We further h
ighlight how data augmentation, multi-scale processing, and reservoir sampling c
an be used to improve model robustness and generation quality. Thirdly, we empir
ically evaluate stability of model architectures and show improved performance o
n a host of benchmarks and use cases, such as image generation, OOD detection, a
nd compositional generation.
****

Learning Online Data Association
Yilun Du,Joshua B. Tenenbaum,Tomas Perez,Leslie Pack Kaelbling
When an agent interacts with a complex environment, it receives a stream of perc

epts in which it may detect entities, such as objects or people. To build up a c oherent, low-variance estimate of the underlying state, it is necessary to fuse information from multiple detections over time. To do this fusion, the agent mus t decide which detections to associate with one another. We address this data-as sociation problem in the setting of an online filter, in which each observation is processed by aggregating into an existing object hypothesis. Classic methods with strong probabilistic foundations exist, but they are computationally expens ive and require models that can be difficult to acquire. In this work, we use th e deep-learning tools of sparse attention and representation learning to learn a machine that processes a stream of detections and outputs a set of hypotheses a bout objects in the world. We evaluate this approach on simple clustering proble ms, problems with dynamics, and a complex image-based domain. We find that it ge neralizes well from short to long observation sequences and from a few to many h ypotheses, outperforming other learning approaches and classical non-learning me thods.

****

ATOM3D: Tasks On Molecules in Three Dimensions

Raphael John Lamarre Townshend,Martin Vogele,Patricia Suriana,Alex Derry,Alex Po wers,Yianni Laloudakis,Sidhika Balachandar,Brandon M Anderson,Stephan Eismann,Ri si Kondor,Russ Altman,Ron O. Dror

While a variety of methods have been developed for predicting molecular properti es, deep learning networks that operate directly on three-dimensional molecular structure have recently demonstrated particular promise. In this work we present ATOM3D, a collection of both novel and existing datasets spanning several key c lasses of biomolecules, to systematically assess such learning methods. We devel op three-dimensional molecular learning networks for each of these tasks, findin g that they consistently improve performance relative to one- and two-dimensiona l methods. The specific choice of architecture proves to be critical for perform ance, with three-dimensional convolutional networks excelling at tasks involving complex geometries, while graph networks perform well on systems requiring deta iled positional information. Furthermore, equivariant networks show significant promise but are currently unable to scale. Our results indicate many molecular p roblems stand to gain from three-dimensional molecular learning. All code and da tasets are available at github.com/xxxxxxx/xxxxxx.

****

Frequency Regularized Deep Convolutional Dictionary Learning and Application to Blind Denoising

Nikola Pavle Janjusevic,Amirhossein Khalilian-Gourtani,Yao Wang

Sparse representation via a learned dictionary is a powerful prior for natural i mages. In recent years, unrolled sparse coding algorithms (e.g. LISTA) have prov en to be useful for constructing interpretable deep-learning networks that perfo rm on par with state-of-the-art models on image-restoration tasks. In this study we are concerned with extending the work of such convolutional dictionary learn ing (CDL) models. We propose to construct strided convolutional dictionaries wit h a single analytic low-pass filter and a set of learned filters regularized to occupy the complementary frequency space. By doing so, we address the necessary modeling assumptions of natural images with respect to convolutional sparse codi ng and reduce the mutual coherence and redundancy of the learned filters. We sho w improved denoising performance at reduced computational complexity when compar ed to other CDL methods, and competitive results when compared to popular deep-l earning models. We further propose to parameterize the thresholds in the soft-th resholding operator of LISTA to be proportional to the estimated noise-variance from an input image. We demonstrate that this parameterization enhances robustne ss to noise-level mismatch between training and inference.

****

On the Power of Abstention and Data-Driven Decision Making for Adversarial Robus tness

Nina Balcan,Avrim Blum,Dravyansh Sharma,Hongyang Zhang

We formally define a feature-space attack where the adversary can perturb datapo ints by arbitrary amounts but in restricted directions. By restricting the attac

k to a small random subspace, our model provides a clean abstraction for non-Lip schitz networks which map small input movements to large feature movements. We p rove that classifiers with the ability to abstain are provably more powerful tha n those that cannot in this setting. Specifically, we show that no matter how we ll-behaved the natural data is, any classifier that cannot abstain will be defea ted by such an adversary. However, by allowing abstention, we give a parameteri zed algorithm with provably good performance against such an adversary when clas ses are reasonably well-separated in feature space and the dimension of the feat ure space is high. We further use a data-driven method to set our algorithm para meters to optimize over the accuracy vs. abstention trade-off with strong theore tical guarantees. Our theory has direct applications to the technique of contras tive learning, where we empirically demonstrate the ability of our algorithms to obtain high robust accuracy with only small amounts of abstention in both super vised and self-supervised settings. Our results provide a first formal abstentio n-based gap, and a first provable optimization for the induced trade-off in an a dversarial defense setting.
****

IsarStep: a Benchmark for High-level Mathematical Reasoning
Wenda Li,Lei Yu,Yuhuai Wu,Lawrence C. Paulson
A well-defined benchmark is essential for measuring and accelerating research pr ogress of machine learning models. In this paper, we present a benchmark for hig h-level mathematical reasoning and study the reasoning capabilities of neural se quence-to-sequence models. We build a non-synthetic dataset from the largest rep ository of proofs written by human experts in a theorem prover. The dataset has a broad coverage of undergraduate and research-level mathematical and computer s cience theorems. In our defined task, a model is required to fill in a missing i ntermediate proposition given surrounding proofs. This task provides a starting point for the long-term goal of having machines generate human-readable proofs a utomatically. Our experiments and analysis reveal that while the task is challen ging, neural models can capture non-trivial mathematical reasoning. We further d esign a hierarchical transformer that outperforms the transformer baseline.
****

HW-NAS-Bench: Hardware-Aware Neural Architecture Search Benchmark
Chaojian Li,Zhongzhi Yu,Yonggan Fu,Yongan Zhang,Yang Zhao,Haoran You,Qixuan Yu,Y ue Wang,Cong Hao,Yingyan Lin
HardWare-aware Neural Architecture Search (HW-NAS) has recently gained tremendou s attention by automating the design of deep neural networks deployed in more re source-constrained daily life devices. Despite its promising performance, develo ping optimal HW-NAS solutions can be prohibitively challenging as it requires cr oss-disciplinary knowledge in the algorithm, micro-architecture, and device-spec ific compilation. First, to determine the hardware-cost to be incorporated into the NAS process, existing works mostly adopt either pre-collected hardware-cost look-up tables or device-specific hardware-cost models. The former can be time-c onsuming due to the required knowledge of the device’s compilation method and ho w to set up the measurement pipeline, while building the latter is often a barri er for non-hardware experts like NAS researchers. Both of them limit the develop ment of HW-NAS innovations and impose a barrier-to-entry to non-hardware experts . Second, similar to generic NAS, it can be notoriously difficult to benchmark H W-NAS algorithms due to their significant required computational resources and t he differences in adopted search spaces, hyperparameters, and hardware devices. To this end, we develop HW-NAS-Bench, the first public dataset for HW-NAS resear ch which aims to democratize HW-NAS research to non-hardware experts and make HW -NAS research more reproducible and accessible. To design HW-NAS-Bench, we caref ully collected the measured/estimated hardware performance (e.g., energy cost an d latency) of all the networks in the search spaces of both NAS-Bench-201 and FB Net, on six hardware devices that fall into three categories (i.e., commercial e dge devices, FPGA, and ASIC). Furthermore, we provide a comprehensive analysis o f the collected measurements in HW-NAS-Bench to provide insights for HW-NAS rese arch. Finally, we demonstrate exemplary user cases to (1) show that HW-NAS-Bench allows non-hardware experts to perform HW-NAS by simply querying our pre-measur

ed dataset and (2) verify that dedicated device-specific HW-NAS can indeed lead to optimal accuracy-cost trade-offs. The codes and all collected data are available at https://github.com/RICE-EIC/HW-NAS-Bench.
****
Factorizing Declarative and Procedural Knowledge in Structured, Dynamical Environments
Anirudh Goyal,Alex Lamb,Phanideep Gampa,Philippe Beaudoin,Charles Blundell,Sergey Levine,Yoshua Bengio,Michael Curtis Mozer
Modeling a structured, dynamic environment like a video game requires keeping track of the objects and their states (declarative knowledge) as well as predicting how objects behave (procedural knowledge). Black-box models with a monolithic hidden state often fail to apply procedural knowledge consistently and uniformly, i.e., they lack systematicity. For example, in a video game, correct prediction of one enemy's trajectory does not ensure correct prediction of another's. We address this issue via an architecture that factorizes declarative and procedural knowledge and that imposes modularity within each form of knowledge. The architecture consists of active modules called object files that maintain the state of a single object and invoke passive external knowledge sources called schemata that prescribe state updates. To use a video game as an illustration, two enemies of the same type will share schemata but will have separate object files to encode their distinct state (e.g., health, position). We propose to use attention to determine which object files to update, the selection of schemata, and the propagation of information between object files. The resulting architecture is a drop-in replacement conforming to the same input-output interface as normal recurrent networks (e.g., LSTM, GRU) yet achieves substantially better generalization on environments that have multiple object tokens of the same type, including a challenging intuitive physics benchmark.

****
Function Contrastive Learning of Transferable Representations
Muhammad Waleed Gondal,Shruti Joshi,Nasim Rahaman,Stefan Bauer,Manuel Wuthrich,Bernhard Schölkopf
Few-shot-learning seeks to find models that are capable of fast-adaptation to novel tasks which are not encountered during training. Unlike typical few-shot learning algorithms, we propose a contrastive learning method which is not trained to solve a set of tasks, but rather attempts to find a good representation of the underlying data-generating processes (\emph{functions}). This allows for finding representations which are useful for an entire series of tasks sharing the same function. In particular, our training scheme is driven by the self-supervision signal indicating whether two sets of samples stem from the same underlying function. Our experiments on a number of synthetic and real-world datasets show that the representations we obtain can outperform strong baselines in terms of downstream performance and noise robustness, even when these baselines are trained in an end-to-end manner.
****
Optimal Transport Graph Neural Networks
Gary Bécigneul,Octavian-Eugen Ganea,Benson Chen,Regina Barzilay,Tommi S. Jaakkola
Current graph neural network (GNN) architectures naively average or sum node embeddings into an aggregated graph representation---potentially losing structural or semantic information. We here introduce OT-GNN, a model that computes graph embeddings using parametric prototypes that highlight key facets of different graph aspects. Towards this goal, we are (to our knowledge) the first to successfully combine optimal transport with parametric graph models. Graph representations are obtained from Wasserstein distances between the set of GNN node embeddings and "prototype" point clouds as free parameters. We theoretically prove that, unlike traditional sum aggregation, our function class on point clouds satisfies a fundamental universal approximation theorem. Empirically, we address an inherent collapse optimization issue by proposing a noise contrastive regularizer to steer the model towards truly exploiting the optimal transport geometry. Finally,

we consistently report better generalization performance on several molecular pr
operty prediction tasks, while exhibiting smoother graph representations.
****
Provable Rich Observation Reinforcement Learning with Combinatorial Latent State
s
Dipendra Misra,Qinghua Liu,Chi Jin,John Langford
We propose a novel setting for reinforcement learning that combines two common r
eal-world difficulties: presence of observations (such as camera images) and fac
tored states (such as location of objects). In our setting, the agent receives o
bservations generated stochastically from a "latent" factored state. These obser
vations are "rich enough" to enable decoding of the latent state and remove part
ial observability concerns. Since the latent state is combinatorial, the size of
 state space is exponential in the number of latent factors. We create a learnin
g algorithm FactoRL (Fact-o-Rel) for this setting, which uses noise-contrastive
 learning to identify latent structures in emission processes and discover a fac
torized state space. We derive polynomial sample complexity guarantees for Facto
RL which polynomially depend upon the number factors, and very weakly depend on
the size of the observation space. We also provide a guarantee of polynomial ti
me complexity when given access to an efficient planning algorithm.
****
LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial
 Recognition
Valeriia Cherepanova,Micah Goldblum,Harrison Foley,Shiyuan Duan,John P Dickerson
,Gavin Taylor,Tom Goldstein
Facial recognition systems are increasingly deployed by private corporations, go
vernment agencies, and contractors for consumer services and mass surveillance p
rograms alike. These systems are typically built by scraping social media profi
les for user images. Adversarial perturbations have been proposed for bypassing
 facial recognition systems. However, existing methods fail on full-scale syste
ms and commercial APIs. We develop our own adversarial filter that accounts for
 the entire image processing pipeline and is demonstrably effective against indu
strial-grade pipelines that include face detection and large scale databases. A
dditionally, we release an easy-to-use webtool that significantly degrades the a
ccuracy of Amazon Rekognition and the Microsoft Azure Face Recognition API, redu
cing the accuracy of each to below 1%.
****
Neural Networks for Learning Counterfactual G-Invariances from Single Environmen
ts
S Chandra Mouli,Bruno Ribeiro
Despite —or maybe because of— their astonishing capacity to fit data, neural net
works are believed to have difficulties extrapolating beyond training data distr
ibution. This work shows that, for extrapolations based on finite transformation
 groups, a model’s inability to extrapolate is unrelated to its capacity. Rather
, the shortcoming is inherited from a learning hypothesis: Examples not explicit
ly observed with infinitely many training examples have underspecified outcomes
in the learner’s model. In order to endow neural networks with the ability to ex
trapolate over group transformations, we introduce a learning framework counterf
actually-guided by the learning hypothesis that any group invariance to (known)
transformation groups is mandatory even without evidence, unless the learner dee
ms it inconsistent with the training data. Unlike existing invariance-driven met
hods for (counterfactual) extrapolations, this framework allows extrapolations f
rom a single environment. Finally, we introduce sequence and image extrapolation
 tasks that validate our framework and showcase the shortcomings of traditional
approaches.
****
Online Limited Memory Neural-Linear Bandits
Tom Zahavy,Ofir Nabati,Leor Cohen,Shie Mannor
We study neural-linear bandits for solving problems where both exploration and r
epresentation learning play an important role. Neural-linear bandits leverage th
e representation power of deep neural networks and combine it with efficient exp

loration mechanisms, designed for linear contextual bandits, on top of the last hidden layer. Since the representation is optimized during learning, information regarding exploration with “old” features is lost. We propose the first limited memory neural- linear bandit that is resilient to this catastrophic forgetting phenomenon by solving a semi-definite program. We then approximate the semi-definite program using stochastic gradient descent to make the algorithm practical and adjusted for online usage. We perform simulations on a variety of data sets, including regression, classification, and sentiment analysis. In addition, we evaluate our algorithm in a challenging uplink rate-control application. The bandit controls the transmission rates of data segments over cellular links to achieve optimal throughput. We observe that our algorithm achieves superior performance and shows resilience to catastrophic forgetting.
****
Constructing Multiple High-Quality Deep Neural Networks: A TRUST-TECH Based Approach
Zhiyong Hao,Hsiao-Dong Chiang,Bin Wang
The success of deep neural networks relied heavily on efficient stochastic gradient descent-like training methods. However, these methods are sensitive to initialization and hyper-parameters.
In this paper, a systematical method for finding multiple high-quality local optimal deep neural networks from a single training session, using the TRUST-TECH (TRansformation Under Stability-reTaining Equilibria Characterization) method, is introduced.
To realize effective TRUST-TECH searches to train deep neural networks on large datasets, a dynamic search paths (DSP) method is proposed to provide an improved search guidance in TRUST-TECH method.
The proposed DSP-TT method is implemented such that the computation graph remains constant during the search process, with only minor GPU memory overhead and requires just one training session to obtain multiple local optimal solutions (LOS). To take advantage of these LOSs, we also propose an improved ensemble method. Experiments on image classification datasets show that our method improves the testing performance by a substantial margin. Specifically, our fully-trained DSP-TT ResNet ensmeble improves the SGD baseline by 20\% (CIFAR10) and 15\%(CIFAR100). Furthermore, our method shows several advantages over other ensembling methods.
****
Simple Spectral Graph Convolution
Hao Zhu,Piotr Koniusz
Graph Convolutional Networks (GCNs) are leading methods for learning graph representations. However, without specially designed architectures, the performance of GCNs degrades quickly with increased depth. As the aggregated neighborhood size and neural network depth are two completely orthogonal aspects of graph representation, several methods focus on summarizing the neighborhood by aggregating K-hop neighborhoods of nodes while using shallow neural networks. However, these methods still encounter oversmoothing, and suffer from high computation and storage costs. In this paper, we use a modified Markov Diffusion Kernel to derive a variant of GCN called Simple Spectral Graph Convolution (SSGC). Our spectral analysis shows that our simple spectral graph convolution used in SSGC is a trade-off of low- and high-pass filter bands which capture the global and local contexts of each node. We provide two theoretical claims which demonstrate that we can aggregate over a sequence of increasingly larger neighborhoods compared to competitors while limiting severe oversmoothing. Our experimental evaluations show that SSGC with a linear learner is competitive in text and node classification tasks. Moreover, SSGC is comparable to other state-of-the-art methods for node clustering and community prediction tasks.
****
Regularized Inverse Reinforcement Learning
Wonseok Jeon,Chen-Yang Su,Paul Barde,Thang Doan,Derek Nowrouzezahrai,Joelle Pineau
Inverse Reinforcement Learning (IRL) aims to facilitate a learner’s ability to i

mitate expert behavior by acquiring reward functions that explain the expert’s d ecisions. Regularized IRLapplies strongly convex regularizers to the learner’s p olicy in order to avoid the expert’s behavior being rationalized by arbitrary co nstant rewards, also known as degenerate solutions. We propose tractable solutio ns, and practical methods to obtain them, for regularized IRL. Current methods a re restricted to the maximum-entropy IRL framework, limiting them to Shannon-ent ropy regularizers, as well as proposing solutions that are intractable in practi ce. We present theoretical backing for our proposed IRL method’s applicability to both discrete and continuous controls, empirically validating our performance on a variety of tasks.
****

Response Modeling of Hyper-Parameters for Deep Convolutional Neural Networks
Mathieu Tuli,Mahdi S. Hosseini,Konstantinos N Plataniotis
Hyper-parameter optimization (HPO) is critical in training high performing Deep Neural Networks (DNN). Current methodologies fail to define an analytical respon se surface and remain a training bottleneck due to their use of additional inter nal hyper-parameters and lengthy evaluation cycles. We demonstrate that the low-rank factorization of the convolution weights of intermediate layers of a CNN ca n define an analytical response surface. We quantify how this surface acts as an auxiliary to optimizing training metrics. We introduce a dynamic tracking algor ithm -- autoHyper -- that performs HPO on the order of hours for various dataset s including ImageNet and requires no manual tuning. Our method -- using a single RTX2080Ti -- is able to select a learning rate within 59 hours for AdaM on ResN et34 applied to ImageNet and improves in testing accuracy by 4.93% over the defa ult learning rate. In contrast to previous methods, we empirically prove that ou r algorithm and response surface generalize well across model, optimizer, and da taset selection removing the need for extensive domain knowledge to achieve high levels of performance.
****

Why Convolutional Networks Learn Oriented Bandpass Filters: Theory and Empirical Support
Isma Hadji,Richard Wildes
It has been repeatedly observed that convolutional architectures when applied to image understanding tasks learn oriented bandpass filters. A standard explanati on of this result is that these filters reflect the structure of the images that they have been exposed to during training: Natural images typically are locally composed of oriented contours at various scales and oriented bandpass filters a re matched to such structure. We offer an alternative explanation based not on t he structure of images, but rather on the structure of convolutional architectur es. In particular, complex exponentials are the eigenfunctions of convolution. T hese eigenfunctions are defined globally; however, convolutional architectures o perate locally. To enforce locality, one can apply a windowing function to the eigenfunctions, which leads to oriented bandpass filters as the natural operator s to be learned with convolutional architectures. From a representational point of view, these filters allow for a local systematic way to characterize and oper ate on an image or other signal. We offer empirical support for the hypothesis t hat convolutional networks learn such filters at all of their convolutional laye rs. While previous research has shown evidence of filters having oriented bandpa ss characteristics at early layers, ours appears to be the first study to docume nt the predominance of such filter characteristics at all layers. Previous studi es have missed this observation because they have concentrated on the cumulative compositional effects of filtering across layers, while we examine the filter c haracteristics that are present at each layer.
****

Towards Learning to Remember in Meta Learning of Sequential Domains
Zhenyi Wang,Tiehang Duan,Donglin Zhan,Changyou Chen
Meta-learning has made rapid progress in past years, with recent extensions made to avoid catastrophic forgetting in the learning process, namely continual meta learning. It is desirable to generalize the meta learner’s ability to continuou sly learn in sequential domains, which is largely unexplored to-date. We found t

hrough extensive empirical verification that significant improvement
is needed for current continual learning techniques to be applied in the sequent
ial domain meta learning setting. To tackle the problem, we adapt existing dynam
ic learning rate adaptation techniques to meta learn both model parameters and l
earning rates. Adaptation on parameters ensures good generalization performance,
 while adaptation on learning rates is made to avoid
catastrophic forgetting of past domains. Extensive experiments on a sequence of
commonly used real-domain data demonstrate the effectiveness of our proposed met
hod, outperforming current strong baselines in continual learning. Our code is m
ade publicly available online (anonymous)
****

Regret Bounds and Reinforcement Learning Exploration of EXP-based Algorithms
Mengfan Xu,Diego Klabjan
EXP-based algorithms are often used for exploration in multi-armed bandit. We re
visit the EXP3.P algorithm and establish both the lower and upper bounds of regr
et in the Gaussian multi-armed bandit setting, as well as a more general distrib
ution option. The analyses do not require bounded rewards compared to classical
regret assumptions. We also extend EXP4 from multi-armed bandit to reinforcement
 learning to incentivize exploration by multiple agents. The resulting algorithm
 has been tested on hard-to-explore games and it shows an improvement on explora
tion compared to state-of-the-art.
****

Anatomy of Catastrophic Forgetting: Hidden Representations and Task Semantics
Vinay Venkatesh Ramasesh,Ethan Dyer,Maithra Raghu
Catastrophic forgetting is a recurring challenge to developing versatile deep le
arning models. Despite its ubiquity, there is limited understanding of its conne
ctions to neural network (hidden) representations and task semantics. In this pa
per, we address this important knowledge gap. Through quantitative analysis of n
eural representations, we find that deeper layers are disproportionately respons
ible for forgetting, with sequential training resulting in an erasure of earlier
 task representational subspaces. Methods to mitigate forgetting stabilize these
 deeper layers, but show diversity on precise effects, with some increasing feat
ure reuse while others store task representations orthogonally, preventing inter
ference. These insights also enable the development of an analytic argument and
empirical picture relating forgetting to task semantic similarity, where we find
 that maximal forgetting occurs for task sequences with intermediate similarity.
****

TaskSet: A Dataset of Optimization Tasks
Luke Metz,Niru Maheswaranathan,Ruoxi Sun,C. Daniel Freeman,Ben Poole,Jascha Sohl
-Dickstein
We present TaskSet, a dataset of tasks for use in training and evaluating optimi
zers. TaskSet is unique in its size and diversity, containing over a thousand ta
sks ranging from image classification with fully connected or convolutional neur
al networks, to variational autoencoders, to non-volume preserving flows on a va
riety of datasets. As an example application of such a dataset we explore meta-l
earning an ordered list of hyperparameters to try sequentially. By learning this
 hyperparameter list from data generated using TaskSet we achieve large speedups
 in sample efficiency over random search. Next we use the diversity of the TaskS
et and our method for learning hyperparameter lists to empirically explore the g
eneralization of these lists to new optimization tasks in a variety of settings
including ImageNet classification with Resnet50 and LM1B language modeling with
transformers. As part of this work we have opensourced code for all tasks, as we
ll as ~29 million training curves for these problems and the corresponding hyper
parameters.
****

On Fast Adversarial Robustness Adaptation in Model-Agnostic Meta-Learning
Ren Wang,Kaidi Xu,Sijia Liu,Pin-Yu Chen,Tsui-Wei Weng,Chuang Gan,Meng Wang
Model-agnostic meta-learning (MAML) has emerged as one of the most successful me
ta-learning techniques in few-shot learning. It enables us to learn a $\textit{m
eta-initialization}$ of model parameters (that we call $\textit{meta-model}$) to

rapidly adapt to new tasks using a small amount of labeled training data. Despite the generalization power of the meta-model, it remains elusive that how $\textit{adversarial robustness}$ can be maintained by MAML in few-shot learning. In addition to generalization, robustness is also desired for a meta-model to defend adversarial examples (attacks). Toward promoting adversarial robustness in MAML, we first study $\textit{when}$ a robustness-promoting regularization should be incorporated, given the fact that MAML adopts a bi-level (fine-tuning vs. meta-update) learning procedure. We show that robustifying the meta-update stage is sufficient to make robustness adapted to the task-specific fine-tuning stage even if the latter uses a standard training protocol. We also make additional justification on the acquired robustness adaptation by peering into the interpretability of neurons' activation maps. Furthermore, we investigate $\textit{how}$ robust regularization can $\textit{efficiently}$ be designed in MAML. We propose a general but easily-optimized robustness-regularized meta-learning framework, which allows the use of unlabeled data augmentation, fast adversarial attack generation, and computationally-light fine-tuning. In particular, we for the first time show that the auxiliary contrastive learning task can enhance the adversarial robustness of MAML. Finally, extensive experiments are conducted to demonstrate the effectiveness of our proposed methods in robust few-shot learning.

****

Safety Verification of Model Based Reinforcement Learning Controllers

akshita gupta,Inseok Hwang

Model-based reinforcement learning (RL) has emerged as a promising tool for developing controllers for real world systems (e.g., robotics, autonomous driving, etc.). However, real systems often have constraints imposed on their state space which must be satisfied to ensure the safety of the system and its environment. Developing a verification tool for RL algorithms is challenging because the non-linear structure of neural networks impedes analytical verification of such models or controllers. To this end, we present a novel safety verification framework for model-based RL controllers using reachable set analysis. The proposed framework can efficiently handle models and controllers which are represented using neural networks. Additionally, if a controller fails to satisfy the safety constraints in general, the proposed framework can also be used to identify the subset of initial states from which the controller can be safely executed.

****

Automatic Data Augmentation for Generalization in Reinforcement Learning

Roberta Raileanu,Maxwell Goldstein,Denis Yarats,Ilya Kostrikov,Rob Fergus

Deep reinforcement learning (RL) agents often fail to generalize beyond their training environments. To alleviate this problem, recent work has proposed the use of data augmentation. However, different tasks tend to benefit from different types of augmentations and selecting the right one typically requires expert knowledge. In this paper, we introduce three approaches for automatically finding an effective augmentation for any RL task. These are combined with two novel regularization terms for the policy and value function, required to make the use of data augmentation theoretically sound for actor-critic algorithms. We evaluate our method on the Procgen benchmark which consists of 16 procedurally generated environments and show that it improves test performance by 40% relative to standard RL algorithms. Our approach also outperforms methods specifically designed to improve generalization in RL, thus setting a new state-of-the-art on Procgen. In addition, our agent learns policies and representations which are more robust to changes in the environment that are irrelevant for solving the task, such as the background.

****

What are the Statistical Limits of Offline RL with Linear Function Approximation?

Ruosong Wang,Dean Foster,Sham M. Kakade

Offline reinforcement learning seeks to utilize offline (observational) data to guide the learning of (causal) sequential decision making strategies. The hope is that offline reinforcement learning coupled with function approximation methods (to deal with the curse of dimensionality) can provide a means to help allevia

te the excessive sample complexity burden in modern sequential decision making p
roblems. However, the extent to which this broader approach can be effective is
not well understood, where the literature largely consists of sufficient conditi
ons.

This work focuses on the basic question of what are necessary representational a
nd distributional conditions that permit provable sample-efficient offline reinf
orcement learning. Perhaps surprisingly, our main result shows that even if: i)
we have realizability in that the true value function of \emph{every} policy is
linear in a given set of features and 2) our off-policy data has good coverage
over all features (under a strong spectral condition), any algorithm still (info
rmation-theoretically) requires a number of offline samples that is exponential
in the problem horizon to non-trivially estimate the value of \emph{any} given p
olicy. Our results highlight that sample-efficient offline policy evaluation is
not possible unless significantly stronger conditions hold; such conditions incl
ude either having low distribution shift (where the offline data distribution is
 close to the distribution of the policy to be evaluated) or significantly stron
ger representational conditions (beyond realizability).
****
WAVEQ: GRADIENT-BASED DEEP QUANTIZATION OF NEURAL NETWORKS THROUGH SINUSOIDAL RE
GULARIZATION
Ahmed T. Elthakeb,Prannoy Pilligundla,Tarek Elgindi,Fatemehsadat Mireshghallah,C
harles-Alban Deledalle,Hadi Esmaeilzadeh
Deep quantization of neural networks below eight bits can lead to superlinear be
nefits in storage and compute efficiency. However, homogeneously quantizing all
the layers to the same level does not account for the distinction of the layers
and their individual properties. Heterogenous assignment of bitwidths to individ
ual layers is attractive but opens an exponentially large non-contiguous hyperpa
rameter space (${Available Bitwidths}^{\# Layers}$). As such finding the bitwidt
h while also quantizing the network to those levels becomes a major challenge. T
his paper addresses this challenge through a sinusoidal regularization mechanism
, dubbed WaveQ. Adding our parametrized sinusoidal regularizer enables us to not
 only find the quantized weights but also learn the bitwidth of the layers by ma
king the period of the sinusoidal regularizer a trainable parameter. In addition
, the sinusoidal regularizer itself is designed to align its minima on the quant
ization levels. With these two innovations, during training, stochastic gradient
 descent uses the form of the sinusoidal regularizer and its minima to push the
weights to the quantization levels while it is also learning the period which wi
ll determine the bitwidth of each layer separately. As such WaveQ is a gradient-
based mechanism that jointly learns the quantized weights as well as the heterog
eneous bitwidths. We show how WaveQ balance compute efficiency and accuracy, and
 provide a heterogeneous bitwidth assignment for quantization of a large variety
 of deep networks (AlexNet, CIFAR-10, MobileNet, ResNet-18, ResNet-20, SVHN, and
 VGG-11) that virtually preserves the accuracy. WaveQ is versatile and can also
be used with predetermined bitwidths by fixing the period of the sinusoidal regu
larizer. In this case. WaveQ enhances quantized training algorithms (DoReFa and
WRPN) with about 4.8% accuracy improvements on average, and outperforms multiple
 state-of-the-art techniques. Finally, WaveQ applied to quantizing transformers

****
Revisiting BFfloat16 Training
Pedram Zamirai,Jian Zhang,Christopher R Aberger,Christopher De Sa
State-of-the-art generic low-precision training algorithms use a mix of 16-bit a
nd 32-bit precision, creating the folklore that 16-bit precision alone is not en
ough to maximize model accuracy. As a result, deep learning accelerators are for
ced to support both 16-bit and 32-bit compute units which is more costly than on
ly using 16-bit units for hardware design. We ask can we do pure 16-bit training
 which requires only 16-bit compute units, while still matching the model accura
cy attained by 32-bit training. Towards this end, we study pure 16-bit training
algorithms on the widely adopted BFloat16 compute unit. While these units conven

tionally use nearest rounding to cast output to 16-bit precision, we show that n earest rounding for model weight updates can often cancel small updates, which d egrades the convergence and model accuracy. Motivated by this, we identify two s imple existing techniques, stochastic rounding and the Kahan accumulation, to re medy the model accuracy degradation in pure 16-bit training. We empirically show that these two techniques can enable up to 7% absolute validation accuracy gain in pure 16-bit training. This leads to 0.1% lower to 0.2% higher matching valid ation accuracy compared to 32-bit precision training across seven deep learning applications.

****

Formal Language Constrained Markov Decision Processes

Eleanor Quint,Dong Xu,Samuel W Flint,Stephen D Scott,Matthew Dwyer

In order to satisfy safety conditions, an agent may be constrained from acting f reely. A safe controller can be designed a priori if an environment is well unde rstood, but not when learning is employed. In particular, reinforcement learned (RL) controllers require exploration, which can be hazardous in safety critical situations. We study the benefits of giving structure to the constraints of a co nstrained Markov decision process by specifying them in formal languages as a st ep towards using safety methods from software engineering and controller synthes is. We instantiate these constraints as finite automata to efficiently recognise constraint violations. Constraint states are then used to augment the underlyin g MDP state and to learn a dense cost function, easing the problem of quickly le arning joint MDP/constraint dynamics. We empirically evaluate the effect of thes e methods on training a variety of RL algorithms over several constraints specif ied in Safety Gym, MuJoCo, and Atari environments.

****

Latent Space Semi-Supervised Time Series Data Clustering

Andrew Hill,Katerina Kechris,Russell Bowler,Farnoush Kashani

Time series data is abundantly available in the real world, but there is a disti nct lack of large, labeled datasets available for many types of learning tasks. Semi-supervised models, which can leverage small amounts of expert-labeled data along with a larger unlabeled dataset, have been shown to improve performance ov er unsupervised learning models. Existing semi-supervised time series clustering algorithms suffer from lack of scalability as they are limited to perform learn ing operations within the original data space. We propose an autoencoder-based s emi-supervised learning model along with multiple semi-supervised objective func tions which can be used to improve the quality of the autoencoder’s learned late nt space via the addition of a small number of labeled examples. Experiments on a variety of datasets show that our methods can usually improve k-Means clusteri ng performance. Our methods achieve a maximum average ARI of 0.897, a 140% incre ase over an unsupervised CAE model. Our methods also achieve a maximum improveme nt of 44% over a semi-supervised model.

****

Bayesian Learning to Optimize: Quantifying the Optimizer Uncertainty

Yue Cao,Tianlong Chen,Zhangyang Wang,Yang Shen

Optimizing an objective function with uncertainty awareness is well-known to imp rove the accuracy and confidence of optimization solutions. Meanwhile, another r elevant but very different question remains yet open: how to model and quantify the uncertainty of an optimization algorithm itself? To close such a gap, the pr erequisite is to consider the optimizers as sampled from a distribution, rather than a few pre-defined and fixed update rules. We first take the novel angle to consider the algorithmic space of optimizers, each being parameterized by a neu ral network. We then propose a Boltzmann-shaped posterior over this optimizer sp ace, and approximate the posterior locally as Gaussian distributions through var iational inference. Our novel model, Bayesian learning to optimize (BL2O) is th e first study to recognize and quantify the uncertainty of the optimization algo rithm. Our experiments on optimizing test functions, energy functions in protein -protein interactions and loss functions in image classification and data privac y attack demonstrate that, compared to state-of-the-art methods, BL2O improves o

ptimization and uncertainty quantification (UQ) in aforementioned problems as well as calibration and out-of-domain detection in image classification.
****
Overcoming barriers to the training of effective learned optimizers
Luke Metz,Niru Maheswaranathan,C. Daniel Freeman,Ben Poole,Jascha Sohl-Dickstein
In this work we focus on general-purpose learned optimizers capable of training a wide variety of problems with no user-specified hyperparameters. We introduce a new, neural network parameterized, hierarchical optimizer with access to additional features such as validation loss to enable automatic regularization. Most learned optimizers have been trained on only a single task, or a small number of tasks. We train our optimizers on thousands of tasks, making use of orders of magnitude more compute, resulting in optimizers that generalize better to unseen tasks. The learned optimizers not only perform well, but learn behaviors that are distinct from existing first order optimizers. For instance, they generate update steps that have implicit regularization and adapt as the problem hyperparameters (e.g. batch size) or architecture (e.g. neural network width) change. Finally, these learned optimizers show evidence of being useful for out of distribution tasks such as training themselves from scratch.
****
Multi-Agent Imitation Learning with Copulas
Hongwei Wang,Lantao Yu,Zhangjie Cao,Stefano Ermon
Multi-agent imitation learning aims to train multiple agents to perform tasks from demonstrations by learning a mapping between observations and actions, which is essential for understanding physical, social, and team-play systems. However, most existing works on modeling multi-agent interactions typically assume that agents make independent decisions based on their observations, ignoring the complex dependence among agents. In this paper, we propose to use copula, a powerful statistical tool for capturing dependence among random variables, to explicitly model the correlation and coordination in multi-agent systems. Our proposed model is able to separately learn marginals that capture the local behavioral patterns of each individual agent, as well as a copula function that solely and fully captures the dependence structure among agents. Extensive experiments on synthetic and real-world datasets show that our model outperforms state-of-the-art baselines across various scenarios in the action prediction task, and is able to generate new trajectories close to expert demonstrations.
****
The geometry of integration in text classification RNNs
Kyle Aitken,Vinay Venkatesh Ramasesh,Ankush Garg,Yuan Cao,David Sussillo,Niru Maheswaranathan
Despite the widespread application of recurrent neural networks (RNNs), a unified understanding of how RNNs solve particular tasks remains elusive. In particular, it is unclear what dynamical patterns arise in trained RNNs, and how those pat-terns depend on the training dataset or task. This work addresses these questions in the context of text classification, building on earlier work studying the dynamics of binary sentiment-classification networks (Maheswaranathan et al., 2019). We study text-classification tasks beyond the binary case, exploring the dynamics ofRNNs trained on both natural and synthetic datasets. These dynamics, which we find to be both interpretable and low-dimensional, share a common mechanism across architectures and datasets: specifically, these text-classification networks use low-dimensional attractor manifolds to accumulate evidence for each class as they process the text. The dimensionality and geometry of the attractor manifold are determined by the structure of the training dataset, with the dimensionality reflecting the number of scalar quantities the network remembers in order to classify.In categorical classification, for example, we show that this dimensionality is one less than the number of classes. Correlations in the dataset, such as those induced by ordering, can further reduce the dimensionality of the attractor manifold; we show how to predict this reduction using simple word-count statistics computed on the training dataset. To the degree that integration of evidence towards a decision is a common computational primitive, this work continues to lay the foundation for using dynamical systems techniques to s

tudy the inner workings of RNNs.
****
Decomposing Mutual Information for Representation Learning
Alessandro Sordoni,Nouha Dziri,Hannes Schulz,Geoff Gordon,Remi Tachet des Combes,Philip Bachman
Many self-supervised representation learning methods maximize mutual information (MI) across views. In this paper, we transform each view into a set of subviews and then decompose the original MI bound into a sum of bounds involving conditional MI between the subviews. E.g.,~given two views x and y of the same input example, we can split x into two subviews, $x^{\prime}$ and $x^{\prime\prime}$, which depend only on x but are otherwise unconstrained. The following holds: $I(x; y) \geq I(x^{\prime\prime}; y) + I(x^{\prime}; y | x^{\prime\prime})$, due to the chain rule and information processing inequality. By maximizing both terms in the decomposition, our approach explicitly rewards the encoder for any information about y which it extracts from $x^{\prime\prime}$, and for information about y extracted from $x^{\prime}$ in excess of the information from $x^{\prime\prime}$. We provide a novel contrastive lower-bound on conditional MI, that relies on sampling contrast sets from $p(y|x^{\prime\prime})$. By decomposing the original MI into a sum of increasingly challenging MI bounds between sets of increasingly informed views, our representations can capture more of the total information shared between the original views. We empirically test the method in a vision domain and for dialogue generation.
****
Behavioral Cloning from Noisy Demonstrations
Fumihiro Sasaki,Ryota Yamashina
We consider the problem of learning an optimal expert behavior policy given noisy demonstrations that contain observations from both optimal and non-optimal expert behaviors. Popular imitation learning algorithms, such as generative adversarial imitation learning, assume that (clear) demonstrations are given from optimal expert policies but not the non-optimal ones, and thus often fail to imitate the optimal expert behaviors given the noisy demonstrations. Prior works that address the problem require (1) learning policies through environment interactions in the same fashion as reinforcement learning, and (2) annotating each demonstration with confidence scores or rankings. However, such environment interactions and annotations in real-world settings take impractically long training time and a significant human effort. In this paper, we propose an imitation learning algorithm to address the problem without any environment interactions and annotations associated with the non-optimal demonstrations. The proposed algorithm learns ensemble policies with a generalized behavioral cloning (BC) objective function where we exploit another policy already learned by BC. Experimental results show that the proposed algorithm can learn behavior policies that are much closer to the optimal policies than ones learned by BC.
****
Human-interpretable model explainability on high-dimensional data
Damien de Mijolla,Christopher Frye,Markus Kunesch,John Mansir,Ilya Feige
The importance of explainability in machine learning continues to grow, as both neural-network architectures and the data they model become increasingly complex. Unique challenges arise when a model's input features become high-dimensional: on one hand, principled model-agnostic approaches to explainability become too computationally expensive; on the other, more efficient explainability algorithms lack natural interpretations for general users. In this work, we introduce a framework for human-interpretable explainability on high-dimensional data, consisting of two modules. First, we apply a semantically-meaningful latent representation, both to reduce the raw dimensionality of the data, and to ensure its human interpretability. These latent features can be learnt, e.g. explicitly as disentangled representations or implicitly through image-to-image translation, or they can be based on any computable quantities the user chooses. Second, we adapt the Shapley paradigm for model-agnostic explainability to operate on these latent features. This leads to interpretable model explanations that are both theoretically controlled and computationally tractable. We benchmark our approach on syn

thetic data and demonstrate its effectiveness on several image-classification ta
sks.
****

SSW-GAN: Scalable Stage-wise Training of Video GANs
Lluis Castrejon,Nicolas Ballas,Aaron Courville
Current state-of-the-art generative models for videos have high computational re
quirements that impede high resolution generations beyond a few frames. In this
work we propose a stage-wise strategy to train Generative Adversarial Networks (
GANs) for videos. We decompose the generative process to first produce a downsa
mpled video that is then spatially upscaled and temporally interpolated by subse
quent stages. Upsampling stages are applied locally on temporal chunks of previo
us outputs to manage the computational complexity. Stages are defined as Generat
ive Adversarial Networks, which are trained sequentially and independently. We v
alidate our approach on Kinetics-600 and BDD100K, for which we train a three sta
ge model capable of generating 128x128 videos with 100 frames.
****

Towards Robust Neural Networks via Close-loop Control
Zhuotong Chen,Qianxiao Li,Zheng Zhang
Despite their success in massive engineering applications, deep neural networks
are vulnerable to various perturbations due to their black-box nature. Recent st
udy has shown that a deep neural network can misclassify the data even if the in
put data is perturbed by an imperceptible amount. In this paper, we address the
robustness issue of neural networks by a novel close-loop control method from th
e perspective of dynamic systems. Instead of modifying the parameters in a fixed
 neural network architecture, a close-loop control process is added to generate
control signals adaptively for the perturbed or corrupted data. We connect the r
obustness of neural networks with optimal control using the geometrical informat
ion of underlying data to design the control objective. The detailed analysis sh
ows how the embedding manifolds of state trajectory affect error estimation of t
he proposed method. Our approach can simultaneously maintain the performance on
clean data and improve the robustness against many types of data perturbations.
It can also further improve the performance of robustly trained neural networks
against different perturbations. To the best of our knowledge, this is the first
 work that improves the robustness of neural networks with close-loop control.
****

Projected Latent Markov Chain Monte Carlo: Conditional Sampling of Normalizing F
lows
Chris Cannella,Mohammadreza Soltani,Vahid Tarokh
We introduce Projected Latent Markov Chain Monte Carlo (PL-MCMC), a technique fo
r sampling from the exact conditional distributions learned by normalizing flows
. As a conditional sampling method, PL-MCMC enables Monte Carlo Expectation Maxi
mization (MC-EM) training of normalizing flows from incomplete data. Through exp
erimental tests applying normalizing flows to missing data tasks for a variety o
f data sets, we demonstrate the efficacy of PL-MCMC for conditional sampling fro
m normalizing flows.
****

Network Architecture Search for Domain Adaptation
Yichen Li,Xingchao Peng
Deep networks have been used to learn transferable representations for domain ad
aptation. Existing deep domain adaptation methods systematically employ popular
hand-crafted networks designed specifically for image-classification tasks, lead
ing to sub-optimal domain adaptation performance. In this paper, we present Neur
al Architecture Search for Domain Adaptation (NASDA), a principle framework that
 leverages differentiable neural architecture search to derive the optimal netwo
rk architecture for domain adaptation task. NASDA is designed with two novel tra
ining strategies: neural architecture search with multi-kernel Maximum Mean Disc
repancy to derive the optimal architecture, and adversarial training between a f
eature generator and a batch of classifiers to consolidate the feature generator
. We demonstrate experimentally that NASDA leads to state-of-the-art performance
 on several domain adaptation benchmarks.

****
How Does Mixup Help With Robustness and Generalization?
Linjun Zhang,Zhun Deng,Kenji Kawaguchi,Amirata Ghorbani,James Zou
Mixup is a popular data augmentation technique based on on convex combinations of pairs of examples and their labels. This simple technique has shown to substantially improve both the model's robustness as well as the generalization of the trained model. However, it is not well-understood why such improvement occurs. In this paper, we provide theoretical analysis to demonstrate how using Mixup in training helps model robustness and generalization. For robustness, we show that minimizing the Mixup loss corresponds to approximately minimizing an upper bound of the adversarial loss. This explains why models obtained by Mixup training exhibits robustness to several kinds of adversarial attacks such as Fast Gradient Sign Method (FGSM). For generalization, we prove that Mixup augmentation corresponds to a specific type of data-adaptive regularization which reduces overfitting. Our analysis provides new insights and a framework to understand Mixup.

****
Understanding the failure modes of out-of-distribution generalization
Vaishnavh Nagarajan,Anders Andreassen,Behnam Neyshabur
Empirical studies suggest that machine learning models often rely on features, such as the background, that may be spuriously correlated with the label only during training time, resulting in poor accuracy during test-time. In this work, we identify the fundamental factors that give rise to this behavior, by explaining why models fail this way even in easy-to-learn tasks where one would expect these models to succeed. In particular, through a theoretical study of gradient-descent-trained linear classifiers on some easy-to-learn tasks, we uncover two complementary failure modes. These modes arise from how spurious correlations induce two kinds of skews in the data: one geometric in nature and another, statistical. Finally, we construct natural modifications of image classification datasets to understand when these failure modes can arise in practice. We also design experiments to isolate the two failure modes when training modern neural networks on these datasets.
****
CTRLsum: Towards Generic Controllable Text Summarization
Junxian He,Wojciech Maciej Kryscinski,Bryan McCann,Nazneen Rajani,Caiming Xiong
Current summarization systems yield generic summaries that are disconnected from users' preferences and expectations. To address this limitation, we present CTRLsum, a novel framework for controllable summarization. Our approach enables users to control multiple aspects of generated summaries by interacting with the summarization system through textual input in the form of a set of keywords or descriptive prompts. Using a single unified model, CTRLsum is able to achieve a broad scope of summary manipulation at inference time without requiring additional human annotations or pre-defining a set of control aspects during training. We quantitatively demonstrate the effectiveness of our approach on three domains of summarization datasets and five control aspects: 1) entity-centric and 2) length-controllable summarization, 3) contribution summarization on scientific papers, 4) invention purpose summarization on patent filings, and 5) question-guided summarization on news articles in a reading comprehension setting. Moreover, when used in a standard, uncontrolled summarization setting, CTRLsum achieves state-of-the-art results on the CNN/DailyMail dataset.
****
Search Data Structure Learning
Mathieu Duchesneau,Hansenclever Bassani,Alain Tapp
In our modern world, an enormous amount of data surrounds us, and we are rarely interested in more than a handful of data points at once. It is like searching for needles in a haystack, and in many cases, there is no better algorithm than a random search, which might not be viable. Previously proposed algorithms for efficient database access are made for particular applications such as finding the min/max, finding all points within a range or finding the k-nearest neighbours. Consequently, there is a lack of versatility concerning what we can search when

it comes to a gigantic database. In this work, we propose Search Data Structure Learning (SDSL), a generalization of the standard Search Data Structure (SDS) in which the machine has to learn how to search in the database. To evaluate approaches in this field, we propose a novel metric called Sequential Search Work Ratio (SSWR), a natural way of measuring a search's efficiency and quality. Finally, we inaugurate the field with the Efficient Learnable Binary Access (ELBA), a family of models for Search Data Structure Learning. It requires a means to train two parametric functions and a search data structure for binary codes. For the training, we developed a novel loss function, the F-beta Loss. For the SDS, we describe the Multi-Bernoulli Search (MBS), a novel approach for probabilistic binary codes. Finally, we exhibit the F-beta Loss and the MBS synergy by experimentally showing that it is at least twice as better than using the alternative loss functions of MIHash and HashNet and twenty times better than with another SDS based on the Hamming radius.

****

Usable Information and Evolution of Optimal Representations During Training

Michael Kleinman,Alessandro Achille,Daksh Idnani,Jonathan Kao

We introduce a notion of usable information contained in the representation learned by a deep network, and use it to study how optimal representations for the task emerge during training. We show that the implicit regularization coming from training with Stochastic Gradient Descent with a high learning-rate and small batch size plays an important role in learning minimal sufficient representations for the task. In the process of arriving at a minimal sufficient representation, we find that the content of the representation changes dynamically during training. In particular, we find that semantically meaningful but ultimately irrelevant information is encoded in the early transient dynamics of training, before being later discarded. In addition, we evaluate how perturbing the initial part of training impacts the learning dynamics and the resulting representations. We show these effects on both perceptual decision-making tasks inspired by neuroscience literature, as well as on standard image classification tasks.

****

NNGeometry: Easy and Fast Fisher Information Matrices and Neural Tangent Kernels in PyTorch

Thomas George

Fisher Information Matrices (FIM) and Neural Tangent Kernels (NTK) are useful tools in a number of diverse applications related to neural networks. Yet these theoretical tools are often difficult to implement using current libraries for practical size networks, given that they require per-example gradients, and a large amount of memory since they scale as the number of parameters (for the FIM) or the number of examples x cardinality of the output space (for the NTK). NNGeometry is a PyTorch library that offers a simple interface for computing various linear algebra operations such as matrix-vector products, trace, frobenius norm, and so on, where the matrix is either the FIM or the NTK, leveraging recent advances in approximating these matrices. We here present the library and motivate our design choices, then we demonstrate it on actual deep neural networks.

****

Adaptive Extra-Gradient Methods for Min-Max Optimization and Games

Kimon Antonakopoulos,Veronica Belmega,Panayotis Mertikopoulos

We present a new family of min-max optimization algorithms that automatically exploit the geometry of the gradient data observed at earlier iterations to perform more informative extra-gradient steps in later ones.

Thanks to this adaptation mechanism, the proposed method automatically detects whether the problem is smooth or not, without requiring any prior tuning by the optimizer.

As a result, the algorithm simultaneously achieves order-optimal convergence rates, \ie it converges to an ε-optimal solution within $\mathcal{O}(1/\varepsilon)$ iterations in smooth problems, and within $\mathcal{O}(1/\varepsilon^2)$ iterations in non-smooth ones. Importantly, these guarantees do not require any of the standard boundedness or Lipschitz continuity conditions that are typically assumed in the literature; in particular, they apply even to problems

with singularities (such as resource allocation problems and the like). This ada
ptation is achieved through the use of a geometric apparatus based on Finsler me
trics and a suitably chosen mirror-prox template that allows us to derive sharp
convergence rates for the methods at hand.
****

Emergent Symbols through Binding in External Memory
Taylor Whittington Webb,Ishan Sinha,Jonathan Cohen
A key aspect of human intelligence is the ability to infer abstract rules direct
ly from high-dimensional sensory data, and to do so given only a limited amount
of training experience. Deep neural network algorithms have proven to be a power
ful tool for learning directly from high-dimensional data, but currently lack th
is capacity for data-efficient induction of abstract rules, leading some to argu
e that symbol-processing mechanisms will be necessary to account for this capaci
ty. In this work, we take a step toward bridging this gap by introducing the Eme
rgent Symbol Binding Network (ESBN), a recurrent network augmented with an exter
nal memory that enables a form of variable-binding and indirection. This binding
 mechanism allows symbol-like representations to emerge through the learning pro
cess without the need to explicitly incorporate symbol-processing machinery, ena
bling the ESBN to learn rules in a manner that is abstracted away from the parti
cular entities to which those rules apply. Across a series of tasks, we show tha
t this architecture displays nearly perfect generalization of learned rules to n
ovel entities given only a limited number of training examples, and outperforms
a number of other competitive neural network architectures.
****

Online Learning of Graph Neural Networks: When Can Data Be Permanently Deleted
Lukas Paul Achatius Galke,Benedikt Franke,Tobias Zielke,Ansgar Scherp
Online learning of graph neural networks (GNNs) faces the challenges of distribu
tion shift and ever growing and changing training data, when temporal graphs evo
lve over time. This makes it inefficient to train over the complete graph whenev
er new data arrives. Deleting old data at some point in time may be preferable t
o maintain a good performance and to account for distribution shift. We systemat
ically analyze these issues by incrementally training and evaluating GNNs in a s
liding window over temporal graphs. We experiment with three representative GNN
architectures and two scalable GNN techniques, on three new datasets. In our exp
eriments, the GNNs face the challenge that new vertices, edges, and even classes
 appear and disappear over time. Our results show that no more than 50% of the
GNN's receptive field is necessary to retain at least 95% accuracy compared to t
raining over a full graph. In most cases, i.e., 14 out 18 experiments, we even o
bserve that a temporal window of size 1 is sufficient to retain at least 90%.
****

Optimistic Policy Optimization with General Function Approximations
Qi Cai,Zhuoran Yang,Csaba Szepesvari,Zhaoran Wang
Although policy optimization with neural networks has a track record of achievin
g state-of-the-art results in reinforcement learning on various domains, the the
oretical understanding of the computational and sample efficiency of policy opti
mization remains restricted to linear function approximations with finite-dimens
ional feature representations, which hinders the design of principled, effective
, and efficient algorithms. To this end, we propose an optimistic model-based po
licy optimization algorithm, which allows general function approximations while
incorporating~exploration. In the episodic setting, we establish a $\sqrt{T}$-re
gret that scales polynomially in the eluder dimension of the general model class
. Here T is the number of steps taken by the agent. In particular, we speciali
ze such a regret to handle two nonparametric model classes; one based on reprodu
cing kernel Hilbert spaces and another based on overparameterized neural network
s.
****

Shapley explainability on the data manifold
Christopher Frye,Damien de Mijolla,Tom Begley,Laurence Cowton,Megan Stanley,Ilya
 Feige
Explainability in AI is crucial for model development, compliance with regulatio

n, and providing operational nuance to predictions. The Shapley framework for ex
plainability attributes a model’s predictions to its input features in a mathema
tically principled and model-agnostic way. However, general implementations of S
hapley explainability make an untenable assumption: that the model’s features ar
e uncorrelated. In this work, we demonstrate unambiguous drawbacks of this assum
ption and develop two solutions to Shapley explainability that respect the data
manifold. One solution, based on generative modelling, provides flexible access
to data imputations; the other directly learns the Shapley value-function, provi
ding performance and stability at the cost of flexibility. While “off-manifold”
Shapley values can (i) give rise to incorrect explanations, (ii) hide implicit m
odel dependence on sensitive attributes, and (iii) lead to unintelligible explan
ations in higher-dimensional data, on-manifold explainability overcomes these pr
oblems.

****
Learning Lagrangian Fluid Dynamics with Graph Neural Networks
Zijie Li,Amir Barati Farimani
We present a data-driven model for fluid simulation under Lagrangian representat
ion. Our model uses graphs to describe the fluid field, where physical quantitie
s are encoded as node and edge features. Instead of directly predicting the acce
leration or position correction given the current state, we decompose the simula
tion scheme into separate parts - advection, collision, and pressure projection.
 For these different reasoning tasks, we propose two kinds of graph neural netwo
rk structures, node-focused networks, and edge-focused networks. By introducing
physics prior knowledge, our model can be efficient in terms of training and inf
erence. Our tests show that the learned model can produce accurate results and r
emain stable in scenarios with a large amount of particles and different geometr
ies. Unlike many previous works, further tests demonstrate that our model is abl
e to retain many important physical properties of incompressible fluids, such as
 minor divergence and reasonable pressure distribution. Additionally, our model
can adopt a range of time step sizes different from ones using in the training s
et, which indicates its robust generalization capability.
****
Reinforcement Learning with Random Delays
Yann Bouteiller,Simon Ramstedt,Giovanni Beltrame,Christopher Pal,Jonathan Binas
Action and observation delays commonly occur in many Reinforcement Learning appl
ications, such as remote control scenarios. We study the anatomy of randomly del
ayed environments, and show that partially resampling trajectory fragments in hi
ndsight allows for off-policy multi-step value estimation. We apply this princip
le to derive Delay-Correcting Actor-Critic (DCAC), an algorithm based on Soft Ac
tor-Critic with significantly better performance in environments with delays. Th
is is shown theoretically and also demonstrated practically on a delay-augmented
 version of the MuJoCo continuous control benchmark.
****
On Proximal Policy Optimization's Heavy-Tailed Gradients
Saurabh Garg,Joshua Zhanson,Emilio Parisotto,Adarsh Prasad,J Zico Kolter,Zachary
 Chase Lipton,Sivaraman Balakrishnan,Ruslan Salakhutdinov,Pradeep Kumar Ravikuma
r
Modern policy gradient algorithms, notably Proximal Policy Optimization (PPO), r
ely on an arsenal of heuristics, including loss clipping and gradient clipping,
to ensure successful learning. These heuristics are reminiscent of techniques fr
om robust statistics, commonly used for estimation in outlier-rich ("heavy-taile
d") regimes. In this paper, we present a detailed empirical study to characteriz
e the heavy-tailed nature of the gradients of the PPO surrogate reward function.

We demonstrate pronounced heavy-tailedness of the gradients, specifically for th
e actor network, which increases as the current policy diverges from the behavi
oral one (i.e., as the agent goes further off policy). Further examination impl
icates the likelihood ratios and advantages in the surrogate reward as the main
sources to the observed heavy-tailedness. Subsequently, we study the effects of

the standard PPO clipping heuristics, demonstrating how these tricks primarily s erve to offset heavy-tailedness in gradients. Motivated by these connections, we propose incorporating GMOM (a high-dimensional robust estimator) into PPO as a substitute for three clipping tricks, achieving performance close to PPO (with a ll heuristics enabled) on a battery of MuJoCo continuous control tasks.
****

Individually Fair Gradient Boosting
Alexander Vargo,Fan Zhang,Mikhail Yurochkin,Yuekai Sun
We consider the task of enforcing individual fairness in gradient boosting. Grad ient boosting is a popular method for machine learning from tabular data, which arise often in applications where algorithmic fairness is a concern. At a high l evel, our approach is a functional gradient descent on a (distributionally) robu st loss function that encodes our intuition of algorithmic fairness for the ML t ask at hand. Unlike prior approaches to individual fairness that only work with smooth ML models, our approach also works with non-smooth models such as decisio n trees. We show that our algorithm converges globally and generalizes. We also demonstrate the efficacy of our algorithm on three ML problems susceptible to al gorithmic bias.
****

Shape or Texture: Understanding Discriminative Features in CNNs
Md Amirul Islam,Matthew Kowal,Patrick Esser,Sen Jia,Björn Ommer,Konstantinos G. Derpanis,Neil Bruce
Contrasting the previous evidence that neurons in the later layers of a Convolut ional Neural Network (CNN) respond to complex object shapes, recent studies have shown that CNNs actually exhibit a 'texture bias': given an image with both tex ture and shape cues (e.g., a stylized image), a CNN is biased towards predicting the category corresponding to the texture. However, these previous studies cond uct experiments on the final classification output of the network, and fail to r obustly evaluate the bias contained (i) in the latent representations, and (ii) on a per-pixel level. In this paper, we design a series of experiments that over come these issues. We do this with the goal of better understanding what type of shape information contained in the network is discriminative, where shape infor mation is encoded, as well as when the network learns about object shape during training. We show that a network learns the majority of overall shape informatio n at the first few epochs of training and that this information is largely encod ed in the last few layers of a CNN. Finally, we show that the encoding of shape does not imply the encoding of localized per-pixel semantic information. The exp erimental results and findings provide a more accurate understanding of the beha viour of current CNNs, thus helping to inform future design choices.
****

Learning to Reason in Large Theories without Imitation
Kshitij Bansal,Christian Szegedy,Markus Norman Rabe,Sarah M. Loos,Viktor Toman
In this paper, we demonstrate how to do automated higher-order logic theorem pro ving in the presence of a large knowledge base of potential premises without lea rning from human proofs. We augment the exploration of premises based on a simpl e tf-idf (term frequency-inverse document frequency) based lookup in a deep rein forcement learning scenario. Our experiments show that our theorem prover traine d with this exploration mechanism but no human proofs, dubbed DeepHOL Zero, outp erforms provers that are trained only on human proofs. It approaches the perform ance of a prover trained by a combination of imitation and reinforcement learnin g. We perform multiple experiments to understand the importance of the underlyin g assumptions that make our exploration approach work, thus explaining our desig n choices.
****

NOVAS: Non-convex Optimization via Adaptive Stochastic Search for End-to-end Lea rning and Control
Ioannis Exarchos,Marcus Aloysius Pereira,Ziyi Wang,Evangelos Theodorou
In this work we propose the use of adaptive stochastic search as a building bloc k for general, non-convex optimization operations within deep neural network arc hitectures. Specifically, for an objective function located at some layer in the

network and parameterized by some network parameters, we employ adaptive stocha
stic search to perform optimization over its output. This operation is different
iable and does not obstruct the passing of gradients during backpropagation, thu
s enabling us to incorporate it as a component in end-to-end learning. We study
the proposed optimization module's properties and benchmark it against two exist
ing alternatives on a synthetic energy-based structured prediction task, and fur
ther showcase its use in stochastic optimal control applications.
****

PolyRetro: Few-shot Polymer Retrosynthesis via Domain Adaptation
Binghong Chen,Chengtao Li,Hanjun Dai,Rampi Ramprasad,Le Song
Polymers appear everywhere in our daily lives -- fabrics, plastics, rubbers, etc
. -- and we could hardly live without them. To make polymers, chemists develop
processes that combine smaller building blocks~(monomers) to form long chains or
 complex networks~(polymers). These processes are called polymerizations and wil
l usually take lots of human efforts to develop. Although machine learning model
s for small molecules have generated lots of promising results, the prediction p
roblem for polymerization is new and suffers from the scarcity of polymerization
 datasets available in the field. Furthermore, the problem is made even more cha
llenging by the large size of the polymers and the additional recursive constrai
nts, which are not present in the small molecule problem. In this paper, we make
 an initial step towards this challenge and propose a learning-based search fram
ework that can automatically identify a sequence of reactions that lead to the p
olymerization of a target polymer with minimal polymerization data involved. Our
 method transfers models trained on small molecule datasets for retrosynthesis t
o check the validity of polymerization reaction. Furthermore, our method also in
corporates a template prior learned on a limited amount of polymer data into the
 framework to adapt the model from small molecule to the polymer domain. We demo
nstrate that our method is able to propose high-quality polymerization plans for
 a dataset of 52 real-world polymers, of which a significant portion successfull
y recovers the currently-in-used polymerization processes in the real world.
****

SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness
Mikhail Yurochkin,Yuekai Sun
In this paper, we cast fair machine learning as invariant machine learning. We f
irst formulate a version of individual fairness that enforces invariance on cert
ain sensitive sets. We then design a transport-based regularizer that enforces t
his version of individual fairness and develop an algorithm to minimize the regu
larizer efficiently. Our theoretical results guarantee the proposed approach tra
ins certifiably fair ML models. Finally, in the experimental studies we demonstr
ate improved fairness metrics in comparison to several recent fair training proc
edures on three ML tasks that are susceptible to algorithmic bias.
****

Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data
Colin Wei,Kendrick Shen,Yining Chen,Tengyu Ma
Self-training algorithms, which train a model to fit pseudolabels predicted by a
nother previously-learned model, have been very successful for learning with unl
abeled data using neural networks. However, the current theoretical understandin
g of self-training only applies to linear models. This work provides a unified t
heoretical analysis of self-training with deep networks for semi-supervised lear
ning, unsupervised domain adaptation, and unsupervised learning. At the core of
our analysis is a simple but realistic “expansion” assumption, which states that
 a low-probability subset of the data must expand to a neighborhood with large p
robability relative to the subset. We also assume that neighborhoods of examples
 in different classes have minimal overlap. We prove that under these assumption
s, the minimizers of population objectives based on self-training and input-cons
istency regularization will achieve high accuracy with respect to ground-truth l
abels. By using off-the-shelf generalization bounds, we immediately convert this
 result to sample complexity guarantees for neural nets that are polynomial in t
he margin and Lipschitzness. Our results help explain the empirical successes of
 recently proposed self-training algorithms which use input consistency regulari

zation.
****
Offline policy selection under Uncertainty
Mengjiao Yang,Bo Dai,Ofir Nachum,George Tucker,Dale Schuurmans
The presence of uncertainty in policy evaluation significantly complicates the process of policy ranking and selection in real-world settings. We formally consider offline policy selection as learning preferences over a set of policy prospects given a fixed experience dataset. While one can select or rank policies based on point estimates of their policy values or high-confidence intervals, access to the full distribution over one's belief of the policy value enables more flexible selection algorithms under a wider range of downstream evaluation metrics. We propose BayesDICE for estimating this belief distribution in terms of posteriors of distribution correction ratios derived from stochastic constraints (as opposed to explicit likelihood, which is not available). Empirically, BayesDICE is highly competitive to existing state-of-the-art approaches in confidence interval estimation. More importantly, we show how the belief distribution estimated by BayesDICE may be used to rank policies with respect to any arbitrary downstream policy selection metric, and we empirically demonstrate that this selection procedure significantly outperforms existing approaches, such as ranking policies according to mean or high-confidence lower bound value estimates.
****
Negative Data Augmentation
Abhishek Sinha,Kumar Ayush,Jiaming Song,Burak Uzkent,Hongxia Jin,Stefano Ermon
Data augmentation is often used to enlarge datasets with synthetic samples generated in accordance with the underlying data distribution. To enable a wider range of augmentations, we explore negative data augmentation strategies (NDA) that intentionally create out-of-distribution samples. We show that such negative out-of-distribution samples provide information on the support of the data distribution, and can be leveraged for generative modeling and representation learning. We introduce a new GAN training objective where we use NDA as an additional source of synthetic data for the discriminator. We prove that under suitable conditions, optimizing the resulting objective still recovers the true data distribution but can directly bias the generator towards avoiding samples that lack the desired structure. Empirically, models trained with our method achieve improved conditional/unconditional image generation along with improved anomaly detection capabilities. Further, we incorporate the same negative data augmentation strategy in a contrastive learning framework for self-supervised representation learning on images and videos, achieving improved performance on downstream image classification, object detection, and action recognition tasks. These results suggest that prior knowledge on what does not constitute valid data is an effective form of weak supervision across a range of unsupervised learning tasks.
****
Boundary Effects in CNNs: Feature or Bug?
Md Amirul Islam,Matthew Kowal,Sen Jia,Konstantinos G. Derpanis,Neil Bruce
Recent studies have shown that the addition of zero padding drives convolutional neural networks (CNNs) to encode a significant amount of absolute position information in their internal representations, while a lack of padding precludes position encoding. Additionally, various studies have used image patches on background canvases (e.g., to accommodate that inputs to CNNs must be rectangular) without consideration that different backgrounds may contain varying levels of position information according to their color. These studies give rise to deeper questions about the role of boundary information in CNNs, that are explored in this paper: (i) What boundary heuristics (e.g., padding type, canvas color) enable optimal encoding of absolute position information for a particular downstream task?; (ii) Where in the latent representations do boundary effects destroy semantic and location information?; (iii) Does encoding position information affect the learning of semantic representations?; (iv) Does encoding position information always improve performance? To provide answers to these questions, we perform the largest case study to date on the role that padding and border heuristics play in CNNs. We first show that zero padding injects optimal position information in

to CNNs relative to other common padding types. We then design a series of novel tasks which allow us to accurately quantify boundary effects as a function of the distance to the border. A number of semantic objectives reveal the destructive effect of dealing with the border on semantic representations. Further, we demonstrate that the encoding of position information improves separability of learned semantic features. Finally, we demonstrate the implications of these findings on a number of real-world tasks to show that position information can act as a feature or a bug.
****
Molecule Optimization by Explainable Evolution
Binghong Chen,Tianzhe Wang,Chengtao Li,Hanjun Dai,Le Song
Optimizing molecules for desired properties is a fundamental yet challenging task in chemistry, material science, and drug discovery. This paper develops a novel algorithm for optimizing molecular properties via an Expectation-Maximization (EM) like explainable evolutionary process. The algorithm is designed to mimic human experts in the process of searching for desirable molecules and alternate between two stages: the first stage on explainable local search which identifies rationales, i.e., critical subgraph patterns accounting for desired molecular properties, and the second stage on molecule completion which explores the larger space of molecules containing good rationales. We test our approach against various baselines on a real-world multi-property optimization task where each method is given the same number of queries to the property oracle. We show that our evolution-by-explanation algorithm is 79% better than the best baseline in terms of a generic metric combining aspects such as success rate, novelty, and diversity. Human expert evaluation on optimized molecules shows that 60% of top molecules obtained from our methods are deemed successful.
****
Learning to Plan Optimistically: Uncertainty-Guided Deep Exploration via Latent Model Ensembles
Tim Seyde,Wilko Schwarting,Sertac Karaman,Daniela Rus
Learning complex behaviors through interaction requires coordinated long-term planning. Random exploration and novelty search lack task-centric guidance and waste effort on non-informative interactions. Instead, decision making should target samples with the potential to optimize performance far into the future, while only reducing uncertainty where conducive to this objective. This paper presents latent optimistic value exploration (LOVE), a strategy that enables deep exploration through optimism in the face of uncertain long-term rewards. We combine finite-horizon rollouts from a latent model with value function estimates to predict infinite-horizon returns and recover associated uncertainty through ensembling. Policy training then proceeds on an upper confidence bound (UCB) objective to identify and select the interactions most promising to improve long-term performance. We apply LOVE to continuous visual control tasks and demonstrate improved sample complexity on a selection of benchmarking tasks.
****
Consistent Instance Classification for Unsupervised Representation Learning
Depu Meng,Zigang Geng,Zhirong Wu,Bin Xiao,Houqiang Li,Jingdong Wang
In this paper, we address the problem of learning the representations from images without human annotations. We study the instance classification solution, which regards each instance as a category, and improve the optimization and feature quality. The proposed consistent instance classification (ConIC) approach simultaneously optimizes the classification loss and an additional consistency loss explicitly penalizing the feature dissimilarity between the augmented views from the same instance. The benefit of optimizing the consistency loss is that the learned features for augmented views from the same instance are more compact and accordingly the classification loss optimization becomes easier, thus boosting the quality of the learned representations. This differs from InstDisc and MoCo that use an estimated prototype as the classifier weight to ease the optimization. Different from SimCLR that directly compares different instances, our approach does not require large batch size. Experimental results demonstrate competitive performance for linear evaluation and better performance than InstDisc, MoCo and

SimCLR at downstream tasks, such as detection and segmentation, as well as competitive or superior performance compared to other methods with stronger training setting.
****
GraphLog: A Benchmark for Measuring Logical Generalization in Graph Neural Networks
Koustuv Sinha,Shagun Sodhani,Joelle Pineau,William L. Hamilton
Relational inductive biases have a key role in building learning agents that can generalize and reason in a compositional manner. While relational learning algorithms such as graph neural networks (GNNs) show promise, we do not understand their effectiveness to adapt to new tasks. In this work, we study the logical generalization capabilities of GNNs by designing a benchmark suite grounded in first-order logic. Our benchmark suite, GraphLog, requires that learning algorithms perform rule induction in different synthetic logics, represented as knowledge graphs.
GraphLog consists of relation prediction tasks on 57 distinct procedurally generated logical worlds. We use GraphLog to evaluate GNNs in three different setups: single-task supervised learning, multi-task (with pretraining), and continual learning. Unlike previous benchmarks, GraphLog enables us to precisely control the logical relationship between the different worlds by controlling the underlying first-order logic rules. We find that models' ability to generalize and adapt strongly correlates to the availability of diverse sets of logical rules during multi-task training. We also find the severe catastrophic forgetting effect in continual learning scenarios, and GraphLog provides a precise mechanism to control the distribution shift. Overall, our results highlight new challenges for the design of GNN models, opening up an exciting area of research in generalization using graph-structured data.
****
Enhanced First and Zeroth Order Variance Reduced Algorithms for Min-Max Optimization
Tengyu Xu,Zhe Wang,Yingbin Liang,H. Vincent Poor
Min-max optimization captures many important machine learning problems such as robust adversarial learning and inverse reinforcement learning, and nonconvex-strongly-concave min-max optimization has been an active line of research. Specifically, a novel variance reduction algorithm SREDA was proposed recently by (Luo et al. 2020) to solve such a problem, and was shown to achieve the optimal complexity dependence on the required accuracy level ϵ. Despite the superior theoretical performance, the convergence guarantee of SREDA requires stringent initialization accuracy and an ϵ-dependent stepsize for controlling the per-iteration progress, so that SREDA can run very slowly in practice. This paper develops a novel analytical framework that guarantees the SREDA's optimal complexity performance for a much enhanced algorithm SREDA-Boost, which has less restrictive initialization requirement and an accuracy-independent (and much bigger) stepsize. Hence, SREDA-Boost runs substantially faster in experiments than SREDA. We further apply SREDA-Boost to propose a zeroth-order variance reduction algorithm named ZO-SREDA-Boost for the scenario that has access only to the information about function values not gradients, and show that ZO-SREDA-Boost outperforms the best known complexity dependence on ϵ. This is the first study that applies the variance reduction technique to zeroth-order algorithm for min-max optimization problems.
****
Estimating Lipschitz constants of monotone deep equilibrium models
Chirag Pabbaraju,Ezra Winston,J Zico Kolter
Several methods have been proposed in recent years to provide bounds on the Lipschitz constants of deep networks, which can be used to provide robustness guarantees, generalization bounds, and characterize the smoothness of decision boundaries. However, existing bounds get substantially weaker with increasing depth of the network, which makes it unclear how to apply such bounds to recently proposed models such as the deep equilibrium (DEQ) model, which can be viewed as representing an infinitely-deep network. In this paper, we show that monotone DEQs, a

recently-proposed subclass of DEQs, have Lipschitz constants that can be bounded as a simple function of the strong monotonicity parameter of the network. We derive simple-yet-tight bounds on both the input-output mapping and the weight-output mapping defined by these networks, and demonstrate that they are small relative to those for comparable standard DNNs. We show that one can use these bounds to design monotone DEQ models, even with e.g. multi-scale convolutional structure, that still have constraints on the Lipschitz constant. We also highlight how to use these bounds to develop PAC-Bayes generalization bounds that do not depend on any depth of the network, and which avoid the exponential depth-dependence of comparable DNN bounds.
****

A Transformer-based Framework for Multivariate Time Series Representation Learning
George Zerveas,Srideepika Jayaraman,Dhaval Patel,Anuradha Bhamidipaty,Carsten Eickhoff
In this work we propose for the first time a transformer-based framework for unsupervised representation learning of multivariate time series. Pre-trained models can be potentially used for downstream tasks such as regression and classification, forecasting and missing value imputation. We evaluate our models on several benchmark datasets for multivariate time series regression and classification and show that they exceed current state-of-the-art performance, even when the number of training samples is very limited, while at the same time offering computational efficiency. We show that unsupervised pre-training of our transformer models offers a substantial performance benefit over fully supervised learning, even without leveraging additional unlabeled data, i.e., by reusing the same data samples through the unsupervised objective.
****

Learning from multiscale wavelet superpixels using GNN with spatially heterogeneous pooling
Maxime Bassenne,Varun Vasudevan,Lei Xing
Neural networks have become the standard for image classification tasks. On one hand, convolutional neural networks (CNNs) achieve state-of-the-art performance by learning from a regular grid representation of images. On the other hand, graph neural networks (GNNs) have shown promise in learning image classification from an embedded superpixel graph. However, in the latter, studies have been restricted to SLIC superpixels, where 1) a single target number of superpixels is arbitrarily defined for an entire dataset irrespective of differences across images and 2) the superpixels in a given image are of similar size despite intrinsic multiscale structure. In this study, we investigate learning from a new principled representation in which individual images are represented by an image-specific number of multiscale superpixels. We propose WaveMesh, a wavelet-based superpixeling algorithm, where the number and sizes of superpixels in an image are systematically computed based on the image content. We also present WavePool, a spatially heterogeneous pooling scheme tailored to WaveMesh superpixels. We study the feasibility of learning from the WaveMesh superpixel representation using SplineCNN, a state-of-the-art network for image graph classification. We show that under the same network architecture and training settings, SplineCNN with original Graclus-based pooling learns from WaveMesh superpixels on-par with SLIC superpixels. Additionally, we observe that the best performance is achieved when replacing Graclus-based pooling with WavePool while using WaveMesh superpixels.
****

Implicit Gradient Regularization
David Barrett,Benoit Dherin
Gradient descent can be surprisingly good at optimizing deep neural networks without overfitting and without explicit regularization. We find that the discrete steps of gradient descent implicitly regularize models by penalizing gradient descent trajectories that have large loss gradients. We call this Implicit Gradient Regularization (IGR) and we use backward error analysis to calculate the size of this regularization. We confirm empirically that implicit gradient regularization biases gradient descent toward flat minima, where test errors are small and

solutions are robust to noisy parameter perturbations. Furthermore, we demonstrate that the implicit gradient regularization term can be used as an explicit regularizer, allowing us to control this gradient regularization directly. More broadly, our work indicates that backward error analysis is a useful theoretical approach to the perennial question of how learning rate, model size, and parameter regularization interact to determine the properties of overparameterized models optimized with gradient descent.
****
Structured Prediction as Translation between Augmented Natural Languages
Giovanni Paolini,Ben Athiwaratkun,Jason Krone,Jie Ma,Alessandro Achille,RISHITA ANUBHAI,Cicero Nogueira dos Santos,Bing Xiang,Stefano Soatto
We propose a new framework, Translation between Augmented Natural Languages (TANL), to solve many structured prediction language tasks including joint entity and relation extraction, nested named entity recognition, relation classification, semantic role labeling, event extraction, coreference resolution, and dialogue state tracking. Instead of tackling the problem by training task-specific discriminative classifiers, we frame it as a translation task between augmented natural languages, from which the task-relevant information can be easily extracted. Our approach can match or outperform task-specific models on all tasks, and in particular achieves new state-of-the-art results on joint entity and relation extraction (CoNLL04, ADE, NYT, and ACE2005 datasets), relation classification (FewRel and TACRED), and semantic role labeling (CoNLL-2005 and CoNLL-2012). We accomplish this while using the same architecture and hyperparameters for all tasks, and even when training a single model to solve all tasks at the same time (multi-task learning). Finally, we show that our framework can also significantly improve the performance in a low-resource regime, thanks to better use of label semantics.
****
Variable-Shot Adaptation for Online Meta-Learning
Tianhe Yu,Xinyang Geng,Chelsea Finn,Sergey Levine
Few-shot meta-learning methods consider the problem of learning new tasks from a small, fixed number of examples, by meta-learning across static data from a set of previous tasks. However, in many real world settings, it is more natural to view the problem as one of minimizing the total amount of supervision --- both the number of examples needed to learn a new task and the amount of data needed for meta-learning. Such a formulation can be studied in a sequential learning setting, where tasks are presented in sequence. When studying meta-learning in this online setting, a critical question arises: can meta-learning improve over the sample complexity and regret of standard empirical risk minimization methods, when considering both meta-training and adaptation together? The answer is particularly non-obvious for meta-learning algorithms with complex bi-level optimizations that may demand large amounts of meta-training data. To answer this question, we extend previous meta-learning algorithms to handle the variable-shot settings that naturally arise in sequential learning: from many-shot learning at the start, to zero-shot learning towards the end. On sequential learning problems, we find that meta-learning solves the full task set with fewer overall labels and achieves greater cumulative performance, compared to standard supervised methods. These results suggest that meta-learning is an important ingredient for building learning systems that continuously learn and improve over a sequence of problems.
****
Faster Binary Embeddings for Preserving Euclidean Distances
Jinjie Zhang,Rayan Saab
We propose a fast, distance-preserving, binary embedding algorithm to transform a high-dimensional dataset $\mathcal{T}\subseteq\mathbb{R}^n$ into binary sequences in the cube $\{\pm 1\}^m$. When $\mathcal{T}$ consists of well-spread (i.e., non-sparse) vectors, our embedding method applies a stable noise-shaping quantization scheme to $A x$ where $A\in\mathbb{R}^{m\times n}$ is a sparse Gaussian random matrix. This contrasts with most binary embedding methods, which usually use $x\mapsto \mathrm{sign}(Ax)$ for the embedding. Moreover, we show that Euclid

ean distances among the elements of $\mathcal{T}$ are approximated by the ℓ_1 norm on the images of $\{\pm 1\}^m$ under a fast linear transformation. This again contrasts with standard methods, where the Hamming distance is used instead. Our method is both fast and memory efficient, with time complexity $O(m)$ and space complexity $O(m)$ on well-spread data. When the data is not well-spread, we show that the approach still works provided that data is transformed via a Walsh-Hadamard matrix, but now the cost is $O(n\log n)$ per data point. Further, we prove that the method is accurate and its associated error is comparable to that of a continuous valued Johnson-Lindenstrauss embedding plus a quantization error that admits a polynomial decay as the embedding dimension m increases. ■Thus the length of the binary codes required to achieve a desired accuracy is quite small, and we show it can even be compressed further without compromising the accuracy. To illustrate our results, we test the proposed method on natural images and show that it achieves strong performance.
****

Scalable Transfer Learning with Expert Models
Joan Puigcerver,Carlos Riquelme Ruiz,Basil Mustafa,Cedric Renggli,André Susano Pinto,Sylvain Gelly,Daniel Keysers,Neil Houlsby
Transfer of pre-trained representations can improve sample efficiency and reduce computational requirements for new tasks. However, representations used for transfer are usually generic, and are not tailored to a particular distribution of downstream tasks. We explore the use of expert representations for transfer with a simple, yet effective, strategy. We train a diverse set of experts by exploiting existing label structures, and use cheap-to-compute performance proxies to select the relevant expert for each target task. This strategy scales the process of transferring to new tasks, since it does not revisit the pre-training data during transfer. Accordingly, it requires little extra compute per target task, and results in a speed-up of 2-3 orders of magnitude compared to competing approaches. Further, we provide an adapter-based architecture able to compress many experts into a single model. We evaluate our approach on two different data sources and demonstrate that it outperforms baselines on over 20 diverse vision tasks in both cases.
****

A Closer Look at Codistillation for Distributed Training
Shagun Sodhani,Olivier Delalleau,Mido Assran,Koustuv Sinha,Nicolas Ballas,Michael Rabbat
Codistillation has been proposed as a mechanism to share knowledge among concurrently trained models by encouraging them to represent the same function through an auxiliary loss. This contrasts with the more commonly used fully-synchronous data-parallel stochastic gradient descent methods, where different model replicas average their gradients (or parameters) at every iteration and thus maintain identical parameters. We investigate codistillation in a distributed training setup, complementing previous work which focused on extremely large batch sizes. Surprisingly, we find that even at moderate batch sizes, models trained with codistillation can perform as well as models trained with synchronous data-parallel methods, despite using a much weaker synchronization mechanism. These findings hold across a range of batch sizes and learning rate schedules, as well as different kinds of models and datasets. Obtaining this level of accuracy, however, requires properly accounting for the regularization effect of codistillation, which we highlight through several empirical observations. Overall, this work contributes to a better understanding of codistillation and how to best take advantage of it in a distributed computing environment.
****

A Primal Approach to Constrained Policy Optimization: Global Optimality and Finite-Time Analysis
Tengyu Xu,Yingbin Liang,Guanghui Lan
Safe reinforcement learning (SRL) problems are typically modeled as constrained Markov Decision Process (CMDP), in which an agent explores the environment to maximize the expected total reward and meanwhile avoids violating certain constraints on a number of expected total costs. In general, such SRL problems have nonc

onvex objective functions subject to multiple nonconvex constraints, and hence a re very challenging to solve, particularly to provide a globally optimal policy. Many popular SRL algorithms adopt a primal-dual structure which utilizes the up dating of dual variables for satisfying the constraints. In contrast, we propose a primal approach, called constraint-rectified policy optimization (CRPO), whic h updates the policy alternatingly between objective improvement and constraint satisfaction. CRPO provides a primal-type algorithmic framework to solve SRL pro blems, where each policy update can take any variant of policy optimization step . To demonstrate the theoretical performance of CRPO, we adopt natural policy gr adient (NPG) for each policy update step and show that CRPO achieves an $\mathca l{O}(1/\sqrt{T})$ convergence rate to the global optimal policy in the constrain ed policy set and an $\mathcal{O}(1/\sqrt{T})$ error bound on constraint satisfa ction. This is the first finite-time analysis of SRL algorithms with global opti mality guarantee. Our empirical results demonstrate that CRPO can outperform the existing primal-dual baseline algorithms significantly.

****

Signatory: differentiable computations of the signature and logsignature transfo rms, on both CPU and GPU

Patrick Kidger,Terry Lyons

Signatory is a library for calculating and performing functionality related to t he signature and logsignature transforms. The focus is on machine learning, and as such includes features such as CPU parallelism, GPU support, and backpropagat ion. To our knowledge it is the first GPU-capable library for these operations. Signatory implements new features not available in previous libraries, such as e fficient precomputation strategies. Furthermore, several novel algorithmic impro vements are introduced, producing substantial real-world speedups even on the CP U without parallelism. The library operates as a Python wrapper around C++, and is compatible with the PyTorch ecosystem. It may be installed directly via \text tt{pip}. Source code, documentation, examples, benchmarks and tests may be found at \texttt{\url{https://github.com/patrick-kidger/signatory}}. The license is A pache-2.0.

****

Efficient Competitive Self-Play Policy Optimization

Yuanyi Zhong,Yuan Zhou,Jian Peng

Reinforcement learning from self-play has recently reported many successes. Self -play, where the agents compete with themselves, is often used to generate train ing data for iterative policy improvement. In previous work, heuristic rules are designed to choose an opponent for the current learner. Typical rules include c hoosing the latest agent, the best agent, or a random historical agent. However, these rules may be inefficient in practice and sometimes do not guarantee conve rgence even in the simplest matrix games. This paper proposes a new algorithmic framework for competitive self-play reinforcement learning in two-player zero-su m games. We recognize the fact that the Nash equilibrium coincides with the sadd le point of the stochastic payoff function, which motivates us to borrow ideas f rom classical saddle point optimization literature. Our method simultaneously tr ains several agents and intelligently takes each other as opponents based on a s imple adversarial rule derived from a principled perturbation-based saddle optim ization method. We prove theoretically that our algorithm converges to an approx imate equilibrium with high probability in convex-concave games under standard a ssumptions. Beyond the theory, we further show the empirical superiority of our method over baseline methods relying on the aforementioned opponent-selection he uristics in matrix games, grid-world soccer, Gomoku, and simulated robot sumo, w ith neural net policy function approximators.

****

Tight Second-Order Certificates for Randomized Smoothing

Alexander Levine,Aounon Kumar,Tom Goldstein,Soheil Feizi

Randomized smoothing is a popular way of providing robustness guarantees against adversarial attacks: randomly-smoothed functions have a universal Lipschitz-lik e bound, allowing for robustness certificates to be easily computed. In this wor k, we show that there also exists a universal curvature-like bound for Gaussian

random smoothing: given the exact value and gradient of a smoothed function, we compute a lower bound on the distance of a point to its closest adversarial example, called the Second-order Smoothing (SoS) robustness certificate. In addition to proving the correctness of this novel certificate, we show that SoS certificates are realizable and therefore tight. Interestingly, we show that the maximum achievable benefits, in terms of certified robustness, from using the additional information of the gradient norm are relatively small: because our bounds are tight, this is a fundamental negative result. The gain of SoS certificates further diminishes if we consider the estimation error of the gradient norms, for which we have developed an estimator. We therefore additionally develop a variant of Gaussian smoothing, called Gaussian dipole smoothing, which provides similar bounds to randomized smoothing with gradient information, but with much-improved sample efficiency. This allows us to achieve (marginally) improved robustness certificates on high-dimensional datasets such as CIFAR-10 and ImageNet.
****
Distance-Based Regularisation of Deep Networks for Fine-Tuning
Henry Gouk,Timothy Hospedales,massimiliano pontil
We investigate approaches to regularisation during fine-tuning of deep neural networks. First we provide a neural network generalisation bound based on Rademacher complexity that uses the distance the weights have moved from their initial values. This bound has no direct dependence on the number of weights and compares favourably to other bounds when applied to convolutional networks. Our bound is highly relevant for fine-tuning, because providing a network with a good initialisation based on transfer learning means that learning can modify the weights less, and hence achieve tighter generalisation. Inspired by this, we develop a simple yet effective fine-tuning algorithm that constrains the hypothesis class to a small sphere centred on the initial pre-trained weights, thus obtaining provably better generalisation performance than conventional transfer learning. Empirical evaluation shows that our algorithm works well, corroborating our theoretical results. It outperforms both state of the art fine-tuning competitors, and penalty-based alternatives that we show do not directly constrain the radius of the search space.
****
Neural SDEs Made Easy: SDEs are Infinite-Dimensional GANs
Patrick Kidger,James Foster,Xuechen Li,Harald Oberhauser,Terry Lyons
Several authors have introduced \emph{Neural Stochastic Differential Equations} (Neural SDEs), often involving complex theory with various limitations. Here, we aim to introduce a generic, user friendly approach to neural SDEs. Our central contribution is the observation that an SDE is a map from Wiener measure (Brownian motion) to a solution distribution, which may be sampled from, but which does not admit a straightforward notion of probability density -- and that this is just the familiar formulation of a GAN. This produces a continuous-time generative model, arbitrary drift and diffusions are admissible, and in the infinite data limit any SDE may be learnt. After that, we construct a new scheme for sampling \emph{and reconstructing} Brownian motion, with constant average-case time and memory costs, adapted to the access patterns of an SDE solver. Finally, we demonstrate that the adjoint SDE (used for backpropagation) may be constructed via rough path theory, without the previous theoretical complexity of two-sided filtrations.
****
Decoupling Global and Local Representations via Invertible Generative Flows
Xuezhe Ma,Xiang Kong,Shanghang Zhang,Eduard H Hovy
In this work, we propose a new generative model that is capable of automatically decoupling global and local representations of images in an entirely unsupervised setting, by embedding a generative flow in the VAE framework to model the decoder.
Specifically, the proposed model utilizes the variational auto-encoding framework to learn a (low-dimensional) vector of latent variables to capture the global information of an image, which is fed as a conditional input to a flow-based invertible decoder with architecture borrowed from style transfer literature.

Experimental results on standard image benchmarks demonstrate the effectiveness of our model in terms of density estimation, image generation and unsupervised representation learning.
Importantly, this work demonstrates that with only architectural inductive biases, a generative model with a likelihood-based objective is capable of learning decoupled representations, requiring no explicit supervision.
The code for our model is available at \url{https://github.com/XuezheMax/wolf}.
****
Understanding Over-parameterization in Generative Adversarial Networks
Yogesh Balaji,Mohammadmahdi Sajedi,Neha Mukund Kalibhat,Mucong Ding,Dominik Stöger,Mahdi Soltanolkotabi,Soheil Feizi
A broad class of unsupervised deep learning methods such as Generative Adversarial Networks (GANs) involve training of overparameterized models where the number of parameters of the model exceeds a certain threshold. Indeed, most successful GANs used in practice are trained using overparameterized generator and discriminator networks, both in terms of depth and width. A large body of work in supervised learning have shown the importance of model overparameterization in the convergence of the gradient descent (GD) to globally optimal solutions. In contrast, the unsupervised setting and GANs in particular involve non-convex concave mini-max optimization problems that are often trained using Gradient Descent/Ascent (GDA).
The role and benefits of model overparameterization in the convergence of GDA to a global saddle point in non-convex concave problems is far less understood. In this work, we present a comprehensive analysis of the importance of model overparameterization in GANs both theoretically and empirically. We theoretically show that in an overparameterized GAN model with a 1-layer neural network generator and a linear discriminator, GDA converges to a global saddle point of the underlying non-convex concave min-max problem. To the best of our knowledge, this is the first result for global convergence of GDA in such settings. Our theory is based on a more general result that holds for a broader class of nonlinear generators and discriminators that obey certain assumptions (including deeper generators and random feature discriminators). Our theory utilizes and builds upon a novel connection with the convergence analysis of linear time-varying dynamical systems which may have broader implications for understanding the convergence behavior of GDA for non-convex concave problems involving overparameterized models. We also empirically study the role of model overparameterization in GANs using several large-scale experiments on CIFAR-10 and Celeb-A datasets. Our experiments show that overparameterization improves the quality of generated samples across various model architectures and datasets. Remarkably, we observe that overparameterization leads to faster and more stable convergence behavior of GDA across the board.
****
Filtered Inner Product Projection for Crosslingual Embedding Alignment
Vin Sachidananda,Ziyi Yang,Chenguang Zhu
Due to widespread interest in machine translation and transfer learning, there are numerous algorithms for mapping multiple embeddings to a shared representation space. Recently, these algorithms have been studied in the setting of bilingual lexicon induction where one seeks to align the embeddings of a source and a target language such that translated word pairs lie close to one another in a common representation space. In this paper, we propose a method, Filtered Inner Product Projection (FIPP), for mapping embeddings to a common representation space. As semantic shifts are pervasive across languages and domains, FIPP first identifies the common geometric structure in both embeddings and then, only on the common structure, aligns the Gram matrices of these embeddings. FIPP is applicable even when the source and target embeddings are of differing dimensionalities. Additionally, FIPP provides computational benefits in ease of implementation and is faster to compute than current approaches. Following the baselines in Glavas et al. 2019, we evaluate FIPP both in the context of bilingual lexicon induction and downstream language tasks. We show that FIPP outperforms existing methods on the XLING BLI dataset for most language pairs while also providing robust perfo

rmance across downstream tasks.
****
Deep Partition Aggregation: Provable Defenses against General Poisoning Attacks
Alexander Levine,Soheil Feizi
Adversarial poisoning attacks distort training data in order to corrupt the test-time behavior of a classifier. A provable defense provides a certificate for each test sample, which is a lower bound on the magnitude of any adversarial distortion of the training set that can corrupt the test sample's classification. We propose two novel provable defenses against poisoning attacks: (i) Deep Partition Aggregation (DPA), a certified defense against a general poisoning threat model, defined as the insertion or deletion of a bounded number of samples to the training set --- by implication, this threat model also includes arbitrary distortions to a bounded number of images and/or labels; and (ii) Semi-Supervised DPA (SS-DPA), a certified defense against label-flipping poisoning attacks. DPA is an ensemble method where base models are trained on partitions of the training set determined by a hash function. DPA is related to both subset aggregation, a well-studied ensemble method in classical machine learning, as well as to randomized smoothing, a popular provable defense against evasion (inference) attacks. Our defense against label-flipping poison attacks, SS-DPA, uses a semi-supervised learning algorithm as its base classifier model: each base classifier is trained using the entire unlabeled training set in addition to the labels for a partition. SS-DPA significantly outperforms the existing certified defense for label-flipping attacks (Rosenfeld et al., 2020) on both MNIST and CIFAR-10: provably tolerating, for at least half of test images, over 600 label flips (vs. < 200 label flips) on MNIST and over 300 label flips (vs. 175 label flips) on CIFAR-10. Against general poisoning attacks where no prior certified defenses exists, DPA can certify $\geq$ 50% of test images against over 500 poison image insertions on MNIST, and nine insertions on CIFAR-10. These results establish new state-of-the-art provable defenses against general and label-flipping poison attacks. Code is available at https://github.com/alevine0/DPA
****
Growing Efficient Deep Networks by Structured Continuous Sparsification
Xin Yuan,Pedro Henrique Pamplona Savarese,Michael Maire
We develop an approach to growing deep network architectures over the course of training, driven by a principled combination of accuracy and sparsity objectives. Unlike existing pruning or architecture search techniques that operate on full-sized models or supernet architectures, our method can start from a small, simple seed architecture and dynamically grow and prune both layers and filters. By combining a continuous relaxation of discrete network structure optimization with a scheme for sampling sparse subnetworks, we produce compact, pruned networks, while also drastically reducing the computational expense of training. For example, we achieve 49.7% inference FLOPs and 47.4% training FLOPs savings compared to a baseline ResNet-50 on ImageNet, while maintaining 75.2% top-1 validation accuracy --- all without any dedicated fine-tuning stage. Experiments across CIFAR, ImageNet, PASCAL VOC, and Penn Treebank, with convolutional networks for image classification and semantic segmentation, and recurrent networks for language modeling, demonstrate that we both train faster and produce more efficient networks than competing architecture pruning or search methods.
****
Learning representations from temporally smooth data
Shima Rahimi Moghaddam,Fanjun Bu,Christopher Honey
Events in the real world are correlated across nearby points in time, and we must learn from this temporally “smooth” data. However, when neural networks are trained to categorize or reconstruct single items, the common practice is to randomize the order of training items. What are the effects of temporally smooth training data on the efficiency of learning? We first tested the effects of smoothness in training data on incremental learning in feedforward nets and found that smoother data slowed learning. Moreover, sampling so as to minimize temporal smoothness produced more efficient learning than sampling randomly. If smoothness generally impairs incremental learning, then how can networks be modified to bene

fit from smoothness in the training data? We hypothesized that two simple brain-inspired mechanisms -- leaky memory in activation units and memory-gating -- could enable networks to exploit the redundancies in smooth data. Across all levels of data smoothness, these brain-inspired architectures achieved more efficient category learning than feedforward networks. Finally, we investigated how these brain-inspired mechanisms altered the internal representations learned by the networks. We found that networks with multi-scale leaky memory and memory-gating could learn internal representations that “un-mixed” data sources which vary on fast and slow timescales across training samples. Altogether, we identified simple mechanisms enabling neural networks to learn more quickly from temporally smooth data, and to generate internal representations that separate timescales in the training signal.

****

Beyond GNNs: A Sample Efficient Architecture for Graph Problems

Pranjal Awasthi,Abhimanyu Das,Sreenivas Gollapudi

Despite their popularity in learning problems over graph structured data, existing Graph Neural Networks (GNNs) have inherent limitations for fundamental graph problems such as shortest paths, k-connectivity, minimum spanning tree and minimum cuts. In all these instances, it is known that one needs GNNs of high depth, scaling at a polynomial rate with the number of nodes n, to provably encode the solution space. This in turn affects their statistical efficiency thus requiring a significant amount of training data in order to obtain networks with good performance. In this work we propose a new hybrid architecture to overcome this limitation. Our proposed architecture that we call as GNNplus networks involve a combination of multiple parallel low depth GNNs along with simple pooling layers involving low depth fully connected networks. We provably demonstrate that for many graph problems, the solution space can be encoded by GNNplus networks using depth that scales only poly-logarithmically in the number of nodes. This significantly improves the amount of training data needed that we establish via improved generalization bounds. Finally, we empirically demonstrate the effectiveness of our proposed architecture for a variety of graph problems.

****

Neural Nonnegative CP Decomposition for Hierarchical Tensor Analysis

Joshua Vendrow,Jamie Haddock,Deanna Needell

There is a significant demand for topic modeling on large-scale data with complex multi-modal structure in applications such as multi-layer network analysis, temporal document classification, and video data analysis; frequently this multi-modal data has latent hierarchical structure. We propose a new hierarchical nonnegative CANDECOMP/PARAFAC (CP) decomposition (hierarchical NCPD) model and a training method, Neural NCPD, for performing hierarchical topic modeling on multi-modal tensor data. Neural NCPD utilizes a neural network architecture and backpropagation to mitigate error propagation through hierarchical NCPD.

****

Mitigating bias in calibration error estimation

Rebecca Roelofs,Nicholas Cain,Jonathon Shlens,Michael Curtis Mozer

Building reliable machine learning systems requires that we correctly understand their level of confidence. Calibration focuses on measuring the degree of accuracy in a model's confidence and most research in calibration focuses on techniques to improve an empirical estimate of calibration error, $\mathrm{ECE}_\mathrm{BIN}$. Using simulation, we show that $\mathrm{ECE}_\mathrm{BIN}$ can systematically underestimate or overestimate the true calibration error depending on the nature of model miscalibration, the size of the evaluation data set, and the number of bins. Critically, $\mathrm{ECE}_\mathrm{BIN}$ is more strongly biased for perfectly calibrated models. We propose a simple alternative calibration error metric, $\mathrm{ECE}_\mathrm{SWEEP}$, in which the number of bins is chosen to be as large as possible while preserving monotonicity in the calibration function. Evaluating our measure on distributions fit to neural network confidence scores on CIFAR-10, CIFAR-100, and ImageNet, we show that $\mathrm{ECE}_\mathrm{SWEEP}$ produces a less biased estimator of calibration error and therefore should be

used by any researcher wishing to evaluate the calibration of models trained on similar datasets.
****
Coordinated Multi-Agent Exploration Using Shared Goals
Iou-Jen Liu,Unnat Jain,Alex Schwing
Exploration is critical for good results of deep reinforcement learning algorithms and has drawn much attention. However, existing multi-agent deep reinforcement learning algorithms still use mostly noise-based techniques. It was recognized recently that noise-based exploration is suboptimal in multi-agent settings, and exploration methods that consider agents' cooperation have been developed. However, existing methods suffer from a common challenge: agents struggle to identify states that are worth exploring, and don't coordinate their exploration efforts toward those states. To address this shortcoming, in this paper, we proposed coordinated multi-agent exploration (CMAE): agents share a common goal while exploring. The goal is selected by a normalized entropy-based technique from multiple projected state spaces. Then, agents are trained to reach the goal in a coordinated manner. We demonstrated that our approach needs only $1\%-5\%$ of the environment steps to achieve similar or better returns than state-of-the-art baselines on various sparse-reward tasks, including a sparse-reward version of the Starcraft multi-agent challenge (SMAC).
****
Cross-Node Federated Graph Neural Network for Spatio-Temporal Data Modeling
Chuizheng Meng,Sirisha Rambhatla,Yan Liu
Vast amount of data generated from networks of sensors, wearables, and the Internet of Things (IoT) devices underscores the need for advanced modeling techniques that leverage the spatio-temporal structure of decentralized data due to the need for edge computation and licensing (data access) issues. While federated learning (FL) has emerged as a framework for model training without requiring direct data sharing and exchange, effectively modeling the complex spatio-temporal dependencies to improve forecasting capabilities still remains an open problem. On the other hand, state-of-the-art spatio-temporal forecasting models assume unfettered access to the data, neglecting constraints on data sharing. To bridge this gap, we propose a federated spatio-temporal model -- Cross-Node Federated Graph Neural Network (CNFGNN) -- which explicitly encodes the underlying graph structure using graph neural network (GNN)-based architecture under the constraint of cross-node federated learning, which requires that data in a network of nodes is generated locally on each node and remains decentralized. CNFGNN operates by disentangling the temporal dynamics modeling on devices and spatial dynamics on the server, utilizing alternating optimization to reduce the communication cost, facilitating computations on the edge devices. Experiments on the traffic flow forecasting task show that CNFGNN achieves the best forecasting performance in both transductive and inductive learning settings with no extra computation cost on edge devices, while incurring modest communication cost.
****
HalentNet: Multimodal Trajectory Forecasting with Hallucinative Intents
Deyao Zhu,Mohamed Zahran,Li Erran Li,Mohamed Elhoseiny
Motion forecasting is essential for making intelligent decisions in robotic navigation. As a result, the multi-agent behavioral prediction has become a core component of modern human-robot interaction applications such as autonomous driving. Due to various intentions and interactions among agents, agent trajectories can have multiple possible futures. Hence, the motion forecasting model's ability to cover possible modes becomes essential to enable accurate prediction. Towards this goal, we introduce HalentNet to better model the future motion distribution in addition to a traditional trajectory regression learning objective by incorporating generative augmentation losses. We model intents with unsupervised discrete random variables whose training is guided by a collaboration between two key signals: A discriminative loss that encourages intents' diversity and a hallucinative loss that explores intent transitions (i.e., mixed intents) and encourages their smoothness. This regulates the neural network behavior to be more accurately predictive on uncertain scenarios due to the active yet careful exploratio

n of possible future agent behavior. Our model's learned representation leads to better and more semantically meaningful coverage of the trajectory distribution. Our experiments show that our method can improve over the state-of-the-art trajectory forecasting benchmarks, including vehicles and pedestrians, for about 20% on average FDE and 50% on road boundary violation rate when predicting 6 seconds future. We also conducted human experiments to show that our predicted trajectories received 39.6% more votes than the runner-up approach and 32.2% more votes than our variant without hallucinative mixed intent loss. The code will be released soon.

****

Self-supervised Contrastive Zero to Few-shot Learning from Small, Long-tailed Text data

Nils Rethmeier,Isabelle Augenstein

For natural language processing (NLP) ‘text-to-text’ tasks, prevailing approaches heavily rely on pretraining large self-supervised models on massive external datasources. However, this methodology is being critiqued for: exceptional compute and pretraining data requirements; diminishing returns on both large and small datasets; and importantly, favourable evaluation settings that overestimate performance differences. The core belief behind current methodology, coined 'the bitter lesson' by R. Sutton, is that 'compute scale-up beats data and compute-efficient algorithms', neglecting that progress in compute hardware scale-up is based almost entirely on the miniaturisation of resource consumption. We thus approach pretraining from a miniaturisation perspective, such as not to require massive external data sources and models, or learned translations from continuous input embeddings to discrete labels. To minimise overly favourable evaluation, we examine learning on a long-tailed, low-resource, multi-label text classification dataset with noisy, highly sparse labels and many rare concepts. To this end, we propose a novel 'dataset-internal' contrastive autoencoding approach to self-supervised pretraining and demonstrate marked improvements in zero-shot, few-shot and solely supervised learning performance; even under an unfavorable low-resource scenario, and without defaulting to large-scale external datasets for self-supervision. We also find empirical evidence that zero and few-shot learning markedly benefit from adding more 'dataset-internal', self-supervised training signals, which is of practical importance when retrieving or computing on large external sources of such signals is infeasible.

****

Reverse engineering learned optimizers reveals known and novel mechanisms

Niru Maheswaranathan,David Sussillo,Luke Metz,Ruoxi Sun,Jascha Sohl-Dickstein

Learned optimizers are algorithms that can themselves be trained to solve optimization problems. In contrast to baseline optimizers (such as momentum or Adam) that use simple update rules derived from intuitive principles, learned optimizers use flexible, high-dimensional, nonlinear parameterizations. Although this can lead to optimizers with better performance in certain settings, their inner workings remain a mystery. How is it that a learned optimizer is able to outperform a well tuned baseline? Has it learned a sophisticated method for combining existing optimization techniques, or is it implementing completely new behavior? In this work, we address these questions by visualizing and understanding learned optimizers. We study learned optimizers trained from scratch on three disparate tasks, and discovered that they have learned interpretable mechanisms, including: momentum, gradient clipping, schedules, and a new form of learning rate adaptation. Moreover, we show how the dynamics of trained learned optimizers enables these behaviors. Our results elucidate the previously murky understanding of what learned optimizers learn, and establishes tools for interpreting future learned optimizers.

****

Motion Forecasting with Unlikelihood Training

Deyao Zhu,Mohamed Zahran,Li Erran Li,Mohamed Elhoseiny

Motion forecasting is essential for making safe and intelligent decisions in robotic applications such as autonomous driving. State-of-the-art methods formulate it as a sequence-to-sequence prediction problem, which is solved in an encoder-

decoder framework with a maximum likelihood estimation objective. In this paper, we show that the likelihood objective itself results in a model assigning too much probability to trajectories that are unlikely given the contextual information such as maps and states of surrounding agents. This is despite the fact that many state-of-the-art models do take contextual information as part of their input. We propose a new objective, unlikelihood training, which forces generated trajectories that conflicts with contextual information to be assigned a lower probability by our model. We demonstrate that our method can significantly improve state-of-art models' performance on challenging real-world trajectory forecasting datasets (nuScenes and Argoverse) by 8% and reduce the standard deviation by up to 50%. The code will be made available.
****
Adversarial Masking: Towards Understanding Robustness Trade-off for Generalization
Minhao Cheng,Zhe Gan,Yu Cheng,Shuohang Wang,Cho-Jui Hsieh,Jingjing Liu
Adversarial training is a commonly used technique to improve model robustness against adversarial examples. Despite its success as a defense mechanism, adversarial training often fails to generalize well to unperturbed test data. While previous work assumes it is caused by the discrepancy between robust and non-robust features, in this paper, we introduce \emph{Adversarial Masking}, a new hypothesis is that this trade-off is caused by different feature maskings applied. Specifically, the rescaling operation in the batch normalization layer, when combined together with ReLU activation, serves as a feature masking layer to select different features for model training. By carefully manipulating different maskings, a well-balanced trade-off can be achieved between model performance on unperturbed and perturbed data. Built upon this hypothesis, we further propose Robust Masking (RobMask), which constructs unique masking for every specific attack perturbation by learning a set of primary adversarial feature maskings. By incorporating different feature maps after the masking, we can distill better features to help model generalization. Sufficiently, adversarial training can be treated as an effective regularizer to achieve better generalization. Experiments on multiple benchmarks demonstrate that RobMask achieves significant improvement on clean test accuracy compared to strong state-of-the-art baselines.
****
Generating unseen complex scenes: are we there yet?
Arantxa Casanova,Michal Drozdzal,Adriana Romero
Although recent complex scene conditional generation models generate increasingly appealing scenes, it is very hard to assess which models perform better and why. This is often due to models being trained to fit different data splits, and defining their own experimental setups. In this paper, we propose a methodology to compare complex scene conditional generation models, and provide an in-depth analysis that assesses the ability of each model to (1) fit the training distribution and hence perform well on seen conditionings, (2) to generalize to unseen conditionings composed of seen object combinations, and (3) generalize to unseen conditionings composed of unseen object combinations. As a result, we observe that recent methods are able to generate recognizable scenes given seen conditionings, and exploit compositionality to generalize to unseen conditionings with seen object combinations. However, all methods suffer from noticeable image quality degradation when asked to generate images from conditionings composed of unseen object combinations. Moreover, through our analysis, we identify the advantages of different pipeline components, and find that (1) encouraging compositionality through instance-wise spatial conditioning normalizations increases robustness to both types of unseen conditionings, (2) using semantically aware losses such as the scene-graph perceptual similarity helps improve some dimensions of the generation process, and (3) enhancing the quality of generated masks and the quality of the individual objects are crucial steps to improve robustness to both types of unseen conditionings.
****
INT: An Inequality Benchmark for Evaluating Generalization in Theorem Proving
Yuhuai Wu,Albert Jiang,Jimmy Ba,Roger Baker Grosse

In learning-assisted theorem proving, one of the most critical challenges is to generalize to theorems unlike those seen at training time. In this paper, we introduce INT, an INequality Theorem proving benchmark designed to test agents' generalization ability. INT is based on a theorem generator, which provides theoretically infinite data and allows us to measure 6 different types of generalization, each reflecting a distinct challenge, characteristic of automated theorem proving. In addition, provides a fast theorem proving environment with sequence-based and graph-based interfaces, conducive to performing learning-based research. We introduce base-lines with architectures including transformers and graph neural networks (GNNs)for INT. Using INT, we find that transformer-based agents achieve stronger test performance for most of the generalization tasks, despite having much larger out-of-distribution generalization gaps than GNNs. We further find that the addition of Monte Carlo Tree Search (MCTS) at test time helps to prove new theorems.

****

Bayesian Few-Shot Classification with One-vs-Each Pólya-Gamma Augmented Gaussian Processes

Jake Snell,Richard Zemel

Few-shot classification (FSC), the task of adapting a classifier to unseen classes given a small labeled dataset, is an important step on the path toward human-like machine learning. Bayesian methods are well-suited to tackling the fundamental issue of overfitting in the few-shot scenario because they allow practitioners to specify prior beliefs and update those beliefs in light of observed data. Contemporary approaches to Bayesian few-shot classification maintain a posterior distribution over model parameters, which is slow and requires storage that scales with model size. Instead, we propose a Gaussian process classifier based on a novel combination of Pólya-Gamma augmentation and the one-vs-each softmax approximation that allows us to efficiently marginalize over functions rather than model parameters. We demonstrate improved accuracy and uncertainty quantification on both standard few-shot classification benchmarks and few-shot domain transfer tasks.

****

Learn what you can't learn: Regularized Ensembles for Transductive out-of-distribution detection

Alexandru ■ifrea,Eric Petru Stavarache,Fanny Yang

Machine learning models are often used in practice once they achieve good generalization results on in-distribution (ID) holdout data. To predict test sets in the wild, they should detect samples they cannot predict well. We show that current out-of-distribution (OOD) detection algorithms for neural networks produce unsatisfactory results in a variety of OOD detection scenarios, e.g.\ when OOD data consists of unseen classes or corrupted measurements. This paper studies how such ``hard'' OOD scenarios can benefit from tuning the detection method after observing a batch of the test data. This \emph{transductive} setting is relevant when the advantage of even a slightly delayed OOD detection outweighs the financial cost for additional tuning. We propose a novel method that uses an artificial labeling scheme for the test data and early stopping regularization to obtain ensembles of models that produce contradictory predictions only on the OOD samples in a test batch. We show via comprehensive experiments that our approach is indeed able to significantly outperform both inductive and transductive baselines on difficult OOD detection scenarios, such as unseen classes on CIFAR-10/CIFAR-100, severe corruptions (CIFAR-C), and strong covariate shift ImageNet vs ObjectNet.

****

World Model as a Graph: Learning Latent Landmarks for Planning

Lunjun Zhang,Ge Yang,Bradly C Stadie

Planning, the ability to analyze the structure of a problem in the large and decompose it into interrelated subproblems, is a hallmark of human intelligence. While deep reinforcement learning (RL) has shown great promise for solving relatively straightforward control tasks, it remains an open problem how to best incorporate planning into existing deep RL paradigms to handle increasingly complex en

vironments. One prominent framework, Model-Based RL, learns a world model and pl ans using step-by-step virtual rollouts. This type of world model quickly diverg es from reality when the planning horizon increases, thus struggling at long-hor izon planning. How can we learn world models that endow agents with the ability to do temporally extended reasoning? In this work, we propose to learn graph-str uctured world models composed of sparse, multi-step transitions. We devise a nov el algorithm to learn latent landmarks that are scattered (in terms of reachabil ity) across the goal space as the nodes on the graph. In this same graph, the ed ges are the reachability estimates distilled from Q-functions. On a variety of h igh-dimensional continuous control tasks ranging from robotic manipulation to na vigation, we demonstrate that our method, named L^{3}P, significantly outperform s prior work, and is oftentimes the only method capable of leveraging both the r obustness of model-free RL and generalization of graph-search algorithms. We bel ieve our work is an important step towards scalable planning in reinforcement le arning.
****
Out-of-distribution Prediction with Invariant Risk Minimization: The Limitation and An Effective Fix
Ruocheng Guo,Pengchuan Zhang,Hao Liu,Emre Kiciman
This work considers the out-of-distribution (OOD) prediction problem where (1)~t he training data are from multiple domains and (2)~the test domain is unseen in the training. DNNs fail in OOD prediction because they are prone to pick up spur ious correlations. Recently, Invariant Risk Minimization (IRM) is proposed to ad dress this issue. Its effectiveness has been demonstrated in the colored MNIST e xperiment. Nevertheless, we find that the performance of IRM can be dramatically degraded under \emph{strong Λ spuriousness} -- when the spurious correl ation between the spurious features and the class label is strong due to the str ong causal influence of their common cause, the domain label, on both of them (s ee Fig. 1). In this work, we try to answer the questions: why does IRM fail in t he aforementioned setting? Why does IRM work for the original colored MNIST data set? Then, we propose a simple and effective approach to fix the problem of IRM. We combine IRM with conditional distribution matching to avoid a specific type of spurious correlation under strong Λ spuriousness. Empirically, we des ign a series of semi synthetic datasets -- the colored MNIST plus, which exposes the problems of IRM and demonstrates the efficacy of the proposed method.
****
Learning with Plasticity Rules: Generalization and Robustness
Rares C Cristian,Max Dabagia,Christos Papadimitriou,Santosh Vempala
Brains learn robustly, and generalize effortlessly between different learning t asks; in contrast, robustness and generalization across tasks are well known wea knesses of artificial neural nets (ANNs). How can we use our accelerating under standing of the brain to improve these and other aspects of ANNs? Here we hypoth esize that (a) Brains employ synaptic plasticity rules that serve as proxies for GD; (b) These rules themselves can be learned by GD on the rule parameters; and (c) This process may be a missing ingredient for the development of ANNs that g eneralize well and are robust to adversarial perturbations. We provide both empi rical and theoretical evidence for this hypothesis. In our experiments, plastici ty rules for the synaptic weights of recurrent neural nets (RNNs) are learned th rough GD and are found to perform reasonably well (with no backpropagation). We find that plasticity rules learned by this process generalize from one type of data/classifier to others (e.g., rules learned on synthetic data work well on MN IST/Fashion MNIST) and converge with fewer updates. Moreover, the classifiers le arned using plasticity rules exhibit surprising levels of tolerance to adversari al perturbations. In the special case of the last layer of a classification netw ork, we show analytically that GD on the plasticity rule recovers (and improves upon) the perceptron algorithm and the multiplicative weights method. Finally, w e argue that applying GD to learning rules is biologically plausible, in the sen se that it can be learned over evolutionary time: we describe a genetic setting where natural selection of a numerical parameter over a sequence of generations provably simulates a simple variant of GD.

****
Later Span Adaptation for Language Understanding
Rongzhou Bao,Zhuosheng Zhang,hai zhao
Pre-trained contextualized language models (PrLMs) broadly use fine-grained tokens (words or sub-words) as minimal linguistic unit in pre-training phase. Introducing span-level information in pre-training has shown capable of further enhancing PrLMs. However, such methods require enormous resources and are lack of adaptivity due to huge computational requirement from pre-training. Instead of too early fixing the linguistic unit input as nearly all previous work did, we propose a novel method that combines span-level information into the representations generated by PrLMs during fine-tuning phase for better flexibility. In this way, the modeling procedure of span-level texts can be more adaptive to different downstream tasks. In detail, we divide the sentence into several spans according to the segmentation generated by a pre-sampled dictionary. Based on the sub-token-level representation provided by PrLMs, we enhance the connection between the tokens in each span and gain a representation with enhanced span-level information. Experiments are conducted on GLUE benchmark and prove that our approach could remarkably enhance the performance of PrLMs in various natural language understanding tasks.
****
A Hypergradient Approach to Robust Regression without Correspondence
Yujia Xie,Yixiu Mao,Simiao Zuo,Hongteng Xu,Xiaojing Ye,Tuo Zhao,Hongyuan Zha
We consider a regression problem, where the correspondence between the input and output data is not available. Such shuffled data are commonly observed in many real world problems. Take flow cytometry as an example: the measuring instruments are unable to preserve the correspondence between the samples and the measurements. Due to the combinatorial nature of the problem, most of the existing methods are only applicable when the sample size is small, and are limited to linear regression models. To overcome such bottlenecks, we propose a new computational framework --- ROBOT --- for the shuffled regression problem, which is applicable to large data and complex models. Specifically, we propose to formulate regression without correspondence as a continuous optimization problem. Then by exploiting the interaction between the regression model and the data correspondence, we propose to develop a hypergradient approach based on differentiable programming techniques. Such a hypergradient approach essentially views the data correspondence as an operator of the regression model, and therefore it allows us to find a better descent direction for the model parameters by differentiating through the data correspondence. ROBOT is quite general, and can be further extended to an inexact correspondence setting, where the input and output data are not necessarily exactly aligned. Thorough numerical experiments show that ROBOT achieves better performance than existing methods in both linear and nonlinear regression tasks, including real-world applications such as flow cytometry and multi-object tracking.
****
Neural Dynamical Systems: Balancing Structure and Flexibility in Physical Prediction
Viraj Mehta,Ian Char,Willie Neiswanger,Youngseog Chung,Andrew Oakleigh Nelson,Mark D Boyer,Egemen Kolemen,Jeff Schneider
We introduce Neural Dynamical Systems (NDS), a method of learning dynamical models in various gray-box settings which incorporates prior knowledge in the form of systems of ordinary differential equations. NDS uses neural networks to estimate free parameters of the system, predicts residual terms, and numerically integrates over time to predict future states. A key insight is that many real dynamical systems of interest are hard to model because the dynamics may vary across rollouts. We mitigate this problem by taking a trajectory of prior states as the input to NDS and train it to dynamically estimate system parameters using the preceding trajectory. We find that NDS learns dynamics with higher accuracy and fewer samples than a variety of deep learning methods that do not incorporate the prior knowledge and methods from the system identification literature which do. We demonstrate these advantages first on synthetic dynamical systems and th

en on real data captured from deuterium shots from a nuclear fusion reactor. Fin ally, we demonstrate that these benefits can be utilized for control in small-sc ale experiments.
****

Parametric Density Estimation with Uncertainty using Deep Ensembles
Abel Peirson,Taylor Howell,Marius Aurel Tirlea
In parametric density estimation, the parameters of a known probability density are typically recovered from measurements by maximizing the log-likelihood. Prio r knowledge of measurement uncertainties is not included in this method -- poten tially producing degraded or even biased parameter estimates.
We propose an efficient two-step, general-purpose approach for parametric densit y estimation using deep ensembles.
Feature predictions and their uncertainties are returned by a deep ensemble and then combined in an importance weighted maximum likelihood estimation to recover parameters representing a known density along with their respective errors. To compare the bias-variance tradeoff of different approaches, we define an appropr iate figure of merit.
We illustrate a number of use cases for our method in the physical sciences and demonstrate state-of-the-art results for X-ray polarimetry that outperform curre nt classical and deep learning methods.
****

Learning to Infer Run-Time Invariants from Source code
Vincent Josua Hellendoorn,Premkumar Devanbu,Alex Polozov,Mark Marron
Source code is notably different from natural language in that it is meant to be executed. Experienced developers infer complex "invariants" about run-time stat e while reading code, which helps them to constrain and predict program behavior . Knowing these invariants can be helpful; yet developers rarely encode these ex plicitly, so machine-learning methods don't have much aligned data to learn from . We propose an approach that adapts cues within existing if-statements regardi ng explicit run-time expectations to generate aligned datasets of code and impli cit invariants. We also propose a contrastive loss to inhibit generation of illo gical invariants. Our model learns to infer a wide vocabulary of invariants for arbitrary code, which can be used to detect and repair real bugs. This is entire ly complementary to established approaches, which either use logical engines tha t scale poorly, or run-time traces that are expensive to obtain; when present, t hat data can complement our tool, as we demonstrate in conjunction with Daikon, an existing tool. Our results show that neural models can derive useful represen tations of run-time behavior directly from source code.
****

C-Learning: Horizon-Aware Cumulative Accessibility Estimation
Panteha Naderian,Gabriel Loaiza-Ganem,Harry J. Braviner,Anthony L. Caterini,Jess e C. Cresswell,Tong Li,Animesh Garg
Multi-goal reaching is an important problem in reinforcement learning needed to achieve algorithmic generalization. Despite recent advances in this field, curre nt algorithms suffer from three major challenges: high sample complexity, learni ng only a single way of reaching the goals, and difficulties in solving complex motion planning tasks. In order to address these limitations, we introduce the concept of cumulative accessibility functions, which measure the reachability of a goal from a given state within a specified horizon. We show that these functi ons obey a recurrence relation, which enables learning from offline interactions . We also prove that optimal cumulative accessibility functions are monotonic in the planning horizon. Additionally, our method can trade off speed and reliabil ity in goal-reaching by suggesting multiple paths to a single goal depending on the provided horizon. We evaluate our approach on a set of multi-goal discrete a nd continuous control tasks. We show that our method outperforms state-of-the-ar t goal-reaching algorithms in success rate, sample complexity, and path optimali ty. Our code is available at https://github.com/layer6ai-labs/CAE, and additiona l visualizations can be found at https://sites.google.com/view/learning-cae/.
****

Minimum Width for Universal Approximation

Sejun Park,Chulhee Yun,Jaeho Lee,Jinwoo Shin
The universal approximation property of width-bounded networks has been studied as a dual of classical universal approximation results on depth-bounded networks . However, the critical width enabling the universal approximation has not been exactly characterized in terms of the input dimension d_x and the output dimension d_y. In this work, we provide the first definitive result in this direction for networks using the ReLU activation functions: The minimum width required for the universal approximation of the L^p functions is exactly $\max\{d_x+1,d_y\}$. We also prove that the same conclusion does not hold for the uniform approximation with ReLU, but does hold with an additional threshold activation function. Our proof technique can be also used to derive a tighter upper bound on the minimum width required for the universal approximation using networks with general activation functions.
****

MIROSTAT: A NEURAL TEXT DECODING ALGORITHM THAT DIRECTLY CONTROLS PERPLEXITY
Sourya Basu,Govardana Sachitanandam Ramachandran,Nitish Shirish Keskar,Lav R. Varshney
Neural text decoding algorithms strongly influence the quality of texts generated using language models, but popular algorithms like top-k, top-p (nucleus), and temperature-based sampling may yield texts that have objectionable repetition or incoherence. Although these methods generate high-quality text after ad hoc parameter tuning that depends on the language model and the length of generated text, not much is known about the control they provide over the statistics of the output. This is important, however, since recent reports show that humans prefer when perplexity is neither too much nor too little and since we experimentally show that cross-entropy (log of perplexity) has a near-linear relation with repetition. First, we provide a theoretical analysis of perplexity in top-k, top-p, and temperature sampling, under Zipfian statistics. Then, we use this analysis to design a feedback-based adaptive top-k text decoding algorithm called mirostat that generates text (of any length) with a predetermined target value of perplexity without any tuning. Experiments show that for low values of k and p, perplexity drops significantly with generated text length and leads to excessive repetitions (the boredom trap). Contrarily, for large values of k and p, perplexity increases with generated text length and leads to incoherence (confusion trap). Mirostat avoids both traps. Specifically, we show that setting target perplexity value beyond a threshold yields negligible sentence-level repetitions. Experiments with
human raters for fluency, coherence, and quality further verify our findings.
****

Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime
Andrea Agazzi,Jianfeng Lu
We study the problem of policy optimization for infinite-horizon discounted Markov Decision Processes with softmax policy and nonlinear function approximation trained with policy gradient algorithms. We concentrate on the training dynamics in the mean-field regime, modeling e.g. the behavior of wide single hidden layer neural networks, when exploration is encouraged through entropy regularization. The dynamics of these models is established as a Wasserstein gradient flow of distributions in parameter space. We further prove global optimality of the fixed points of this dynamics under mild conditions on their initialization.
****

Natural Compression for Distributed Deep Learning
Samuel Horváth,Chen-Yu Ho,Ludovit Horváth,Atal Narayan Sahu,Marco Canini,Peter Richtarik
Modern deep learning models are often trained in parallel over a collection of distributed machines to reduce training time. In such settings, communication of model updates among machines becomes a significant performance bottleneck and various lossy update compression techniques have been proposed to alleviate this problem. In this work, we introduce a new, simple yet theoretically and practically effective compression technique: {\em natural compression (C_{nat})}. Our t

echnique is applied individually to all entries of the to-be-compressed update v
ector and works by randomized rounding to the nearest (negative or positive) pow
er of two, which can be computed in a ``natural'' way by ignoring the mantissa.
We show that compared to no compression, C_{nat} increases the second moment
of the compressed vector by not more than the tiny factor $\nicefrac{9}{8}$, wh
ich means that the effect of C_{nat} on the convergence speed of popular train
ing algorithms, such as distributed SGD, is negligible. However, the communicat
ions savings enabled by C_{nat} are substantial,
leading to {\em 3-$4\times$ improvement in overall theoretical running time}.
For applications requiring more aggressive compression, we generalize C_{nat}
to {\em natural dithering}, which we prove is {\em exponentially better} than th
e common random dithering technique. Our compression operators can be used on th
eir own or in combination with existing operators for a more aggressive combined
 effect, and offer new state-of-the-art both in theory and practice.
****

The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers
Preetum Nakkiran,Behnam Neyshabur,Hanie Sedghi
We propose a new framework for reasoning about generalization in deep learning.
The core idea is to couple the Real World, where optimizers take stochastic grad
ient steps on the empirical loss, to an Ideal World, where optimizers take steps
 on the population loss. This leads to an alternate decomposition of test error
into: (1) the Ideal World test error plus (2) the gap between the two worlds. If
 the gap (2) is universally small, this reduces the problem of generalization in
 offline learning to the problem of optimization in online learning.
We then give empirical evidence that this gap between worlds can be small in rea
listic deep learning settings, in particular supervised image classification. Fo
r example, CNNs generalize better than MLPs on image distributions in the Real W
orld, but this is "because" they optimize faster on the population loss in the I
deal World. This suggests our framework is a useful tool for understanding gener
alization in deep learning, and lays the foundation for future research in this
direction.
****

AdaLead: A simple and robust adaptive greedy search algorithm for sequence desig
n
Sam Sinai,Richard Wang,Alexander Whatley,Stewart Slocum,Elina Locane,Eric Kelsic
Efficient design of biological sequences will have a great impact across many in
dustrial and healthcare domains. However, discovering improved sequences require
s solving a difficult optimization problem. Traditionally, this challenge was ap
proached by biologists through a model-free method known as “directed evolution”
, the iterative process of random mutation and selection. As the ability to buil
d models that capture the sequence-to-function map improves, such models can be
used as oracles to screen sequences before running experiments. In recent years,
 interest in better algorithms that effectively use such oracles to outperform m
odel-free approaches has intensified. These span from approaches based on Bayesi
an Optimization, to regularized generative models and adaptations of reinforceme
nt learning. In this work, we implement an open-source Fitness Landscape EXplora
tion Sandbox (FLEXS) environment to test and evaluate these algorithms based on
their optimality, consistency, and robustness. Using FLEXS, we develop an easy-t
o-implement, scalable, and robust evolutionary greedy algorithm (AdaLead). Despi
te its simplicity, we show that AdaLead is a remarkably strong benchmark that ou
t-competes more complex state of the art approaches in a variety of biologically
 motivated sequence design challenges.
****

Grounding Language to Entities for Generalization in Reinforcement Learning
H. J. Austin Wang,Karthik R Narasimhan
In this paper, we consider the problem of leveraging textual descriptions to imp
rove generalization of control policies to new scenarios. Unlike prior work in t
his space, we do not assume access to any form of prior knowledge connecting tex
t and state observations, and learn both symbol grounding and control policy sim
ultaneously. This is challenging due to a lack of concrete supervision, and inco

rrect groundings can result in worse performance than policies that do not use t
he text at all.
We develop a new model, EMMA (Entity Mapper with Multi-modal Attention) which us
es a multi-modal entity-conditioned attention module that allows for selective f
ocus over relevant sentences in the manual for each entity in the environment. E
MMA is end-to-end differentiable and can learn a latent grounding of entities an
d dynamics from text to observations using environment rewards as the only sourc
e of supervision.
To empirically test our model, we design a new framework of 1320 games and colle
ct text manuals with free-form natural language via crowd-sourcing. We demonstra
te that EMMA achieves successful zero-shot generalization to unseen games with n
ew dynamics, obtaining significantly higher rewards compared to multiple baselin
es. The grounding acquired by EMMA is also robust to noisy descriptions and ling
uistic variation.
****

Efficient Differentiable Neural Architecture Search with Model Parallelism
Yi-Wei Chen,Qingquan Song,Xia Hu
Neural architecture search (NAS) automatically designs effective network archite
ctures. Differentiable NAS with supernets that encompass all potential architect
ures in a large graph cuts down search overhead to few GPU days or less. However
, these algorithms consume massive GPU memory, which will restrain NAS from larg
e batch sizes and large search spaces (e.g., more candidate operations, diverse
cell structures, and large depth of supernets). In this paper, we present binary
 neural architecture search (NASB) with consecutive model parallel (CMP) to tack
le the problem of insufficient GPU memory. CMP aggregates memory from multiple G
PUs for supernets. It divides forward/backward phases into several sub-tasks and
 executes the same type of sub-tasks together to reduce waiting cycles. This app
roach improves the hardware utilization of model parallel, but it utilizes large
 GPU memory. NASB is proposed to reduce memory footprint, which excludes inactiv
e operations from computation graphs and computes those operations on the fly fo
r inactive architectural gradients in backward phases. Experiments show that NAS
B-CMP runs 1.2× faster than other model parallel approaches and outperforms stat
e-of-the-art differentiable NAS. NASB can also save twice GPU memory more than P
C-DARTS. Finally, we apply NASB-CMP to complicated supernet architectures. Altho
ugh deep supernets with diverse cell structures do not improve NAS performance,
NASB-CMP shows its potential to explore supernet architecture design in large se
arch space.
****

On the Certified Robustness for Ensemble Models and Beyond
Zhuolin Yang,Linyi Li,Xiaojun Xu,Bhavya Kailkhura,Bo Li
Recent studies show that deep neural networks (DNN) are vulnerable to adversaria
l examples, which aim to mislead DNNs to make arbitrarily incorrect predictions.
 To defend against such attacks, both empirical and theoretical defense approach
es have been proposed for a single ML model. In this work, we aim to explore and
 characterize the robustness conditions for ensemble ML models. We prove that th
e diversified gradient and large confidence margin are sufficient and necessary
conditions for certifiably robust ensemble models under the model-smoothness ass
umption. We also show that an ensemble model can achieve higher certified robust
ness than a single base model based on these conditions. To our best knowledge,
this is the first work providing tight conditions for the ensemble robustness. I
nspired by our analysis, we propose the lightweight Diversity Regularized Traini
ng (DRT) for ensemble models. We derive the certified robustness of DRT based en
sembles such as standard Weighted Ensemble and Max-Margin Ensemble following the
 sufficient and necessary conditions. Besides, to efficiently calculate the mode
l-smoothness, we leverage adapted randomized model smoothing to obtain the certi
fied robustness for different ensembles in practice. We show that the certified
robustness of ensembles, on the other hand, verifies the necessity of DRT. To co
mpare different ensembles, we prove that when the adversarial transferability am
ong base models is high, Max-Margin Ensemble can achieve higher certified robust
ness than Weighted Ensemble; vice versa. Extensive experiments show that ensembl

e models trained with DRT can achieve the state-of-the-art certified robustness under various settings. Our work will shed light on future analysis for robust e nsemble models.
****
Analyzing Attention Mechanisms through Lens of Sample Complexity and Loss Landsc ape
Bingyuan Liu,Yogesh Balaji,Lingzhou Xue,Martin Renqiang Min
Attention mechanisms have advanced state-of-the-art deep learning models in many machine learning tasks. Despite significant empirical gains, there is a lack of theoretical analyses on their effectiveness. In this paper, we address this pro blem by studying the sample complexity and loss landscape of attention-based neu ral networks. Our results show that, under mild assumptions, every local minimum of the attention model has low prediction error, and attention models require l ower sample complexity than models without attention. Besides revealing why popu lar self-attention works, our theoretical results also provide guidelines for de signing future attention models. Experiments on various datasets validate our th eoretical findings.
****
Investigating and Simplifying Masking-based Saliency Methods for Model Interpret ability
Jason Phang,Jungkyu Park,Krzysztof J. Geras
Saliency maps that identify the most informative regions of an image for a class ifier are valuable for model interpretability. A common approach to creating sal iency maps involves generating input masks that mask out portions of an image to maximally deteriorate classification performance, or mask in an image to preser ve classification performance. Many variants of this approach have been proposed in the literature, such as counterfactual generation and optimizing over a Gumb el-Softmax distribution. Using a general formulation of masking-based saliency m ethods, we conduct an extensive evaluation study of a number of recently propose d variants to understand which elements of these methods meaningfully improve pe rformance. Surprisingly, we find that a well-tuned, relatively simple formulatio n of a masking-based saliency model outperforms many more complex approaches. We find that the most important ingredients for high quality saliency map generati on are (1) using both masked-in and masked-out objectives and (2) training the c lassifier alongside the masking model. Strikingly, we show that a masking model can be trained with as few as 10 examples per class and still generate saliency maps with only a 0.7-point increase in localization error.
****
A Better Alternative to Error Feedback for Communication-Efficient Distributed L earning
Samuel Horváth,Peter Richtarik
Modern large-scale machine learning applications require stochastic optimization algorithms to be implemented on distributed computing systems. A key bottleneck of such systems is the communication overhead for exchanging information across the workers, such as stochastic gradients. Among the many techniques proposed t o remedy this issue, one of the most successful is the framework of compressed c ommunication with error feedback (EF). EF remains the only known technique that can deal with the error induced by contractive compressors which are not unbiase d, such as Top-K or PowerSGD. In this paper, we propose a new and theoretical ly and practically better alternative to EF for dealing with contractive compres sors. In particular, we propose a construction which can transform any contracti ve compressor into an induced unbiased compressor. Following this transformation , existing methods able to work with unbiased compressors can be applied. We sho w that our approach leads to vast improvements over EF, including reduced memory requirements, better communication complexity guarantees and fewer assumptions. We further extend our results to federated learning with partial participation following an arbitrary distribution over the nodes and demonstrate the benefits thereof. We perform several numerical experiments which validate our theoretical findings.
****

Boosting One-Point Derivative-Free Online Optimization via Residual Feedback
Yan Zhang,Yi Zhou,Kaiyi Ji,Michael Zavlanos
Zeroth-order optimization (ZO) typically relies on two-point feedback to estimat e the unknown gradient of the objective function, which queries the objective fu nction value twice at each time instant. However, if the objective function is t ime-varying, as in online optimization, two-point feedback can not be used. In t his case, the gradient can be estimated using one-point feedback that queries a single function value at each time instant, although at the expense of producing gradient estimates with large variance. In this work, we propose a new one-poin t feedback method for online optimization that estimates the objective function gradient using the residual between two feedback points at consecutive time inst ants. We study the regret bound of ZO with residual feedback for both convex and nonconvex online optimization problems. Specifically, for both Lipschitz and sm ooth functions, we show that using residual feedback produces gradient estimates with much smaller variance compared to conventional one-point feedback methods, which improves the learning rate. Our regret bound for ZO with residual feedbac k is tighter than the existing regret bound for ZO with conventional one-point f eedback and relies on weaker assumptions, which suggests that ZO with our propos ed residual feedback can better track the optimizer of online optimization probl ems. We provide numerical experiments that demonstrate that ZO with residual fee dback significantly outperforms existing one-point feedback methods in practice.
****
StructFormer: Joint Unsupervised Induction of Dependency and Constituency Struct ure from Masked Language Modeling
Yikang Shen,Yi Tay,Che Zheng,Dara Bahri,Donald Metzler,Aaron Courville
There are two major classes of natural language grammars --- the dependency gram mar that models one-to-one correspondences between words and the constituency gr ammar that models the assembly of one or several corresponded words. While previ ous unsupervised parsing methods mostly focus on only inducing one class of gram mars, we introduce a novel model, StructFormer, that can induce dependency and c onstituency structure at the same time. To achieve this, we propose a new parsin g framework that can jointly generates constituency tree and dependency graph. T hen we integrate the induced dependency relations into transformer, in a differe ntiable manner, through a novel dependency-constrained self-attention mechanism. Experimental results show that our model can achieve strong results on unsuperv ised constituency parsing, unsupervised dependency parsing and masked language m odeling at the same time.
****
DrNAS: Dirichlet Neural Architecture Search
Xiangning Chen,Ruochen Wang,Minhao Cheng,Xiaocheng Tang,Cho-Jui Hsieh
This paper proposes a novel differentiable architecture search method by formula ting it into a distribution learning problem. We treat the continuously relaxed architecture mixing weight as random variables, modeled by Dirichlet distributio n. With recently developed pathwise derivatives, the Dirichlet parameters can be easily optimized with gradient-based optimizer in an end-to-end manner. This fo rmulation improves the generalization ability and induces stochasticity that nat urally encourages exploration in the search space. Furthermore, to alleviate the large memory consumption of differentiable NAS, we propose a simple yet effecti ve progressive learning scheme that enables searching directly on large-scale ta sks, eliminating the gap between search and evaluation phases. Extensive experim ents demonstrate the effectiveness of our method. Specifically, we obtain a test error of 2.46\% for CIFAR-10, 23.7\% for ImageNet under the mobile setting. On NAS-Bench-201, we also achieve state-of-the-art results on all three datasets an d provide insights for the effective design of neural architecture search algori thms.
****
Improving Few-Shot Visual Classification with Unlabelled Examples
Peyman Bateni,Jarred Barber,Jan-Willem van de Meent,Frank Wood
We propose a transductive meta-learning method that uses unlabelled instances to improve few-shot image classification performance. Our approach combines a regu

larized Mahalanobis-distance-based soft k-means clustering procedure with a modi
fied state of the art neural adaptive feature extractor to achieve improved test
-time classification accuracy using unlabelled data. We evaluate our method on t
ransductive few-shot learning tasks, in which the goal is to jointly predict lab
els for query (test) examples given a set of support (training) examples. We ach
ieve new state of the art performance on the Meta-Dataset and the mini-ImageNet
and tiered-ImageNet benchmarks.
****

Offline Model-Based Optimization via Normalized Maximum Likelihood Estimation
Justin Fu,Sergey Levine
In this work we consider data-driven optimization problems where one must maximi
ze a function given only queries at a fixed set of points. This problem setting
emerges in many domains where function evaluation is a complex and expensive pro
cess, such as in the design of materials, vehicles, or neural network architectu
res. Because the available data typically only covers a small manifold of the po
ssible space of inputs, a principal challenge is to be able to construct algorit
hms that can reason about uncertainty and out-of-distribution values, since a na
ive optimizer can easily exploit an estimated model to return adversarial inputs
. We propose to tackle the MBO problem by leveraging the normalized maximum-like
lihood (NML) estimator, which provides a principled approach to handling uncerta
inty and out-of-distribution inputs. While in the standard formulation NML is in
tractable, we propose a tractable approximation that allows us to scale our meth
od to high-capacity neural network models. We demonstrate that our method can ef
fectively optimize high-dimensional design problems in a variety of disciplines
such as chemistry, biology, and materials engineering.
****

Viewmaker Networks: Learning Views for Unsupervised Representation Learning
Alex Tamkin,Mike Wu,Noah Goodman
Many recent methods for unsupervised representation learning train models to be
invariant to different "views," or distorted versions of an input. However, desi
gning these views requires considerable trial and error by human experts, hinder
ing widespread adoption of unsupervised representation learning methods across d
omains and modalities. To address this, we propose viewmaker networks: generativ
e models that learn to produce useful views from a given input. Viewmakers are s
tochastic bounded adversaries: they produce views by generating and then adding
an ℓ_p-bounded perturbation to the input, and are trained adversarially wit
h respect to the main encoder network. Remarkably, when pretraining on CIFAR-10,
 our learned views enable comparable transfer accuracy to the well-tuned SimCLR
augmentations---despite not including transformations like cropping or color jit
ter. Furthermore, our learned views significantly outperform baseline augmentati
ons on speech recordings (+9 points on average) and wearable sensor data (+17 po
ints on average). Viewmaker views can also be combined with handcrafted views: t
hey improve robustness to common image corruptions and can increase transfer per
formance in cases where handcrafted views are less explored. These results sugge
st that viewmakers may provide a path towards more general representation learni
ng algorithms---reducing the domain expertise and effort needed to pretrain on a
 much wider set of domains. Code is available at https://github.com/alextamkin/v
iewmaker.
****

Robust Pruning at Initialization
Soufiane Hayou,Jean-Francois Ton,Arnaud Doucet,Yee Whye Teh
Overparameterized Neural Networks (NN) display state-of-the-art performance. How
ever, there is a growing need for smaller, energy-efficient, neural networks to
be able to use machine learning applications on devices with limited computation
al resources. A popular approach consists of using pruning techniques. While the
se techniques have traditionally focused on pruning pre-trained NN (LeCun et al.
,1990; Hassibi et al., 1993), recent work by Lee et al. (2018) has shown promisi
ng results when pruning at initialization. However, for Deep NNs, such procedure
s remain unsatisfactory as the resulting pruned networks can be difficult to tra
in and, for instance, they do not prevent one layer from being fully pruned. In

this paper, we provide a comprehensive theoretical analysis of Magnitude and Gra
dient based pruning at initialization and training of sparse architectures. Thi
s allows us to propose novel principled approaches which we validate experimenta
lly on a variety of NN architectures.
****

Improving Graph Neural Network Expressivity via Subgraph Isomorphism Counting
Giorgos Bouritsas,Fabrizio Frasca,Stefanos Zafeiriou,Michael M. Bronstein
While Graph Neural Networks (GNNs) have achieved remarkable results in a variety
 of applications, recent studies exposed important shortcomings in their ability
 to capture the structure of the underlying graph. It has been shown that the ex
pressive power of standard GNNs is bounded by the Weisfeiler-Lehman (WL) graph i
somorphism test, from which they inherit proven limitations such as the inabilit
y to detect and count graph substructures. On the other hand, there is significa
nt empirical evidence, e.g. in network science and bioinformatics, that substruc
tures are often informative for downstream tasks, suggesting that it is desirabl
e to design GNNs capable of leveraging this important source of information. To
this end, we propose a novel topologically-aware message passing scheme based on
 substructure encoding. We show that our architecture allows incorporating domai
n-specific inductive biases and that it is strictly more expressive than the WL
test. Importantly, in contrast to recent works on the expressivity of GNNs, we d
o not attempt to adhere to the WL hierarchy; this allows us to retain multiple a
ttractive properties of standard GNNs such as locality and linear network comple
xity, while being able to disambiguate even hard instances of graph isomorphism.
 We extensively evaluate our method on graph classification and regression tasks
 and show state-of-the-art results on multiple datasets including molecular grap
hs and social networks.
****

Non-robust Features through the Lens of Universal Perturbations
Sung Min Park,Kuo-An Wei,Kai Yuanqing Xiao,Jerry Li,Aleksander Madry
Recent work ties adversarial examples to existence of non-robust features: featu
res which are susceptible to small perturbations and believed to be unintelligib
le to humans, but still useful for prediction. We study universal adversarial pe
rturbations and demonstrate that the above picture is more nuanced. Specifically
, even though universal perturbations---similarly to standard adversarial pertur
bations---do leverage non-robust features, these features tend to be fundamental
ly different from the ``standard'' ones and, in particular, non-trivially human-
aligned. Namely, universal perturbations have more human-aligned locality and sp
atial invariance properties. However, we also show that these human-aligned non-
robust features have much less predictive signal than general non-robust feature
s. Our findings thus take a step towards improving our understanding of these pr
eviously unintelligible features.
****

Single-Photon Image Classification
Thomas Fischbacher,Luciano Sbaiz
Quantum Computing based Machine Learning mainly focuses on quantum computing har
dware that is experimentally challenging to realize due to requiring quantum gat
es that operate at very low temperature. We demonstrate the existence of a "quan
tum computing toy model" that illustrates key aspects of quantum information pro
cessing while being experimentally accessible with room temperature optics. Pond
ering the question of the theoretical classification accuracy performance limit
for MNIST (respectively "Fashion-MNIST") classifiers, subject to the constraint
that a decision has to be made after detection of the very first photon that pas
sed through an image-filter, we show that a machine learning system that is perm
itted to use quantum interference on the photon's state can substantially outper
form any machine learning system that can not. Specifically, we prove that a "c
lassical" MNIST (respectively "Fashion-MNIST") classifier cannot achieve an accu
racy of better than 21.28% (respectively 18.28% for "Fashion-MNIST") if it
 must make a decision after seeing a single photon falling on one of the $28\tim
es 28$ image pixels of a detector array. We further demonstrate that a classifi
er that is permitted to employ quantum interference by optically transforming th

e photon state prior to detection can achieve a classification accuracy of at le ast 41.27% for MNIST (respectively 36.14% for "Fashion-MNIST"). We show in detail how to train the corresponding quantum state transformation with TensorF low and also explain how this example can serve as a teaching tool for the measu rement process in quantum mechanics.

****
On Learning Read-once DNFs With Neural Networks
Ido Bronstein,Alon Brutzkus,Amir Globerson
Learning functions over Boolean variables is a fundamental problem in machine le arning. But not much is known about learning such functions by neural networks. Because learning these functions in the distribution free setting is NP-Hard, th ey are unlikely to be efficiently learnable by networks in this case. However, a ssuming the inputs are sampled from the uniform distribution, an important subse t of functions that are known to be efficiently learnable is read-once DNFs. Her e we focus on this setting where the functions are learned by a convex neural ne twork and gradient descent.
We first observe empirically that the learned neurons are aligned with the terms of the DNF, despite the fact that there are many zero-error networks that do no t have this property. Thus, the learning process has a clear inductive bias towa rds such logical formulas. To gain a better theoretical understanding of this ph enomenon we focus on minimizing the population risk. We show that this risk can be minimized by multiple networks: from ones that memorize data to ones that com pactly represent the DNF. We then set out to understand why gradient descent ``" chooses" the compact representation.
We use a computer assisted proof to prove the inductive bias for relatively smal l DNFs, and use it to design a process for reconstructing the DNF from the learn ed network. We then continue to provide theoretical insights on the learning pro cess and the loss surface to better understand the resulting inductive bias. For example, we show that the neurons in solutions with minimum l_2-norm of the w eights are also aligned with the terms of the DNF. Finally, we empirically show that our results are validated in the empirical case for high dimensional DNFs, more general network architectures and tabular datasets.
****
Generative Auto-Encoder: Controllable Synthesis with Disentangled Exploration
Yunhao Ge,Gan Xin,Zhi Xu,Yao Xiao,Yunkui Pang,Yining HE,Laurent Itti
Autoencoders perform a powerful information compression framework with are const ruction loss and can be a regularization module in different tasks, which has no generative ability itself. We wondering if an autoencoder gains generative abil ity without using GAN and VAE based modification, which are two mature methods.
Here we propose a new method: Disentanglement and ExplorationAutoencoder (DEA E), DEAE using a disentangle and exploration positive iteration achieve semantic controllable synthesis especially controllable mining the novel semantic value.
For instance, given only red, green, and blue object color while mining new ob ject color expressions in the image domain. The encoder of DEAE first turn the i nput sample into a disentangled latent code, then explore the latent codespace b y attribute oriented interpolation. To encourage interpolated latent code succe ssfully output a semantically meaningful sample by the decoder, we propose a reg ularization procedure by ’reuse’ encoder and constrain the output latent value w hich implicitly improves the quality of the interpolated sample. DEAE can become a generative model and synthesis semantic controllable samples by interpolating latent code, which can even synthesis novel attribute value never is shown in t he original dataset. Experiments demonstrate how disentanglement and exploration can boost each other which empowers autoencoder generative ability. We also dem onstrate that DEAE can improve the performance of downstream tasks compared with GANand VAE based generative model, especially in controllable data augmentation , dataset bias elimination (Fairness)
****
Can a Fruit Fly Learn Word Embeddings?
Yuchen Liang,Chaitanya Ryali,Benjamin Hoover,Leopold Grinberg,Saket Navlakha,Moh

ammed J Zaki,Dmitry Krotov
The mushroom body of the fruit fly brain is one of the best studied systems in n euroscience. At its core it consists of a population of Kenyon cells, which rece ive inputs from multiple sensory modalities. These cells are inhibited by the an terior paired lateral neuron, thus creating a sparse high dimensional representa tion of the inputs. In this work we study a mathematical formalization of this n etwork motif and apply it to learning the correlational structure between words and their context in a corpus of unstructured text, a common natural language pr ocessing (NLP) task. We show that this network can learn semantic representation s of words and can generate both static and context-dependent word embeddings. U nlike conventional methods (e.g., BERT, GloVe) that use dense representations fo r word embedding, our algorithm encodes semantic meaning of words and their cont ext in the form of sparse binary hash codes. The quality of the learned represen tations is evaluated on word similarity analysis, word-sense disambiguation, and document classification. It is shown that not only can the fruit fly network mo tif achieve performance comparable to existing methods in NLP, but, additionally , it uses only a fraction of the computational resources (shorter training time and smaller memory footprint).
****

Demystifying Loss Functions for Classification
Simon Kornblith,Honglak Lee,Ting Chen,Mohammad Norouzi
It is common to use the softmax cross-entropy loss to train neural networks on c lassification datasets where a single class label is assigned to each example. H owever, it has been shown that modifying softmax cross-entropy with label smooth ing or regularizers such as dropout can lead to higher performance. In this pape r, we compare a variety of loss functions and output layer regularization strate gies that improve performance on image classification tasks. We find differences in the outputs of networks trained with these different objectives, in terms of accuracy, calibration, out-of-distribution robustness, and predictions. However , differences in hidden representations of networks trained with different objec tives are restricted to the last few layers; representational similarity reveals no differences among network layers that are not close to the output. We show t hat all objectives that improve over vanilla softmax loss produce greater class separation in the penultimate layer of the network, which potentially accounts f or improved performance on the original task, but results in features that trans fer worse to other tasks.
****

VCNet and Functional Targeted Regularization For Learning Causal Effects of Cont inuous Treatments
Lizhen Nie,Mao Ye,qiang liu,Dan Nicolae
Motivated by the rising abundance of observational data with continuous treatmen ts, we investigate the problem of estimating the average dose-response curve (AD RF). Available parametric methods are limited in their model space, and previous attempts in leveraging neural network to enhance model expressiveness relied on partitioning continuous treatment into blocks and using separate heads for each block; this however produces in practice discontinuous ADRFs. Therefore, the qu estion of how to adapt the structure and training of neural network to estimate ADRFs remains open. This paper makes two important contributions. First, we prop ose a novel varying coefficient neural network (VCNet) that improves model expre ssiveness while preserving continuity of the estimated ADRF. Second, to improve finite sample performance, we generalize targeted regularization to obtain a dou bly robust estimator of the whole ADRF curve.
****

Distributed Adversarial Training to Robustify Deep Neural Networks at Scale
Gaoyuan Zhang,Songtao Lu,Sijia Liu,Xiangyi Chen,Pin-Yu Chen,Lee Martie,Lior Hore sh,Mingyi Hong
Current deep neural networks are vulnerable to adversarial attacks, where advers arial perturbations to the inputs can change or manipulate classification. To de fend against such attacks, an effective and popular approach, known as adversari al training, has been shown to mitigate the negative impact of adversarial attac

ks by virtue of a min-max robust training method. While effective, this approach is difficult to scale well to large models on large datasets (e.g., ImageNet) in general. To address this challenge, we propose distributed adversarial training (DAT), a large-batch adversarial training framework implemented over multiple machines. DAT supports one-shot and iterative attack generation methods, gradient quantization, and training over labeled and unlabeled data. Theoretically, we provide, under standard conditions in the optimization theory, the convergence rate of DAT to the first-order stationary points in general non-convex settings. Empirically, on ResNet-18 and -50 under CIFAR-10 and ImageNet, we demonstrate that DAT either matches or outperforms state-of-the-art robust accuracies and achieves a graceful training speedup.

****

Topology-Aware Segmentation Using Discrete Morse Theory

Xiaoling Hu,Yusu Wang,Li Fuxin,Dimitris Samaras,Chao Chen

In the segmentation of fine-scale structures from natural and biomedical images, per-pixel accuracy is not the only metric of concern. Topological correctness, such as vessel connectivity and membrane closure, is crucial for downstream analysis tasks. In this paper, we propose a new approach to train deep image segmentation networks for better topological accuracy. In particular, leveraging the power of discrete Morse theory (DMT), we identify global structures, including 1D skeletons and 2D patches, which are important for topological accuracy. Trained with a novel loss based on these global structures, the network performance is significantly improved especially near topologically challenging locations (such as weak spots of connections and membranes). On diverse datasets, our method achieves superior performance on both the DICE score and topological metrics.

****

A Wigner-Eckart Theorem for Group Equivariant Convolution Kernels

Leon Lang,Maurice Weiler

Group equivariant convolutional networks (GCNNs) endow classical convolutional networks with additional symmetry priors, which can lead to a considerably improved performance. Recent advances in the theoretical description of GCNNs revealed that such models can generally be understood as performing convolutions with G-steerable kernels, that is, kernels that satisfy an equivariance constraint themselves. While the G-steerability constraint has been derived, it has to date only been solved for specific use cases - a general characterization of G-steerable kernel spaces is still missing. This work provides such a characterization for the practically relevant case of G being any compact group. Our investigation is motivated by a striking analogy between the constraints underlying steerable kernels on the one hand and spherical tensor operators from quantum mechanics on the other hand. By generalizing the famous Wigner-Eckart theorem for spherical tensor operators, we prove that steerable kernel spaces are fully understood and parameterized in terms of 1) generalized reduced matrix elements, 2) Clebsch-Gordan coefficients, and 3) harmonic basis functions on homogeneous spaces.

****

Model-Targeted Poisoning Attacks with Provable Convergence

Fnu Suya,Saeed Mahloujifar,David Evans,Yuan Tian

In a poisoning attack, an adversary with control over a small fraction of the training data attempts to select that data in a way that induces a model that misbehaves in a particular way desired by the adversary, such as misclassifying certain inputs. We propose an efficient poisoning attack that can target a desired model based on online convex optimization. Unlike previous model-targeted poisoning attacks, our attack comes with provable convergence to any achievable target classifier. The distance from the induced classifier to the target classifier is inversely proportional to the square root of the number of poisoning points. We also provide a lower bound on the minimum number of poisoning points needed to achieve a given target classifier. Our attack is the first model-targeted poisoning attack that provides provable convergence, and in our experiments it either exceeds or matches the best state-of-the-art attacks in terms of attack success rate and distance to the target model. In addition, as an online attack our attack can incrementally determine nearly optimal poisoning points.

****
Unifying Graph Convolutional Neural Networks and Label Propagation
Hongwei Wang,Jure Leskovec
Label Propagation (LPA) and Graph Convolutional Neural Networks (GCN) are both message passing algorithms on graphs. Both solve the task of node classification but LPA propagates node label information across the edges of the graph, while GCN propagates and transforms node feature information. However, while conceptually similar, it is unclear how LPA and GCN can be combined under a unified framework to improve node classification. Here we study the relationship between LPA and GCN in terms of feature/label influence, in which we characterize how much the initial feature/label of one node influences the final feature/label of another node in GCN/LPA. Based on our theoretical analysis, we propose an end-to-end model that combines GCN and LPA. In our unified model, edge weights are learnable, and the LPA serves as regularization to assist the GCN in learning proper edge weights that lead to improved classification performance. Our model can also be seen as learning the weights for edges based on node labels, which is more task-oriented than existing feature-based attention models and topology-based diffusion models. In a number of experiments on real-world graphs, our model shows superiority over state-of-the-art graph neural networks in terms of node classification accuracy.
****
D4RL: Datasets for Deep Data-Driven Reinforcement Learning
Justin Fu,Aviral Kumar,Ofir Nachum,George Tucker,Sergey Levine
The offline reinforcement learning (RL) problem, also known as batch RL, refers to the setting where a policy must be learned from a static dataset, without additional online data collection. This setting is compelling as it potentially allows RL methods to take advantage of large, pre-collected datasets, much like how the rise of large datasets has fueled results in supervised learning in recent years. However, existing online RL benchmarks are not tailored towards the offline setting, making progress in offline RL difficult to measure. In this work, we introduce benchmarks specifically designed for the offline setting, guided by key properties of datasets relevant to real-world applications of offline RL. Examples of such properties include: datasets generated via hand-designed controllers and human demonstrators, multi-objective datasets where an agent can perform different tasks in the same environment, and datasets consisting of a mixtures of policies. To facilitate research, we release our benchmark tasks and datasets with a comprehensive evaluation of existing algorithms and an evaluation protocol together with an open-source codebase. We hope that our benchmark will focus research effort on methods that drive improvements not just on simulated tasks, but ultimately on the kinds of real-world problems where offline RL will have the largest impact.
****
Non-asymptotic Confidence Intervals of Off-policy Evaluation: Primal and Dual Bounds
Yihao Feng,Ziyang Tang,na zhang,qiang liu
Off-policy evaluation (OPE) is the task of estimating the expected reward of a given policy based on offline data previously collected under different policies. Therefore, OPE is a key step in applying reinforcement learning to real-world domains such as medical treatment, where interactive data collection is expensive or even unsafe. As the observed data tends to be noisy and limited, it is essential to provide rigorous uncertainty quantification, not just a point estimation, when applying OPE to make high stakes decisions. This work considers the problem of constructing non-asymptotic confidence intervals in infinite-horizon off-policy evaluation, which remains a challenging open question. We develop a practical algorithm through a primal-dual optimization-based approach, which leverages the kernel Bellman loss (KBL) of Feng et al. 2019 and a new martingale concentration inequality of KBL applicable to time-dependent data with unknown mixing conditions. Our algorithm makes minimum assumptions on the data and the function class of the Q-function, and works for the behavior-agnostic settings where the data is collected under a mix of arbitrary unknown behavior policies. We pr

esent empirical results that clearly demonstrate the advantages of our approach over existing methods.

****
Multi-Agent Collaboration via Reward Attribution Decomposition
Tianjun Zhang,Huazhe Xu,Xiaolong Wang,Yi Wu,Kurt Keutzer,Joseph E. Gonzalez,Yuandong Tian
Recent advances in multi-agent reinforcement learning (MARL) have achieved super-human performance in games like Quake 3 and Dota 2. Unfortunately, these techniques require orders-of-magnitude more training rounds than humans and don't generalize to new agent configurations even on the same game. In this work, we propose Collaborative Q-learning (CollaQ) that achieves state-of-the-art performance in the StarCraft multi-agent challenge and supports ad hoc team play. We first formulate multi-agent collaboration as a joint optimization on reward assignment and show that each agent has an approximately optimal policy that decomposes into two parts: one part that only relies on the agent's own state, and the other part that is related to states of nearby agents. Following this novel finding, CollaQ decomposes the Q-function of each agent into a self term and an interactive term, with a Multi-Agent Reward Attribution (MARA) loss that regularizes the training. CollaQ is evaluated on various StarCraft maps and shows that it outperforms existing state-of-the-art techniques (i.e., QMIX, QTRAN, and VDN) by improving the win rate by 40% with the same number of samples. In the more challenging ad hoc team play setting (i.e., reweight/add/remove units without re-training or finetuning), CollaQ outperforms previous SoTA by over 30%.
****
Model Patching: Closing the Subgroup Performance Gap with Data Augmentation
Karan Goel,Albert Gu,Yixuan Li,Christopher Re
Classifiers in machine learning are often brittle when deployed. Particularly concerning are models with inconsistent performance on specific subgroups of a class, e.g., exhibiting disparities in skin cancer classification in the presence or absence of a spurious bandage. To mitigate these performance differences, we introduce model patching, a two-stage framework for improving robustness that encourages the model to be invariant to subgroup differences, and focus on class information shared by subgroups. Model patching first models subgroup features within a class and learns semantic transformations between them, and then trains a classifier with data augmentations that deliberately manipulate subgroup features. We instantiate model patching with CAMEL, which (1) uses a CycleGAN to learn the intra-class, inter-subgroup augmentations, and (2) balances subgroup performance using a theoretically-motivated subgroup consistency regularizer, accompanied by a new robust objective. We demonstrate CAMEL’s effectiveness on 3 benchmark datasets, with reductions in robust error of up to 33% relative to the best baseline. Lastly, CAMEL successfully patches a model that fails due to spurious features on a real-world skin cancer dataset.
****
Measuring and mitigating interference in reinforcement learning
Vincent Liu,Adam M White,Hengshuai Yao,Martha White
Catastrophic interference is common in many network-based learning systems, and many proposals exist for mitigating it. But, before we overcome interference we must understand it better. In this work, we first provide a definition and novel measure of interference for value-based control methods such as Fitted Q Iteration and DQN. We systematically evaluate our measure of interference, showing that it correlates with forgetting, across a variety of network architectures. Our new interference measure allows us to ask novel scientific questions about commonly used deep learning architectures and develop new learning algorithms. In particular we show that updates on the last layer result in significantly higher interference than updates internal to the network. Lastly, we introduce a novel online-aware representation learning algorithm to minimize interference, and we empirically demonstrate that it improves stability and has lower interference.
****
Conditional Coverage Estimation for High-quality Prediction Intervals

Ziyi Huang,Henry Lam,Haofeng Zhang
Deep learning has achieved state-of-the-art performance to generate high-quality prediction intervals (PIs) for uncertainty quantification in regression tasks. The high-quality criterion requires PIs to be as narrow as possible, whilst maintaining a pre-specified level of data (marginal) coverage. However, most existing works for high-quality PIs lack accurate information on conditional coverage, which may cause unreliable predictions if it is significantly smaller than the marginal coverage. To address this problem, we propose a novel end-to-end framework which could output high-quality PIs and simultaneously provide their conditional coverage estimation. In doing so, we design a new loss function that is both easy-to-implement and theoretically justified via an exponential concentration bound. Our evaluation on real-world benchmark datasets and synthetic examples shows that our approach not only outperforms the state-of-the-arts on high-quality PIs in terms of average PI width, but also accurately estimates conditional coverage information that is useful in assessing model uncertainty.
****

SOLAR: Sparse Orthogonal Learned and Random Embeddings
Tharun Medini,Beidi Chen,Anshumali Shrivastava
Dense embedding models are commonly deployed in commercial search engines, wherein all the document vectors are pre-computed, and near-neighbor search (NNS) is performed with the query vector to find relevant documents. However, the bottleneck of indexing a large number of dense vectors and performing an NNS hurts the query time and accuracy of these models. In this paper, we argue that high-dimensional and ultra-sparse embedding is a significantly superior alternative to dense low-dimensional embedding for both query efficiency and accuracy. Extreme sparsity eliminates the need for NNS by replacing them with simple lookups, while its high dimensionality ensures that the embeddings are informative even when sparse. However, learning extremely high dimensional embeddings leads to blow up in the model size. To make the training feasible, we propose a partitioning algorithm that learns such high dimensional embeddings across multiple GPUs without any communication. This is facilitated by our novel asymmetric mixture of Sparse, Orthogonal, Learned and Random (SOLAR) Embeddings. The label vectors are random, sparse, and near-orthogonal by design, while the query vectors are learned and sparse. We theoretically prove that our way of one-sided learning is equivalent to learning both query and label embeddings. With these unique properties, we can successfully train 500K dimensional SOLAR embeddings for the tasks of searching through 1.6M books and multi-label classification on the three largest public datasets. We achieve superior precision and recall compared to the respective state-of-the-art baselines for each task with up to 10 times faster speed.
****

Neural representation and generation for RNA secondary structures
Zichao Yan,William L. Hamilton,Mathieu Blanchette
Our work is concerned with the generation and targeted design of RNA, a type of genetic macromolecule that can adopt complex structures which influence their cellular activities and functions. The design of large scale and complex biological structures spurs dedicated graph-based deep generative modeling techniques, which represents a key but underappreciated aspect of computational drug discovery. In this work, we investigate the principles behind representing and generating different RNA structural modalities, and propose a flexible framework to jointly embed and generate these molecular structures along with their sequence in a meaningful latent space. Equipped with a deep understanding of RNA molecular structures, our most sophisticated encoding and decoding methods operate on the molecular graph as well as the junction tree hierarchy, integrating strong inductive bias about RNA structural regularity and folding mechanism such that high structural validity, stability and diversity of generated RNAs are achieved. Also, we seek to adequately organize the latent space of RNA molecular embeddings with regard to the interaction with proteins, and targeted optimization is used to navigate in this latent space to search for desired novel RNA molecules.
****

Shortest-Path Constrained Reinforcement Learning for Sparse Reward Tasks

Sungryull Sohn,Sungtae Lee,Jongwook Choi,Harm van Seijen,Honglak Lee,Mehdi Fatemi

We propose the k-Shortest-Path (k-SP) constraint: a novel constraint on the agent’s trajectory that improves the sample-efficiency in sparse-reward MDPs. We show that any optimal policy necessarily satisfies the k-SP constraint. Notably, the k-SP constraint prevents the policy from exploring state-action pairs along the non-k-SP trajectories (e.g., going back and forth). However, in practice, excluding state-action pairs may hinder convergence of many RL algorithms. To overcome this, we propose a novel cost function that penalizes the policy violating SP constraint, instead of completely excluding it. Our numerical experiment in a tabular RL setting demonstrate that the SP constraint can significantly reduce the trajectory space of policy. As a result, our constraint enables more sample efficient learning by suppressing redundant exploration and exploitation. Our empirical experiment results on MiniGrid and DeepMind Lab show that the proposed method significantly improves proximal policy optimization (PPO) and outperforms existing novelty-seeking exploration methods including count-based exploration, indicating that it improves the sample efficiency by preventing the agent from taking redundant actions.

****

BeBold: Exploration Beyond the Boundary of Explored Regions

Tianjun Zhang,Huazhe Xu,Xiaolong Wang,Yi Wu,Kurt Keutzer,Joseph E. Gonzalez,Yuandong Tian

Efficient exploration under sparse rewards remains a key challenge in deep reinforcement learning. To guide exploration, previous work makes extensive use of intrinsic reward (IR). There are many heuristics for IR, including visitation counts, curiosity, and state-difference. In this paper, we analyze the pros and cons of each method and propose the regulated difference of inverse visitation counts as a simple but effective criterion for IR. The criterion helps the agent explore Beyond the Boundary of explored regions and mitigates common issues in count-based methods, such as short-sightedness and detachment. The resulting method, BeBold, solves the 12 most challenging procedurally-generated tasks in MiniGridwith just 120M environment steps, without any curriculum learning. In comparison, previous SoTA only solves 50%of the tasks. BeBold also achieves SoTAon multiple tasks in NetHack, a popular rogue-like game that contains more challenging procedurally-generated environments.

****

N-Bref : A High-fidelity Decompiler Exploiting Programming Structures

Cheng Fu,Kunlin Yang,Xinyun Chen,Yuandong Tian,Jishen Zhao

Binary decompilation is a powerful technique for analyzing and understanding software, when source code is unavailable. It is a critical problem in the computer security domain. With the success of neural machine translation (NMT), recent efforts on neural-based decompiler show promising results compared to traditional approaches. However, several key challenges remain: (i) Prior neural-based decompilers focus on simplified programs without considering sophisticated yet widely-used data types such as pointers; furthermore, many high-level expressions map to the same low-level code (expression collision), which incurs critical decompiling performance degradation; (ii) State-of-the-art NMT models(e.g., transformer and its variants) mainly deal with sequential data; this is inefficient for decompilation, where the input and output data are highly structured. In this paper, we propose N-Bref, a new framework for neural decompilers that addresses the two aforementioned challenges with two key design principles: (i)N-Bref designs a structural transformer with three key design components for better comprehension of structural data – an assembly encoder, an abstract syntax tree encoder, and a tree decoder, extending transformer models in the context of decompilation. (ii) N-Bref introduces a program generation tool that can control the complexity of code generation and removes expression collisions. Extensive experiments demonstrate that N-Bref outperforms previous neural-based decompilers by a margin of 6.1%/8.8% accuracy in datatype recovery and source code generation. In particular, N-Bref decompiled human-written Leetcode programs with complex library calls and data types in high accuracy.

****
A Mathematical Exploration of Why Language Models Help Solve Downstream Tasks
Nikunj Saunshi,Sadhika Malladi,Sanjeev Arora
Autoregressive language models, pretrained using large text corpora to do well on next word prediction, have been successful at solving many downstream tasks, even with zero-shot usage. However, there is little theoretical understanding of this success. This paper initiates a mathematical study of this phenomenon for the downstream task of text classification by considering the following questions: (1) What is the intuitive connection between the pretraining task of next word prediction and text classification? (2) How can we mathematically formalize this connection and quantify the benefit of language modeling? For (1), we hypothesize, and verify empirically, that classification tasks of interest can be reformulated as sentence completion tasks, thus making language modeling a meaningful pretraining task. With a mathematical formalization of this hypothesis, we make progress towards (2) and show that language models that are ϵ-optimal in cross-entropy (log-perplexity) learn features that can linearly solve such classification tasks with $\mathcal{O}(\sqrt{\epsilon})$ error, thus demonstrating that doing well on language modeling can be beneficial for downstream tasks. We experimentally verify various assumptions and theoretical findings, and also use insights from the analysis to design a new objective function that performs well on some classification tasks.
****
Adding Recurrence to Pretrained Transformers
Davis Yoshida,Allyson Ettinger,Kevin Gimpel
Fine-tuning a pretrained transformer for a downstream task has become a standard method in NLP in the last few years. While the results from these models are impressive, applying them can be extremely computationally expensive, as is pretraining new models with the latest architectures. We present a novel method for applying pretrained transformer language models which lowers their memory requirement both at training and inference time. An additional benefit is that our method removes the fixed context size constraint that most transformer models have, allowing for more flexible use. When applied to the GPT-2 language model, we find that our method attains better perplexity than an unmodified GPT-2 model on the PG-19 and WikiText-103 corpora, for a given amount of computation or memory.
****
Adversarial Problems for Generative Networks
Kalliopi Basioti,George V. Moustakides
We are interested in the design of generative networks. The training of these mathematical structures is mostly performed with the help of adversarial (min-max) optimization problems. We propose a simple methodology for constructing such problems assuring, at the same time, consistency of the corresponding solution. We give characteristic examples developed by our method, some of which can be recognized from other applications and some are introduced here for the first time. We compare various possibilities by applying them to well known datasets using neural networks of different configurations and sizes.
****
Temporal and Object Quantification Nets
Jiayuan Mao,Zhezheng Luo,Chuang Gan,Joshua B. Tenenbaum,Jiajun Wu,Leslie Pack Kaelbling,Tomer Ullman
We aim to learn generalizable representations for complex activities by quantifying over both entities and time, as in “the kicker is behind all the other players,” or “the player controls the ball until it moves toward the goal.” Such a structural inductive bias of object relations, object quantification, and temporal orders will enable the learned representation to generalize to situations with varying numbers of agents, objects, and time courses. In this paper, we present Temporal and Object Quantification Nets (TOQ-Nets), which provide such structural inductive bias for learning composable action concepts from time sequences that describe the properties and relations of multiple entities. We evaluate TOQ-Nets on two benchmarks: trajectory-based soccer event detection, and 6D pose-based manipulation concept learning. We demonstrate that TOQ-Nets can generalize fro

m small amounts of data to scenarios where there are more agents and objects tha
n were present during training. The learned concepts are also robust with respec
t to temporally warped sequences and easily transfer to other prediction tasks i
n a similar domain.
****

D3C: Reducing the Price of Anarchy in Multi-Agent Learning
Ian Gemp,Kevin McKee,Richard Everett,Edgar Alfredo Duenez-Guzman,Yoram Bachrach,
David Balduzzi,Andrea Tacchetti
Even in simple multi-agent systems, fixed incentives can lead to outcomes that a
re poor for the group and each individual agent. We propose a method, D3C, for o
nline adjustment of agent incentives that reduces the loss incurred at a Nash eq
uilibrium. Agents adjust their incentives by learning to mix their incentive wit
h that of other agents, until a compromise is reached in a distributed fashion.
We show that D3C improves outcomes for each agent and the group as a whole in se
veral social dilemmas including a traffic network with Braess’s paradox, a priso
ner’s dilemma, and several reinforcement learning domains.
****

On Alignment in Deep Linear Neural Networks
Adityanarayanan Radhakrishnan,Eshaan Nichani,Daniel Bernstein,Caroline Uhler
    We study the properties of alignment, a form of implicit regularization, in
linear neural networks under gradient descent. We define alignment for fully co
nnected networks with multidimensional outputs and show that it is a natural ext
ension of alignment in networks with 1-dimensional outputs as defined by Ji and
Telgarsky, 2018. While in fully connected networks, there always exists a globa
l minimum corresponding to an aligned solution, we analyze alignment as it relat
es to the training process. Namely, we characterize when alignment is an invari
ant of training under gradient descent by providing necessary and sufficient con
ditions for this invariant to hold. In such settings, the dynamics of gradient d
escent simplify, thereby allowing us to provide an explicit learning rate under
which the network converges linearly to a global minimum. We then analyze netwo
rks with layer constraints such as convolutional networks. In this setting, we p
rove that gradient descent is equivalent to projected gradient descent, and that
 alignment is impossible with sufficiently large datasets.

****

EigenGame: PCA as a Nash Equilibrium
Ian Gemp,Brian McWilliams,Claire Vernade,Thore Graepel
We present a novel view on principal components analysis as a competitive game i
n which each approximate eigenvector is controlled by a player whose goal is to
maximize their own utility function. We analyze the properties of this PCA game
and the behavior of its gradient based updates. The resulting algorithm---which
combines elements from Oja's rule with a generalized Gram-Schmidt orthogonaliza
tion---is naturally decentralized and hence parallelizable through message passi
ng. We demonstrate the scalability of the algorithm with experiments on large im
age datasets and neural network activations. We discuss how this new view of PCA
 as a differentiable game can lead to further algorithmic developments and insig
hts.
****

Noise or Signal: The Role of Image Backgrounds in Object Recognition
Kai Yuanqing Xiao,Logan Engstrom,Andrew Ilyas,Aleksander Madry
We assess the tendency of state-of-the-art object recognition models to depend o
n signals from image backgrounds. We create a toolkit for disentangling foregrou
nd and background signal on ImageNet images, and find that (a) models can achiev
e non-trivial accuracy by relying on the background alone, (b) models often misc
lassify images even in the presence of correctly classified foregrounds--up to 8
8% of the time with adversarially chosen backgrounds, and (c) more accurate mode
ls tend to depend on backgrounds less. Our analysis of backgrounds brings us clo
ser to understanding which correlations machine learning models use, and how the
y determine models' out of distribution performance.

****
Does enhanced shape bias improve neural network robustness to common corruptions ?
Chaithanya Kumar Mummadi,Ranjitha Subramaniam,Robin Hutmacher,Julien Vitay,Volke r Fischer,Jan Hendrik Metzen
Convolutional neural networks (CNNs) learn to extract representations of complex features, such as object shapes and textures to solve image recognition tasks. Recent work indicates that CNNs trained on ImageNet are biased towards features that encode textures and that these alone are sufficient to generalize to unseen test data from the same distribution as the training data but often fail to gen eralize to out-of-distribution data. It has been shown that augmenting the train ing data with different image styles decreases this texture bias in favor of inc reased shape bias while at the same time improving robustness to common corrupti ons, such as noise and blur. Commonly, this is interpreted as shape bias increas ing corruption robustness. However, this relationship is only hypothesized. We p erform a systematic study of different ways of composing inputs based on natural images, explicit edge information, and stylization. While stylization is essent ial for achieving high corruption robustness, we do not find a clear correlation between shape bias and robustness. We conclude that the data augmentation cause d by style-variation accounts for the improved corruption robustness and increa sed shape bias is only a byproduct.
****
Long Live the Lottery: The Existence of Winning Tickets in Lifelong Learning
Tianlong Chen,Zhenyu Zhang,Sijia Liu,Shiyu Chang,Zhangyang Wang
The lottery ticket hypothesis states that a highly sparsified sub-network can be trained in isolation, given the appropriate weight initialization. This paper e xtends that hypothesis from one-shot task learning, and demonstrates for the fir st time that such extremely compact and independently trainable sub-networks can be also identified in the lifelong learning scenario, which we call lifelong ti ckets. We show that the resulting lifelong ticket can further be leveraged to im prove the performance of learning over continual tasks. However, it is highly no n-trivial to conduct network pruning in the lifelong setting. Two critical roadb locks arise: i) As many tasks now arrive sequentially, finding tickets in a gree dy weight pruning fashion will inevitably suffer from the intrinsic bias, that t he earlier emerging tasks impact more; ii) As lifelong learning is consistently challenged by catastrophic forgetting, the compact network capacity of tickets m ight amplify the risk of forgetting. In view of those, we introduce two pruning options, e.g., top-down and bottom-up, for finding lifelong tickets. Compared to the top-down pruning that extends vanilla (iterative) pruning over sequential t asks, we show that the bottom-up one, which can dynamically shrink and (re-)expa nd model capacity, effectively avoids the undesirable excessive pruning in the e arly stage. We additionally introduce lottery teaching that further overcomes fo rgetting via knowledge distillation aided by external unlabeled data. Unifying t hose ingredients, we demonstrate the existence of very competitive lifelong tick ets, e.g., achieving 3-8% of the dense model size with even higher accuracy, com pared to strong class-incremental learning baselines on CIFAR-10/CIFAR-100/Tiny-ImageNet datasets. Codes available at https://github.com/VITA-Group/Lifelong-Lea rning-LTH.
****
Do Deeper Convolutional Networks Perform Better?
Eshaan Nichani,Adityanarayanan Radhakrishnan,Caroline Uhler
Over-parameterization is a recent topic of much interest in the machine lear ning community. While over-parameterized neural networks are capable of perfect ly fitting (interpolating) training data, these networks often perform well on t est data, thereby contradicting classical learning theory. Recent work provided an explanation for this phenomenon by introducing the double descent curve, sho wing that increasing model capacity past the interpolation threshold leads to a decrease in test error. In line with this, it was recently shown empirically an d theoretically that increasing neural network capacity through width leads to d ouble descent. In this work, we analyze the effect of increasing depth on test

performance. In contrast to what is observed for increasing width, we demonstra te through a variety of classification experiments on CIFAR10 and ImageNet-32 us ing ResNets and fully-convolutional networks that test performance worsens beyon d a critical depth. We posit an explanation for this phenomenon by drawing intu ition from the principle of minimum norm solutions in linear networks.

****
Compressing gradients in distributed SGD by exploiting their temporal correlatio n
Tharindu Adikari,Stark Draper
We propose SignXOR, a novel compression scheme that exploits temporal correlatio n of gradients for the purpose of gradient compression. Sign-based schemes such as Scaled-sign and SignSGD (Bernstein et al., 2018; Karimireddy et al., 2019) co mpress gradients by storing only the sign of gradient entries. These methods, ho wever, ignore temporal correlations between gradients. The equality or non-equal ity of signs of gradients in two consecutive iterations can be represented by a binary vector, which can be further compressed depending on its entropy. By impl ementing a rate-distortion encoder we increase the temporal correlation of gradi ents, lowering entropy and improving compression. We achieve theoretical converg ence of SignXOR by employing the two-way error-feedback approach introduced by Z heng et al. (2019). Zheng et al. (2019) show that two-way compression with error -feedback achieves the same asymptotic convergence rate as SGD, although converg ence is slower by a constant factor. We strengthen their analysis to show that t he rate of convergence of two-way compression with errorfeedback asymptotically is the same as that of SGD. As a corollary we prove that two-way SignXOR compres sion with error-feedback achieves the same asymptotic rate of convergence as SGD . We numerically evaluate our proposed method on the CIFAR-100 and ImageNet data sets and show that SignXOR requires less than 50% of communication traffic compa red to sending sign of gradients. To the best of our knowledge we are the first to present a gradient compression scheme that exploits temporal correlation of g radients.
****
Learning a Non-Redundant Collection of Classifiers
Daniel Pace,Alessandra Russo,Murray Shanahan
Supervised learning models constructed under the i.i.d. assumption have often be en shown to exploit spurious or brittle predictive signals instead of more robus t ones present in the training data. Inspired by Quality-Diversity algorithms, i n this work we train a collection of classifiers to learn distinct solutions to a classification problem, with the goal of learning to exploit a variety of pred ictive signals present in the training data. We propose an information-theoretic measure of model diversity based on minimizing an estimate of conditional total correlation of final layer representations across models given the label. We co nsider datasets with synthetically injected spurious correlations and evaluate o ur framework's ability to rapidly adapt to a change in distribution that destroy s the spurious correlation. We compare our method to a variety of baselines unde r this evaluation protocol, showing that it is competitive with other approaches while being more successful at isolating distinct signals. We also show that ou r model is competitive with Invariant Risk Minimization under this evaluation pr otocol without requiring access to the environment information required by IRM t o discriminate between spurious and robust signals.
****
Exploring the Uncertainty Properties of Neural Networks’ Implicit Priors in the Infinite-Width Limit
Ben Adlam,Jaehoon Lee,Lechao Xiao,Jeffrey Pennington,Jasper Snoek
Modern deep learning models have achieved great success in predictive accuracy f or many data modalities. However, their application to many real-world tasks is restricted by poor uncertainty estimates, such as overconfidence on out-of-distr ibution (OOD) data and ungraceful failing under distributional shift. Previous b enchmarks have found that ensembles of neural networks (NNs) are typically the b est calibrated models on OOD data. Inspired by this, we leverage recent theoreti

cal advances that characterize the function-space prior of an infinitely-wide NN as a Gaussian process, termed the neural network Gaussian process (NNGP). We use the NNGP with a softmax link function to build a probabilistic model for multi-class classification and marginalize over the latent Gaussian outputs to sample from the posterior. This gives us a better understanding of the implicit prior NNs place on function space and allows a direct comparison of the calibration of the NNGP and its finite-width analogue. We also examine the calibration of previous approaches to classification with the NNGP, which treat classification problems as regression to the one-hot labels. In this case the Bayesian posterior is exact, and we compare several heuristics to generate a categorical distribution over classes. We find these methods are well calibrated under distributional shift. Finally, we consider an infinite-width final layer in conjunction with a pre-trained embedding. This replicates the important practical use case of transfer learning and allows scaling to significantly larger datasets. As well as achieving competitive predictive accuracy, this approach is better calibrated than its finite width analogue.
****

Hierarchical Autoregressive Modeling for Neural Video Compression
Ruihan Yang,Yibo Yang,Joseph Marino,Stephan Mandt
Recent work by Marino et al. (2020) showed improved performance in sequential density estimation by combining masked autoregressive flows with hierarchical latent variable models. We draw a connection between such autoregressive generative models and the task of lossy video compression. Specifically, we view recent neural video compression methods (Lu et al., 2019; Yang et al., 2020b; Agustssonet al., 2020) as instances of a generalized stochastic temporal autoregressive transform, and propose avenues for enhancement based on this insight. Comprehensive evaluations on large-scale video data show improved rate-distortion performance over both state-of-the-art neural and conventional video compression methods.
****

LLBoost: Last Layer Perturbation to Boost Pre-trained Neural Networks
Adityanarayanan Radhakrishnan,Neha Prasad,Caroline Uhler
While deep networks have produced state-of-the-art results in several domains from image classification to machine translation, hyper-parameter selection remains a significant computational bottleneck. In order to produce the best possible model, practitioners often search across random seeds or use ensemble methods. As models get larger, any method to improve neural network performance that involves re-training becomes intractable. For example, computing the training accuracy of FixResNext-101 (829 million parameters) on ImageNet takes roughly 1~day when using 1~GPU.

In this work, we present LLBoost, a theoretically-grounded, computationally-efficient method to boost the validation accuracy of pre-trained over-parameterized models without impacting the original training accuracy. LLBoost adjusts the last layer of a neural network by adding a term that is orthogonal to the training feature matrix, which is constructed by applying all layers but the last to the training data. We provide an efficient implementation of LLBoost on the GPU and demonstrate that LLBoost, run using only 1 GPU, improves the test/validation accuracy of pre-trained models on CIFAR10, ImageNet32, and ImageNet. In the over-parameterized linear regression setting, we prove that LLBoost reduces the generalization error of any interpolating solution with high probability without affecting training error.
****

On the Dynamic Regret of Online Multiple Mirror Descent
Nima Eshraghi,and Ben Liang
We study the problem of online convex optimization, where a learner makes sequential decisions to minimize an accumulation of strongly convex costs over time. The quality of decisions is given in terms of the dynamic regret, which measures the performance of the learner relative to a sequence of dynamic minimizers. Prior works on gradient descent and mirror descent have shown that the dynamic regret can be upper bounded using the path length, which depend on the differences between successive minimizers, and an upper bound using the squared path length h

as also been shown when multiple gradient queries are allowed per round. However, they all require the cost functions to be Lipschitz continuous, which imposes a strong requirement especially when the cost functions are also strongly convex. In this work, we consider Online Multiple Mirror Descent (OMMD), which is based on mirror descent but uses multiple mirror descent steps per online round. Without requiring the cost functions to be Lipschitz continuous, we derive two upper bounds on the dynamic regret based on the path length and squared path length. We further derive a third upper bound that relies on the gradient of cost functions, which can be much smaller than the path length or squared path length, especially when the cost functions are smooth but fluctuate over time. Thus, we show that the dynamic regret of OMMD scales linearly with the minimum among the path length, squared path length, and sum squared gradients. Our experimental results further show substantial improvement on the dynamic regret compared with existing alternatives.
****
DeLighT: Deep and Light-weight Transformer
Sachin Mehta,Marjan Ghazvininejad,Srinivasan Iyer,Luke Zettlemoyer,Hannaneh Hajishirzi
We introduce a deep and light-weight transformer, DeLighT, that delivers similar or better performance than standard transformer-based models with significantly fewer parameters. DeLighT more efficiently allocates parameters both (1) within each Transformer block using the DeLighT transformation, a deep and light-weight transformation and (2) across blocks using block-wise scaling, that allows for shallower and narrower DeLighT blocks near the input and wider and deeper DeLighT blocks near the output. Overall, DeLighT networks are 2.5 to 4 times deeper than standard transformer models and yet have fewer parameters and operations. Experiments on benchmark machine translation and language modeling tasks show that DeLighT matches or improves the performance of baseline Transformers with 2 to 3 times fewer parameters on average.
****
Learning A Minimax Optimizer: A Pilot Study
Jiayi Shen,Xiaohan Chen,Howard Heaton,Tianlong Chen,Jialin Liu,Wotao Yin,Zhangyang Wang
Solving continuous minimax optimization is of extensive practical interest, yet notoriously unstable and difficult. This paper introduces the learning to optimize(L2O) methodology to the minimax problems for the first time and addresses its accompanying unique challenges. We first present Twin-L2O, the first dedicated minimax L2O method consisting of two LSTMs for updating min and max variables separately. The decoupled design is found to facilitate learning, particularly when the min and max variables are highly asymmetric. Empirical experiments on a variety of minimax problems corroborate the effectiveness of Twin-L2O. We then discuss a crucial concern of Twin-L2O, i.e., its inevitably limited generalizability to unseen optimizees. To address this issue, we present two complementary strategies. Our first solution, Enhanced Twin-L2O, is empirically applicable for general minimax problems, by improving L2O training via leveraging curriculum learning. Our second alternative, called Safeguarded Twin-L2O, is a preliminary theoretical exploration stating that under some strong assumptions, it is possible to theoretically establish the convergence of Twin-L2O. We benchmark our algorithms on several testbed problems and compare against state-of-the-art minimax solvers. The code is available at: https://github.com/VITA-Group/L2O-Minimax.
****
First-Order Optimization Algorithms via Discretization of Finite-Time Convergent Flows
Mouhacine Benosman,Orlando Romero,Anoop Cherian
In this paper, we investigate the performance of several discretization algorithms for two first-order finite-time optimization flows. These flows are, namely, 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 introduce three discretization methods for these first-order finite-time flows, and provide converge

nce guarantees. We then apply the proposed algorithms in training neural networks and empirically test their performances on three standard datasets, namely, CIFAR10, SVHN, and MNIST. Our results show that our schemes demonstrate faster convergences against standard optimization alternatives, while achieving equivalent or better accuracy.
****
Beyond Prioritized Replay: Sampling States in Model-Based RL via Simulated Priorities
Jincheng Mei,Yangchen Pan,Martha White,Amir-massoud Farahmand,Hengshuai Yao
The prioritized Experience Replay (ER) method has attracted great attention; however, there is little theoretical understanding of such prioritization strategy and why they help. In this work, we revisit prioritized ER and, in an ideal setting, show equivalence to minimizing cubic loss, providing theoretical insight into why it improves upon uniform sampling. This theoretical equivalence highlights two limitations of current prioritized experience replay methods: insufficient coverage of the sample space and outdated priorities of training samples. This motivates our model-based approach, which does not suffer from these limitations. Our key idea is to actively search for high priority states using gradient ascent. Under certain conditions, we prove that the hypothetical experiences generated from these states are sampled proportionally to approximately true priorities. We also characterize the distance between the sampling distribution of our method and the true prioritized sampling distribution. Our experiments on both benchmark and application-oriented domains show that our approach achieves superior performance over baselines.
****
ALFA: Adversarial Feature Augmentation for Enhanced Image Recognition
Tianlong Chen,Yu Cheng,Zhe Gan,Yu Hu,Zhangyang Wang,Jingjing Liu
Adversarial training is an effective method to combat adversarial attacks in order to create robust neural networks. By using an auxiliary batch normalization on adversarial examples, it has been shown recently to possess great potential in improving the generalization ability of neural networks for image recognition as well. However, crafting pixel-level adversarial perturbations is computationally expensive. To address this issue, we propose AdversariaL Feature Augmentation (ALFA), which advocates adversarial training on the intermediate layers of feature embeddings. ALFA utilizes both clean and adversarial augmented features jointly to enhance standard trained networks. To eliminate laborious tuning of key parameters such as locations and strength of feature augmentations, we further design a learnable adversarial feature augmentation (L-ALFA) framework to automatically adjust the perturbation magnitude of each perturbed feature. Extensive experiments demonstrate that our proposed ALFA and L-ALFA methods achieve significant and consistent generalization improvement over strong baselines on CIFAR-10, CIFAR-100, and ImageNet benchmarks across different backbone networks for image recognition.
****
A Neural Network MCMC sampler that maximizes Proposal Entropy
ZENGYI LI,Yubei Chen,Friedrich Sommer
Markov Chain Monte Carlo (MCMC) methods sample from unnormalized probability distributions and offer guarantees of exact sampling. However, in the continuous case, unfavorable geometry of the target distribution can greatly limit the efficiency of MCMC methods. Augmenting samplers with neural networks can potentially improve their efficiency. Previous neural network based samplers were trained with objectives that either did not explicitly encourage exploration, or used a L2 jump objective which could only be applied to well structured distributions. Thus it seems promising to instead maximize the proposal entropy for adapting the proposal to distributions of any shape. To allow direct optimization of the proposal entropy, we propose a neural network MCMC sampler that has a flexible and tractable proposal distribution. Specifically, our network architecture utilizes the gradient of the target distribution for generating proposals. Our model achieves significantly higher efficiency than previous neural network MCMC techniques in a variety of sampling tasks. Further, the sampler is applied on training of

a convergent energy-based model of natural images. The learned sampler achieves significantly higher proposal entropy and sample quality compared to Langevin dynamics sampler.
****
Online Learning under Adversarial Corruptions
Pranjal Awasthi,Sreenivas Gollapudi,Kostas Kollias,Apaar Sadhwani
We study the design of efficient online learning algorithms tolerant to adversarially corrupted rewards. In particular, we study settings where an online algorithm makes a prediction at each time step, and receives a stochastic reward from the environment that can be arbitrarily corrupted with probability $\epsilon \in [0,\frac 1 2)$. Here ϵ is the noise rate the characterizes the strength of the adversary. As is standard in online learning, we study the design of algorithms with small regret over a period of time steps. However, while the algorithm observes corrupted rewards, we require its regret to be small with respect to the true uncorrupted reward distribution. We build upon recent advances in robust estimation for unsupervised learning problems to design robust online algorithms with near optimal regret in three different scenarios: stochastic multi-armed bandits, linear contextual bandits, and Markov Decision Processes~(MDPs) with stochastic rewards and transitions. Finally, we provide empirical evidence regarding the robustness of our proposed algorithms on synthetic and real datasets.
****
On Disentangled Representations Learned From Correlated Data
Frederik Träuble,Elliot Creager,Niki Kilbertus,Anirudh Goyal,Francesco Locatello,Bernhard Schölkopf,Stefan Bauer
Despite impressive progress in the last decade, it still remains an open challenge to build models that generalize well across multiple tasks and datasets. One path to achieve this is to learn meaningful and compact representations, in which different semantic aspects of data are structurally disentangled. The focus of disentanglement approaches has been on separating independent factors of variation despite the fact that real-world observations are often not structured into meaningful independent causal variables. In this work, we bridge the gap to real-world scenarios by analyzing the behavior of most prominent methods and disentanglement scores on correlated data in a large scale empirical study (including 4260 models). We show that systematically induced correlations in the dataset are being learned and reflected in the latent representations, while widely used disentanglement scores fall short of capturing these latent correlations. Finally, we demonstrate how to disentangle these latent correlations using weak supervision, even if we constrain this supervision to be causally plausible. Our results thus support the argument to learn independent mechanisms rather than independent factors of variations.
****
Playing Nondeterministic Games through Planning with a Learned Model
Thomas Willkens,Jordan Pollack
The MuZero algorithm is known for achieving high-level performance on traditional zero-sum two-player games of perfect information such as chess, Go, and shogi, as well as visual, non-zero sum, single-player environments such as the Atari suite. Despite lacking a perfect simulator and employing a learned model of environmental dynamics, MuZero produces game-playing agents comparable to its predecessor AlphaZero. However, the current implementation of MuZero is restricted only to deterministic environments. This paper presents Nondeterministic MuZero (NDMZ), an extension of MuZero for nondeterministic, two-player, zero-sum games of perfect information. Borrowing from Nondeterministic Monte Carlo Tree Search and the theory of extensive-form games, NDMZ formalizes chance as a player in the game and incorporates it into the MuZero network architecture and tree search. Experiments show that NDMZ is capable of learning effective strategies and an accurate model of the game.
****
Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift
Marvin Mengxin Zhang,Henrik Marklund,Nikita Dhawan,Abhishek Gupta,Sergey Levine,

Chelsea Finn
A fundamental assumption of most machine learning algorithms is that the trainin g and test data are drawn from the same underlying distribution. However, this a ssumption is violated in almost all practical applications: machine learning sys tems are regularly tested under distribution shift, due to temporal correlations , particular end users, or other factors. In this work, we consider the setting where the training data are structured into groups and test time shifts correspo nd to changes in the group distribution. Prior work has approached this problem by attempting to be robust to all possible test time distributions, which may de grade average performance. In contrast, we propose to use ideas from meta-learni ng to learn models that are adaptable, such that they can adapt to shift at test time using a batch of unlabeled test points. We acquire such models by learning to adapt to training batches sampled according to different distributions, whic h simulate structural shifts that may occur at test time. Our primary contributi on is to introduce the framework of adaptive risk minimization (ARM), a formaliz ation of this setting that lends itself to meta-learning. We develop meta-learni ng methods for solving the ARM problem, and compared to a variety of prior metho ds, these methods provide substantial gains on image classification problems in the presence of shift.
****
GRF: Learning a General Radiance Field for 3D Scene Representation and Rendering
Alex Trevithick,Bo Yang
We present a simple yet powerful implicit neural function that can represent and render arbitrarily complex 3D scenes in a single network only from 2D observati ons. The function models 3D scenes as a general radiance field, which takes a se t of 2D images with camera poses and intrinsics as input, constructs an internal representation for each 3D point of the scene, and renders the corresponding ap pearance and geometry of any 3D point viewing from an arbitrary angle. The key t o our approach is to explicitly integrate the principle of multi-view geometry t o obtain the internal representations from observed 2D views, such that the lear ned implicit representations empirically remain multi-view consistent. In additi on, we introduce an effective neural module to learn general features for each p ixel in 2D images, allowing the constructed internal 3D representations to be ge neral as well. Extensive experiments demonstrate the superiority of our approach .
****
Prepare for the Worst: Generalizing across Domain Shifts with Adversarial Batch Normalization
Manli Shu,Zuxuan Wu,Micah Goldblum,Tom Goldstein
Adversarial training is the industry standard for producing models that are robu st to small adversarial perturbations. However, machine learning practitioners need models that are robust to other kinds of changes that occur naturally, such as changes in the style or illumination of input images. Such changes in input distribution have been effectively modeled as shifts in the mean and variance of deep image features. We adapt adversarial training by adversarially perturbin g these feature statistics, rather than image pixels, to produce models that are robust to distributional shifts. We also visualize images from adversarially cr afted distributions. Our method, Adversarial Batch Normalization (AdvBN), signif icantly improves the performance of ResNet-50 on ImageNet-C (+8.1%), Stylized-Im ageNet (+6.7%), and ImageNet-Instagram (+3.9%) over standard training practices. In addition, we demonstrate that AdvBN can also improve generalization on sema ntic segmentation.
****
Language-Mediated, Object-Centric Representation Learning
Ruocheng Wang,Jiayuan Mao,Samuel Gershman,Jiajun Wu
We present Language-mediated, Object-centric Representation Learning (LORL), lea rning disentangled, object-centric scene representations from vision and languag e. LORL builds upon recent advances in unsupervised object segmentation, notably MONet and Slot Attention. Just like these algorithms, LORL also learns an objec t-centric representation by reconstructing the input image. But LORL further lea

rns to associate the learned representations to concepts, i.e., words for object categories, properties, and spatial relationships, from language input. These object-centric concepts derived from language facilitate the learning of object-centric representations. LORL can be integrated with various unsupervised segmentation algorithms that are language-agnostic. Experiments show that LORL consistently improves the performance of MONet and Slot Attention on two datasets via the help of language. We also show that concepts learned by LORL aid downstream tasks such as referential expression interpretation.
****

Weighted Bellman Backups for Improved Signal-to-Noise in Q-Updates
Kimin Lee,Michael Laskin,Aravind Srinivas,Pieter Abbeel
Off-policy deep reinforcement learning (RL) has been successful in a range of challenging domains. However, standard off-policy RL algorithms can suffer from low signal and even instability in Q-learning because target values are derived from current Q-estimates, which are often noisy. To mitigate the issue, we propose ensemble-based weighted Bellman backups, which re-weight target Q-values based on uncertainty estimates from a Q-ensemble. We empirically observe that the proposed method stabilizes and improves learning on both continuous and discrete control benchmarks. We also specifically investigate the signal-to-noise aspect by studying environments with noisy rewards, and find that weighted Bellman backups significantly outperform standard Bellman backups. Furthermore, since our weighted Bellman backups rely on maintaining an ensemble, we investigate how weighted Bellman backups interact with UCB Exploration. By enforcing the diversity between agents using Bootstrap, we show that these different ideas are largely orthogonal and can be fruitfully integrated, together further improving the performance of existing off-policy RL algorithms, such as Soft Actor-Critic and Rainbow DQN, for both continuous and discrete control tasks on both low-dimensional and high-dimensional environments.
****

Multimodal Attention for Layout Synthesis in Diverse Domains
Kamal Gupta,Vijay Mahadevan,Alessandro Achille,Justin Lazarow,Larry S. Davis,Abhinav Shrivastava
We address the problem of scene layout generation for diverse domains such as images, mobile applications, documents and 3D objects. Most complex scenes, natural or human-designed, can be expressed as a meaningful arrangement of simpler compositional graphical primitives. Generating a new layout or extending an existing layout requires understanding the relationships between these primitives. To do this, we propose a multimodal attention framework, MMA, that leverages self-attention to learn contextual relationships between layout elements and generate novel layouts in a given domain. Our framework allows us to generate a new layout either from an empty set or from an initial seed set of primitives, and can easily scale to support an arbitrary of primitives per layout. Further, our analyses show that the model is able to automatically capture the semantic properties of the primitives. We propose simple improvements in both representation of layout primitives, as well as training methods to demonstrate competitive performance in very diverse data domains such as object bounding boxes in natural images (COCO bounding boxes), documents (PubLayNet), mobile applications (RICO dataset) as well as 3D shapes (PartNet).
****

Fuzzy Tiling Activations: A Simple Approach to Learning Sparse Representations Online
Yangchen Pan,Kirby Banman,Martha White
Recent work has shown that sparse representations---where only a small percentage of units are active---can significantly reduce interference. Those works, however, relied on relatively complex regularization or meta-learning approaches, that have only been used offline in a pre-training phase. In this work, we pursue a direction that achieves sparsity by design, rather than by learning. Specifically, we design an activation function that produces sparse representations deterministically by construction, and so is more amenable to online training. The idea relies on the simple approach of binning, but overcomes the two key limitatio

ns of binning: zero gradients for the flat regions almost everywhere, and lost precision---reduced discrimination---due to coarse aggregation. We introduce a Fuzzy Tiling Activation (FTA) that provides non-negligible gradients and produces overlap between bins that improves discrimination. We first show that FTA is robust under covariate shift in a synthetic online supervised learning problem, where we can vary the level of correlation and drift. Then we move to the deep reinforcement learning setting and investigate both value-based and policy gradient algorithms that use neural networks with FTAs, in classic discrete control and Mujoco continuous control environments. We show that algorithms equipped with FTAs are able to learn a stable policy faster without needing target networks on most domains.

****

f-Domain-Adversarial Learning: Theory and Algorithms for Unsupervised Domain Adaptation with Neural Networks

David Acuna,Guojun Zhang,Marc T Law,Sanja Fidler

The problem of unsupervised domain adaptation arises in a variety of practical applications where the distribution of the training samples differs from those used at test time. The existing theory of domain adaptation derived generalization bounds based on divergence measures that are hard to optimize in practice. This has led to a large disconnect between theory and state-of-the-art methods. In this paper, we propose a novel domain-adversarial framework that introduces new theory for domain adaptation and leads to practical learning algorithms with neural networks. In particular, we derive a novel generalization bound that utilizes a new measure of discrepancy between distributions based on a variational characterization of f-divergences. We show that our bound recovers the theoretical results from Ben-David et al. (2010a) as a special case with a particular choice of divergence, and also supports divergences typically used in practice. We derive a general algorithm for domain-adversarial learning for the complete family of f-divergences. We provide empirical results for several f-divergences and show that some, not considered previously in domain-adversarial learning, achieve state-of-the-art results in practice. We provide empirical insights into how choosing a particular divergence affects the transfer performance on real-world datasets. By further recognizing the optimization problem as a Stackelberg game, we utilize the latest optimizers from the game optimization literature, achieving additional performance boosts in our training algorithm. We show that our f-domain adversarial framework achieves state-of-the-art results on the challenging Office-31 and Office-Home datasets without extra hyperparameters.

****

Expectigrad: Fast Stochastic Optimization with Robust Convergence Properties

Brett Daley,Christopher Amato

Many popular adaptive gradient methods such as Adam and RMSProp rely on an exponential moving average (EMA) to normalize their stepsizes. While the EMA makes these methods highly responsive to new gradient information, recent research has shown that it also causes divergence on at least one convex optimization problem. We propose a novel method called Expectigrad, which adjusts stepsizes according to a per-component unweighted mean of all historical gradients and computes a bias-corrected momentum term jointly between the numerator and denominator. We prove that Expectigrad cannot diverge on every instance of the optimization problem known to cause Adam to diverge. We also establish a regret bound in the general stochastic nonconvex setting that suggests Expectigrad is less susceptible to gradient variance than existing methods are. Testing Expectigrad on several high-dimensional machine learning tasks, we find it often performs favorably to state-of-the-art methods with little hyperparameter tuning.

****

FedBE: Making Bayesian Model Ensemble Applicable to Federated Learning

Hong-You Chen,Wei-Lun Chao

Federated learning aims to collaboratively train a strong global model by accessing users' locally trained models but not their own data. A crucial step is therefore to aggregate local models into a global model, which has been shown challenging when users have non-i.i.d. data. In this paper, we propose a novel aggrega

tion algorithm named FedBE, which takes a Bayesian inference perspective by sampling higher-quality global models and combining them via Bayesian model Ensemble, leading to much robust aggregation. We show that an effective model distribution can be constructed by simply fitting a Gaussian or Dirichlet distribution to the local models. Our empirical studies validate FedBE's superior performance, especially when users' data are not i.i.d. and when the neural networks go deeper. Moreover, FedBE is compatible with recent efforts in regularizing users' model training, making it an easily applicable module: you only need to replace the aggregation method but leave other parts of your federated learning algorithm intact.

****

Should Ensemble Members Be Calibrated?

Xixin Wu,Mark Gales

Underlying the use of statistical approaches for a wide range of applications is the assumption that the probabilities obtained from a statistical model are representative of the “true” probability that event, or outcome, will occur. Unfortunately, for modern deep neural networks this is not the case, they are often observed to be poorly calibrated. Additionally, these deep learning approaches make use of large numbers of model parameters, motivating the use of Bayesian, or ensemble approximation, approaches to handle issues with parameter estimation. This paper explores the application of calibration schemes to deep ensembles from both a theoretical perspective and empirically on a standard image classification task, CIFAR-100. The underlying theoretical requirements for calibration, and associated calibration criteria, are first described. It is shown that well calibrated ensemble members will not necessarily yield a well calibrated ensemble prediction, and if the ensemble prediction is well calibrated its performance cannot exceed that of the average performance of the calibrated ensemble members. On CIFAR-100 the impact of calibration for ensemble prediction, and associated calibration is evaluated. Additionally the situation where multiple different topologies are combined together is discussed.

****

Randomized Automatic Differentiation

Deniz Oktay,Nick McGreivy,Joshua Aduol,Alex Beatson,Ryan P Adams

The successes of deep learning, variational inference, and many other fields have been aided by specialized implementations of reverse-mode automatic differentiation (AD) to compute gradients of mega-dimensional objectives. The AD techniques underlying these tools were designed to compute exact gradients to numerical precision, but modern machine learning models are almost always trained with stochastic gradient descent. Why spend computation and memory on exact (minibatch) gradients only to use them for stochastic optimization? We develop a general framework and approach for randomized automatic differentiation (RAD), which can allow unbiased gradient estimates to be computed with reduced memory in return for variance. We examine limitations of the general approach, and argue that we must leverage problem specific structure to realize benefits. We develop RAD techniques for a variety of simple neural network architectures, and show that for a fixed memory budget, RAD converges in fewer iterations than using a small batch size for feedforward networks, and in a similar number for recurrent networks. We also show that RAD can be applied to scientific computing, and use it to develop a low-memory stochastic gradient method for optimizing the control parameters of a linear reaction-diffusion PDE representing a fission reactor.

****

Unsupervised Anomaly Detection by Robust Collaborative Autoencoders

Boyang Liu,Ding Wang,Kaixiang Lin,Pang-Ning Tan,Jiayu Zhou

Unsupervised anomaly detection plays a crucial role in many critical applications. Driven by the success of deep learning, recent years have witnessed growing interests in applying deep neural networks (DNNs) to anomaly detection problems. A common approach is to use autoencoders to learn a feature representation for the normal (non-anomalous) observations in the data. The reconstruction error of the autoencoder is then used as outlier scores to detect anomalies. However, due to the high complexity brought upon by over-parameterization of DNNs, the recon

struction error of the anomalies could also be small, which hampers the effectiv
eness of these methods. To alleviate this problem, we propose a robust framework
 using collaborative autoencoders to jointly identify normal observations from t
he data while learning its feature representation. We investigate the theoretica
l properties of the framework and empirically show its outstanding performance a
s compared to other DNN-based methods. Our experimental results also show the re
siliency of the framework to missing values compared to other baseline methods.
****

JAKET: Joint Pre-training of Knowledge Graph and Language Understanding

Donghan Yu,Chenguang Zhu,Yiming Yang,Michael Zeng

Knowledge graphs (KGs) contain rich information about world knowledge, entities,
 and relations. Thus, they can be great supplements to existing pre-trained lang
uage models. However, it remains a challenge to efficiently integrate informatio
n from KG into language modeling. And the understanding of a knowledge graph req
uires related context. We propose a novel joint pre-training framework, JAKET, t
o model both the knowledge graph and language. The knowledge module and language
 module provide essential information to mutually assist each other: the knowled
ge module produces embeddings for entities in text while the language module gen
erates context-aware initial embeddings for entities and relations in the graph.
 Our design enables the pre-trained model to easily adapt to unseen knowledge gr
aphs in new domains. Experimental results on several knowledge-aware NLP tasks s
how that our proposed framework achieves superior performance by effectively lev
eraging knowledge in language understanding.
****

Memory-Efficient Semi-Supervised Continual Learning: The World is its Own Replay
 Buffer

James Smith,Jonathan C Balloch,Yen-Chang Hsu,Zsolt Kira

Rehearsal is a critical component for class-incremental continual learning, yet
it requires a substantial memory budget. Our work investigates whether we can si
gnificantly reduce this memory budget by leveraging unlabeled data from an agent
's environment in a realistic and challenging continual learning paradigm. Speci
fically, we explore and formalize a novel semi-supervised continual learning (SS
CL) setting, where labeled data is scarce yet non-i.i.d. unlabeled data from the
 agent's environment is plentiful. Importantly, data distributions in the SSCL s
etting are realistic and therefore reflect object class correlations between, an
d among, the labeled and unlabeled data distributions. We show that a strategy b
uilt on pseudo-labeling, consistency regularization, Out-of-Distribution (OoD) d
etection, and knowledge distillation reduces forgetting in this setting. Our app
roach, DistillMatch, increases performance over the state-of-the-art by no less
than 8.7% average task accuracy and up to a 54.5% increase in average task accur
acy in SSCL CIFAR-100 experiments. Moreover, we demonstrate that DistillMatch ca
n save up to 0.23 stored images per processed unlabeled image compared to the ne
xt best method which only saves 0.08. Our results suggest that focusing on reali
stic correlated distributions is a significantly new perspective, which accentua
tes the importance of leveraging the world's structure as a continual learning s
trategy.
****

Provable Robust Learning for Deep Neural Networks under Agnostic Corrupted Super
vision

Boyang Liu,Mengying Sun,Ding Wang,Pang-Ning Tan,Jiayu Zhou

Training deep neural models in the presence of corrupted supervisions is challen
ging as the corrupted data points may significantly impact the generalization pe
rformance. To alleviate this problem, we present an efficient robust algorithm
that achieves strong guarantees without any assumption on the type of corruption
 and provides a unified framework for both classification and regression problem
s. Different from many existing approaches that quantify the quality of individu
al data points (e.g., loss values) and filter out data points accordingly, the p
roposed algorithm focuses on controlling the collective impact of data points on
 the averaged gradient. Even when a corrupted data point failed to be excluded
by the proposed algorithm, the data point will have very limited impacts on the

overall loss, as compared with state-of-the-art filtering data points based on l
oss values. Extensive empirical results on multiple benchmark datasets have demo
nstrated the robustness of the proposed method under different types of corrupti
on.
****

Context-Agnostic Learning Using Synthetic Data
Charles Jin,Martin Rinard
We propose a novel setting for learning, where the input domain is the image of
a map defined on the product of two sets, one of which completely determines the
 labels. Given the ability to sample from each set independently, we present an
algorithm that learns a classifier over the input domain more efficiently than s
ampling from the input domain directly. We apply this setting to visual classifi
cation tasks, where our approach enables us to train classifiers on datasets tha
t consist entirely of a single example of each class. On several standard benchm
arks for real-world image classification, our approach achieves performance comp
etitive with state-of-the-art results from the few-shot learning and domain tran
sfer literature, while using significantly less data.
****

Neural Bayes: A Generic Parameterization Method for Unsupervised Learning
Devansh Arpit,Huan Wang,Caiming Xiong,richard socher,Yoshua Bengio
We introduce a parameterization method called Neural Bayes which allows computin
g statistical quantities that are in general difficult to compute and opens aven
ues for formulating new objectives for unsupervised representation learning. Spe
cifically, given an observed random variable $\mathbf{x}$ and a latent discrete
variable z, we can express $p(\mathbf{x}|z)$, $p(z|\mathbf{x})$ and $p(z)$ in
closed form in terms of a sufficiently expressive function (Eg. neural network)
using our parameterization without restricting the class of these distributions.
 To demonstrate its usefulness, we develop two independent use cases for this pa
rameterization:

1. Disjoint Manifold Separation: Neural Bayes allows us to formulate an objectiv
e which can optimally label samples from disjoint manifolds present in the suppo
rt of a continuous distribution. This can be seen as a specific form of clusteri
ng where each disjoint manifold in the support is a separate cluster. We design
clustering tasks that obey this formulation and empirically show that the model
optimally labels the disjoint manifolds.

2. Mutual Information Maximization (MIM): MIM has become a popular means for sel
f-supervised representation learning. Neural Bayes allows us to compute mutual i
nformation between observed random variables $\mathbf{x}$ and latent discrete ra
ndom variables z in closed form. We use this for learning image representation
s and show its usefulness on downstream classification tasks.
****

Are Neural Nets Modular? Inspecting Functional Modularity Through Differentiable
 Weight Masks
Róbert Csordás,Sjoerd van Steenkiste,Jürgen Schmidhuber
Neural networks (NNs) whose subnetworks implement reusable functions are expecte
d to offer numerous advantages, including compositionality through efficient rec
ombination of functional building blocks, interpretability, preventing catastrop
hic interference, etc. Understanding if and how NNs are modular could provide in
sights into how to improve them. Current inspection methods, however, fail to li
nk modules to their functionality. In this paper, we present a novel method base
d on learning binary weight masks to identify individual weights and subnets res
ponsible for specific functions. Using this powerful tool, we contribute an exte
nsive study of emerging modularity in NNs that covers several standard architect
ures and datasets. We demonstrate how common NNs fail to reuse submodules and of
fer new insights into the related issue of systematic generalization on language
 tasks.
****

Class Imbalance in Few-Shot Learning

Mateusz Ochal,Massimiliano Patacchiola,Jose Vazquez,Amos Storkey,Sen Wang
Few-shot learning aims to train models on a limited number of labeled samples fr
om a support set in order to generalize to unseen samples from a query set. In t
he standard setup, the support set contains an equal amount of data points for e
ach class. This assumption overlooks many practical considerations arising from
the dynamic nature of the real world, such as class imbalance. In this paper, we
 present a detailed study of few-shot class imbalance along three axes: dataset
vs. support set imbalance, effect of different imbalance distributions (linear,
step, random), and effect of rebalancing techniques. We extensively compare over
 10 state-of-the-art few-shot learning methods using backbones of different dept
hs on multiple datasets. Our analysis reveals that 1) compared to the balanced t
ask, the performances of their class-imbalance counterparts always drop, by up t
o 18.0% for optimization-based methods, although feature-transfer and metric-
based methods generally suffer less, 2) strategies used to mitigate imbalance in
 supervised learning can be adapted to the few-shot case resulting in better per
formances, 3) the effects of imbalance at the dataset level are less significant
 than the effects at the support set level. The code to reproduce the experiment
s is released under an open-source license.
****

Calibration tests beyond classification
David Widmann,Fredrik Lindsten,Dave Zachariah
Most supervised machine learning tasks are subject to irreducible prediction err
ors. Probabilistic predictive models address this limitation by providing probab
ility distributions that represent a belief over plausible targets, rather than
point estimates. Such models can be a valuable tool in decision-making under unc
ertainty, provided that the model output is meaningful and interpretable. Calibr
ated models guarantee that the probabilistic predictions are neither over- nor u
nder-confident. In the machine learning literature, different measures and stati
stical tests have been proposed and studied for evaluating the calibration of cl
assification models. For regression problems, however, research has been focused
 on a weaker condition of calibration based on predicted quantiles for real-valu
ed targets. In this paper, we propose the first framework that unifies calibrati
on evaluation and tests for general probabilistic predictive models. It applies
to any such model, including classification and regression models of arbitrary d
imension. Furthermore, the framework generalizes existing measures and provides
a more intuitive reformulation of a recently proposed framework for calibration
in multi-class classification. In particular, we reformulate and generalize the
kernel calibration error, its estimators, and hypothesis tests using scalar-valu
ed kernels, and evaluate the calibration of real-valued regression
problems.
****

Supervision Accelerates Pre-training in Contrastive Semi-Supervised Learning of
Visual Representations
Mido Assran,Nicolas Ballas,Lluis Castrejon,Michael Rabbat
We investigate a strategy for improving the efficiency of contrastive learning o
f visual representations by leveraging a small amount of supervised information
during pre-training. We propose a semi-supervised loss, SuNCEt, based on noise-c
ontrastive estimation and neighbourhood component analysis, that aims to disting
uish examples of different classes in addition to the self-supervised instance-w
ise pretext tasks. On ImageNet, we find that SuNCEt can be used to match the sem
i-supervised learning accuracy of previous contrastive approaches while using le
ss than half the amount of pre-training and compute. Our main insight is that le
veraging even a small amount of labeled data during pre-training, and not only d
uring fine-tuning, provides an important signal that can significantly accelerat
e contrastive learning of visual representations.
****

Meta Back-Translation
Hieu Pham,Xinyi Wang,Yiming Yang,Graham Neubig
Back-translation is an effective strategy to improve the performance of Neural M
achine Translation~(NMT) by generating pseudo-parallel data. However, several re

cent works have found that better translation quality in the pseudo-parallel dat
a does not necessarily lead to a better final translation model, while lower-qua
lity but diverse data often yields stronger results instead.
In this paper we propose a new way to generate pseudo-parallel data for back-tra
nslation that directly optimizes the final model performance. Specifically, we
propose a meta-learning framework where the back-translation model learns to mat
ch the forward-translation model's gradients on the development data with those
on the pseudo-parallel data. In our evaluations in both the standard datasets WM
T En-De'14 and WMT En-Fr'14, as well as a multilingual translation setting, our
method leads to significant improvements over strong baselines.
****

Amortized Conditional Normalized Maximum Likelihood
Aurick Zhou,Sergey Levine
While deep neural networks provide good performance for a range of challenging t
asks, calibration and uncertainty estimation remain major challenges. In this pa
per, we propose the amortized conditional normalized maximum likelihood (ACNML)
method as a scalable general-purpose approach for uncertainty estimation, calibr
ation, and out-of-distribution robustness with deep networks. Our algorithm buil
ds on the conditional normalized maximum likelihood (CNML) coding scheme, which
has minimax optimal properties according to the minimum description length princ
iple, but is computationally intractable to evaluate exactly for all but the sim
plest of model classes. We propose to use approximate Bayesian inference technqi
ues to produce a tractable approximation to the CNML distribution. Our approach
can be combined with any approximate inference algorithm that provides tractable
 posterior densities over model parameters. We demonstrate that ACNML compares f
avorably to a number of prior techniques for uncertainty estimation in terms of
and calibration on out-of-distribution inputs.
****

One Size Doesn't Fit All: Adaptive Label Smoothing
Ujwal Krothapalli,Lynn Abbott
This paper concerns the use of objectness measures to improve the calibration pe
rformance of Convolutional Neural Networks (CNNs). CNNs have proven to be very g
ood classifiers and generally localize objects well; however, the loss functions
 typically used to train classification CNNs do not penalize inability to locali
ze an object, nor do they take into account an object's relative size in the giv
en image. During training on ImageNet-1K almost all approaches use random crops
on the images and this transformation sometimes provides the CNN with background
 only samples. This causes the classifiers to depend on context. Context depende
nce is harmful for safety-critical applications. We present a novel approach to
classification that combines the ideas of objectness and label smoothing during
training. Unlike previous methods, we compute a smoothing factor that is \emph{a
daptive} based on relative object size within an image. This causes our approach
 to produce confidences that are grounded in the size of the object being classi
fied instead of relying on context to make the correct predictions. We present e
xtensive results using ImageNet to demonstrate that CNNs trained using adaptive
label smoothing are much less likely to be overconfident in their predictions. W
e show qualitative results using class activation maps and quantitative results
using classification and transfer learning tasks. Our approach is able to produc
e an order of magnitude reduction in confidence when predicting on context only
images when compared to baselines. Using transfer learning, we gain 0.021AP on
 MS COCO compared to the hard label approach.
****

ABSTRACTING INFLUENCE PATHS FOR EXPLAINING (CONTEXTUALIZATION OF) BERT MODELS
Kaiji Lu,Zifan Wang,Piotr Mardziel,Anupam Datta
While “attention is all you need” may be proving true, we do not yet know why: a
ttention-based transformer models such as BERT are superior but how they context
ualize information even for simple grammatical rules such as subject-verb number
 agreement(SVA) is uncertain. We introduce multi-partite patterns, abstractions
of sets of paths through a neural network model. Patterns quantify and localize
the effect of an input concept (e.g., a subject’s number) on an output concept (

e.g. corresponding verb’s number) to paths passing through a sequence of model components, thus surfacing how BERT contextualizes information. We describe guided pattern refinement, an efficient search procedure for finding sufficient and sparse patterns representative of concept-critical paths. We discover that patterns generate succinct and meaningful explanations for BERT, highlighted by “copy” and “transfer” operations implemented by skip connections and attention heads, respectively. We also show how pattern visualizations help us understand how BERT contextualizes various grammatical concepts, such as SVA across clauses, and why it makes errors in some cases while succeeding in others.
****
Attentional Constellation Nets for Few-Shot Learning
Weijian Xu,yifan xu,Huaijin Wang,Zhuowen Tu
The success of deep convolutional neural networks builds on top of the learning of effective convolution operations, capturing a hierarchy of structured features via filtering, activation, and pooling. However, the explicit structured features, e.g. object parts, are not expressive in the existing CNN frameworks. In this paper, we tackle the few-shot learning problem and make an effort to enhance structured features by expanding CNNs with a constellation model, which performs cell feature clustering and encoding with a dense part representation; the relationships among the cell features are further modeled by an attention mechanism. With the additional constellation branch to increase the awareness of object parts, our method is able to attain the advantages of the CNNs while making the overall internal representations more robust in the few-shot learning setting. Our approach attains a significant improvement over the existing methods in few-shot learning on the CIFAR-FS, FC100, and mini-ImageNet benchmarks.
****
Unsupervised Learning of Global Factors in Deep Generative Models
Ignacio Peis,Pablo M. Olmos,Antonio Artés
We present a novel deep generative model based on non i.i.d. variational autoencoders that captures global dependencies among observations in a fully unsupervised fashion. In contrast to the recent semi-supervised alternatives for global modeling in deep generative models, our approach combines a mixture model in the local or data-dependent space and a global Gaussian latent variable, which lead us to obtain three particular insights. First, the induced latent global space captures interpretable disentangled representations with no user-defined regularization in the evidence lower bound (as in beta-VAE and its generalizations). Second, we show that the model performs domain alignment to find correlations and interpolate between different databases. Finally, we study the ability of the global space to discriminate between groups of observations with non-trivial underlying structures, such as face images with shared attributes or defined sequences of digits images.
****
On Dynamic Noise Influence in Differential Private Learning
Junyuan Hong,Zhangyang Wang,Jiayu Zhou
Protecting privacy in learning while maintaining the model performance has become increasingly critical in many applications that involve sensitive data. Private Gradient Descent (PGD) is a commonly used private learning framework, which adds noise according to the Differential Privacy protocol.Recent studies show that dynamic privacy schedules of decreasing noise magnitudes can improve loss at the final iteration, and yet theoretical understandings of the effectiveness of such schedules and their connections to optimization algorithms remain limited. In this paper, we provide comprehensive analysis of noise influence in dynamic privacy schedules to answer these critical questions. We first present a dynamic noise schedule minimizing the utility upper bound of PGD, and show how the noise influence from each optimization step collectively impacts utility of the final model. Our study also reveals how impacts from dynamic noise influence change when momentum is used. We empirically show the connection exists for general non-convex losses, and the influence is greatly impacted by the loss curvature.
****
Momentum Contrastive Autoencoder

Devansh Arpit,Aadyot Bhatnagar,Huan Wang,Caiming Xiong
Wasserstein autoencoder (WAE) shows that matching two distributions is equivalen t to minimizing a simple autoencoder (AE) loss under the constraint that the lat ent space of this AE matches a pre-specified prior distribution. This latent spa ce distribution matching is a core component in WAE, and is in itself a challeng ing task. In this paper, we propose to use the contrastive learning framework th at has been shown to be effective for self-supervised representation learning, a s a means to resolve this problem. We do so by exploiting the fact that contrast ive learning objectives optimize the latent space distribution to be uniform ove r the unit hyper-sphere, which can be easily sampled from. This results in a sim ple and scalable algorithm that avoids many of the optimization challenges of ex isting generative models, while retaining the advantage of efficient sampling. Q uantitatively, we show that our algorithm achieves a new state-of-the-art FID of 54.36 on CIFAR-10, and performs competitively with existing models on CelebA in terms of FID score. We also show qualitative results on CelebA-HQ in addition t o these datasets, confirming that our algorithm can generate realistic images at multiple resolutions.
****
Uncertainty Estimation in Autoregressive Structured Prediction
Andrey Malinin,Mark Gales
Uncertainty estimation is important for ensuring safety and robustness of AI sys tems. While most research in the area has focused on un-structured prediction t asks, limited work has investigated general uncertainty estimation approaches fo r structured prediction. Thus, this work aims to investigate uncertainty estimat ion for structured prediction tasks within a single unified and interpretable pr obabilistic ensemble-based framework. We consider: uncertainty estimation for s equence data at the token-level and complete sequence-level; interpretations for , and applications of, various measures of uncertainty; and discuss both the the oretical and practical challenges associated with obtaining them. This work also provides baselines for token-level and sequence-level error detection, and sequ ence-level out-of-domain input detection on the WMT’14 English-French and WMT’17 English-German translation and LibriSpeech speech recognition datasets.
****
Regression Prior Networks
Andrey Malinin,Sergey Chervontsev,Ivan Provilkov,Mark Gales
Prior Networks are a recently developed class of models which yield interpretabl e measures of uncertainty and have been shown to outperform state-of-the-art ens emble approaches on a range of tasks. They can also be used to distill an ensemb le of models via \emph{Ensemble Distribution Distillation} (EnD^2), such that its accuracy, calibration and uncertainty estimates are retained within a single model. However, Prior Networks have so far been developed only for classificati on tasks. This work extends Prior Networks and EnD^2 to regression tasks by co nsidering the Normal-Wishart distribution. The properties of Regression Prior Ne tworks are demonstrated on synthetic data, selected UCI datasets and a monocular depth estimation task, where they yield performance competitive with ensemble a pproaches.
****
Benefits of Assistance over Reward Learning
Rohin Shah,Pedro Freire,Neel Alex,Rachel Freedman,Dmitrii Krasheninnikov,Lawrenc e Chan,Michael D Dennis,Pieter Abbeel,Anca Dragan,Stuart Russell
Much recent work has focused on how an agent can learn what to do from human fee dback, leading to two major paradigms. The first paradigm is reward learning, in which the agent learns a reward model through human feedback that is provided e xternally from the environment. The second is assistance, in which the human is modeled as a part of the environment, and the true reward function is modeled as a latent variable in the environment that the agent may make inferences about. The key difference between the two paradigms is that in the reward learning para digm, by construction there is a separation between reward learning and control using the learned reward. In contrast, in assistance these functions are perform ed as needed by a single policy. By merging reward learning and control, assisti

ve agents can reason about the impact of control actions on reward learning, lea
ding to several advantages over agents based on reward learning. We illustrate t
hese advantages in simple environments by showing desirable qualitative behavior
s of assistive agents that cannot be found by agents based on reward learning.
****
Measuring Massive Multitask Language Understanding
Dan Hendrycks,Collin Burns,Steven Basart,Andy Zou,Mantas Mazeika,Dawn Song,Jacob
 Steinhardt
We propose a new test to measure a text model's multitask accuracy. The test cov
ers 57 tasks including elementary mathematics, US history, computer science, law
, and more. To attain high accuracy on this test, models must possess extensive
world knowledge and problem solving ability. We find that while most recent mode
ls have near random-chance accuracy, the very largest GPT-3 model improves over
random chance by almost 20 percentage points on average. However, on every one o
f the 57 tasks, the best models still need substantial improvements before they
can reach expert-level accuracy. Models also have lopsided performance and frequ
ently do not know when they are wrong. Worse, they still have near-random accura
cy on some socially important subjects such as morality and law. By comprehensiv
ely evaluating the breadth and depth of a model's academic and professional unde
rstanding, our test can be used to analyze models across many tasks and to ident
ify important shortcomings.
****
No MCMC for me: Amortized sampling for fast and stable training of energy-based
models
Will Sussman Grathwohl,Jacob Jin Kelly,Milad Hashemi,Mohammad Norouzi,Kevin Swer
sky,David Duvenaud
Energy-Based Models (EBMs) present a flexible and appealing way to represent unc
ertainty. Despite recent advances, training EBMs on high-dimensional data remain
s a challenging problem as the state-of-the-art approaches are costly, unstable,
 and require considerable tuning and domain expertise to apply successfully. In
this work, we present a simple method for training EBMs at scale which uses an e
ntropy-regularized generator to amortize the MCMC sampling typically used in EBM
 training. We improve upon prior MCMC-based entropy regularization methods with
a fast variational approximation. We demonstrate the effectiveness of our approa
ch by using it to train tractable likelihood models. Next, we apply our estimato
r to the recently proposed Joint Energy Model (JEM), where we match the original
 performance with faster and stable training. This allows us to extend JEM model
s to semi-supervised classification on tabular data from a variety of continuous
 domains.
****
Communication-Efficient Sampling for Distributed Training of Graph Convolutional
 Networks
Peng Jiang,Masuma Akter Rumi
Training Graph Convolutional Networks (GCNs) is expensive as it needs to aggrega
te data recursively from neighboring nodes. To reduce the computation overhead,
previous works have proposed various neighbor sampling methods that estimate the
 aggregation result based on a small number of sampled neighbors. Although these
 methods have successfully accelerated the training, they mainly focus on the si
ngle-machine setting. As real-world graphs are large, training GCNs in distribut
ed systems is desirable. However, we found that the existing neighbor sampling m
ethods do not work well in a distributed setting. Specifically, a naive implemen
tation may incur a huge amount of communication of feature vectors among differe
nt machines. To address this problem, we propose a communication-efficient neigh
bor sampling method in this work. Our main idea is to assign higher sampling pro
babilities to the local nodes so that remote nodes are accessed less frequently.
 We present an algorithm that determines the local sampling probabilities and ma
kes sure our skewed neighbor sampling does not affect much to the convergence of
 the training. Our experiments with node classification benchmarks show that our
 method significantly reduces the communication overhead for distributed GCN tra
ining with little accuracy loss.

****
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 fundamentally incorrect. We demonstrate that anyone with access to the classifier, even without access to any original training data or trigger, can construct several alternative triggers that are as effective or more so at eliciting the target class at test time. We construct these alternative triggers by first generating adversarial examples for a smoothed version of the classifier, created with a recent process called Denoised Smoothing, and then extracting colors or cropped portions of adversarial images. We demonstrate the effectiveness of our attack through extensive experiments on ImageNet and TrojAI datasets, including a user study which demonstrates that our method allows users to easily determine the existence of such backdoors in existing poisoned classifiers. Furthermore, we demonstrate that our alternative triggers can in fact look entirely different from the original trigger, highlighting that the backdoor actually learned by the classifier differs substantially from the trigger image itself. 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.
****
Bayesian Meta-Learning for Few-Shot 3D Shape Completion
Masanori Koyama,Toshiki Nakanishi,Shin-ichi Maeda,Vitor Campagnolo Guizilini,Adrien Gaidon
Estimating the 3D shape of real-world objects is a key perceptual challenge. It requires going from partial observations, which are often too sparse and incomprehensible for the human eye, to detailed shape representations that vary significantly across categories and instances. We propose to cast shape completion as a Bayesian meta-learning problem to facilitate the transfer of knowledge learned from observing one object into estimating the shape of another object. To combine the Bayesian framework with an approach that uses implicit 3D object representation, we introduce an encoder that describes the posterior distribution of a latent representation conditioned on sparse point clouds. With its ability to isolate object-specific properties from object-agnostic properties, our meta-learning algorithm enables accurate shape completion of newly-encountered objects from sparse observations. We demonstrate the efficacy of our proposed method with experimental results on the standard ShapeNet and ICL-NUIM benchmarks.
****
A Distributional Approach to Controlled Text Generation
Muhammad Khalifa,Hady Elsahar,Marc Dymetman
We propose a Distributional Approach for addressing Controlled Text Generation from pre-trained Language Models (LM). This approach permits to specify, in a single formal framework, both “pointwise’” and “distributional” constraints over the target LM — to our knowledge, the first model with such generality —while minimizing KL divergence from the initial LM distribution. The optimal target distribution is then uniquely determined as an explicit EBM (Energy-BasedModel) representation. From that optimal representation, we then train a target controlled Autoregressive LM through an adaptive distributional variant of PolicyGradient. We conduct a first set of experiments over pointwise constraints showing the advantages of our approach over a set of baselines, in terms of obtaining a controlled LM balancing constraint satisfaction with divergence from the pretrained LM. We then perform experiments over distributional constraints, a unique feature of our approach, demonstrating its potential as a remedy to the problem of Bias in Language Models. Through an ablation study, we show the effectiveness o

f our adaptive technique for obtaining faster convergence.
Code available at https://github.com/naver/gdc
****

Aligning AI With Shared Human Values
Dan Hendrycks,Collin Burns,Steven Basart,Andrew Critch,Jerry Li,Dawn Song,Jacob Steinhardt
We show how to assess a language model's knowledge of basic concepts of morality. We introduce the ETHICS dataset, a new benchmark that spans concepts in justice, well-being, duties, virtues, and commonsense morality. Models predict widespread moral judgments about diverse text scenarios. This requires connecting physical and social world knowledge to value judgements, a capability that may enable us to steer chatbot outputs or eventually regularize open-ended reinforcement learning agents. With the ETHICS dataset, we find that current language models have a promising but incomplete ability to predict basic human ethical judgements. Our work shows that progress can be made on machine ethics today, and it provides a steppingstone toward AI that is aligned with human values.
****

Convolutional Neural Networks are not invariant to translation, but they can learn to be
Valerio Biscione,Jeffrey Bowers
When seeing a new object, humans can immediately recognize it across different retinal locations: we say that the internal object representation is invariant to translation. It is commonly believed that Convolutional Neural Networks (CNNs) are architecturally invariant to translation thanks to the convolution and/or pooling operations they are endowed with. In fact, several works have found that these networks systematically fail to recognise new objects on untrained locations. In this work we show how, even though CNNs are not 'architecturally invariant' to translation, they can indeed 'learn' to be invariant to translation. We verified that this can be achieved by pretraining on ImageNet, and we found that it is also possible with much simpler datasets in which the items are fully translated across the input canvas. Significantly, simply training everywhere on the canvas was not enough. We investigated how this pretraining affected the internal network representations, finding that the invariance was almost always acquired, even though it was some times disrupted by further training due to catastrophic forgetting/interference.
These experiments show how pretraining a network on an environment with the right 'latent' characteristics (a more naturalistic environment) can result in the network learning deep perceptual rules which would dramatically improve subsequent generalization.
****

Self-Activating Neural Ensembles for Continual Reinforcement Learning
Sam Powers,Abhinav Gupta
The ability for an agent to continuously learn new skills without catastrophically forgetting existing knowledge is of critical importance for the development of generally intelligent agents. Most methods devised to address this problem depend heavily on well-defined task boundaries which simplify the problem considerably. Our task-agnostic method, Self-Activating Neural Ensembles (SANE), uses a hierarchical modular architecture designed to avoid catastrophic forgetting without making any such assumptions. At each timestep a path through the SANE tree is activated; during training only activated nodes are updated, ensuring that unused nodes do not undergo catastrophic forgetting. Additionally, new nodes are created as needed, allowing the system to leverage and retain old skills while growing and learning new ones. We demonstrate our approach on MNIST and a set of grid world environments, demonstrating that SANE does not undergo catastrophic forgetting where existing methods do.
****

Learning Axioms to Compute Verifiable Symbolic Expression Equivalence Proofs Using Graph-to-Sequence Networks
Steven James Kommrusch,Louis-Noel Pouchet,Theo Barolett
We target the problem of proving the semantic equivalence between two complex ex

pressions represented as typed trees, and demonstrate our system on expressions from a rich multi-type symbolic language for linear algebra. We propose the firs t graph-to-sequence deep learning system to generate axiomatic proofs of equival ence between program pairs. We generate expressions which include scalars, vecto rs and matrices and 16 distinct operators combining them, with 147 distinct axio ms of equivalence. We study the robustness of the system to generate proofs of i ncreasing length, demonstrating how incremental graph-to-sequence networks can l earn to represent complex and verifiable symbolic reasoning. It achieves 93% ave rage true positive coverage on 10,000 test cases while ensuring zero false posit ives by design.
****

Higher-order Structure Prediction in Evolving Graph Simplicial Complexes
Manohar Kaul,Masaaki Imaizumi
Dynamic graphs are rife with higher-order interactions, such as co-authorship re lationships and protein-protein interactions in biological networks, that natura lly arise between more than two nodes at once. In spite of the ubiquitous presen ce of such higher-order interactions, limited attention has been paid to the hig her-order counterpart of the popular pairwise link prediction problem. Existing higher-order structure prediction methods are mostly based on heuristic feature extraction procedures, which work well in practice but lack theoretical guarante es. Such heuristics are primarily focused on predicting links in a static snapsh ot of the graph. Moreover, these heuristic-based methods fail to effectively uti lize and benefit from the knowledge of latent substructures already present with in the higher-order structures. In this paper, we overcome these obstacles by ca pturing higher-order interactions succinctly as simplices, model their neighborh ood by face-vectors, and develop a nonparametric kernel estimator for simplices that views the evolving graph from the perspective of a time process (i.e., a se quence of graph snapshots). Our method substantially outperforms several baselin e higher-order prediction methods. As a theoretical achievement, we prove the co nsistency and asymptotic normality in terms of Wasserstein distance of our estim ator using Stein's method.
****

Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking
Michael Sejr Schlichtkrull,Nicola De Cao,Ivan Titov
Graph neural networks (GNNs) have become a popular approach to integrating struc tural inductive biases into NLP models. However, there has been little work on i nterpreting them, and specifically on understanding which parts of the graphs (e .g. syntactic trees or co-reference structures) contribute to a prediction. In t his work, we introduce a post-hoc method for interpreting the predictions of GNN s which identifies unnecessary edges. Given a trained GNN model, we learn a simp le classifier that, for every edge in every layer, predicts if that edge can be dropped. We demonstrate that such a classifier can be trained in a fully differe ntiable fashion, employing stochastic gates and encouraging sparsity through the expected L_0 norm. We use our technique as an attribution method to analyze G NN models for two tasks -- question answering and semantic role labeling -- prov iding insights into the information flow in these models. We show that we can dr op a large proportion of edges without deteriorating the performance of the mode l, while we can analyse the remaining edges for interpreting model predictions.
****

Multi-hop Attention Graph Neural Network
Guangtao Wang,Zhitao Ying,Jing Huang,Jure Leskovec
Self-attention mechanism in graph neural networks (GNNs) led to state-of-the-art performance on many graph representation learning task. Currently, at every lay er, attention is computed between connected pairs of nodes and depends solely on the representation of the two nodes. However, such attention mechanism does not account for nodes that are not directly connected but provide important network context, which could lead to improved predictive performance. Here we propose M ulti-hop Attention Graph Neural Network (MAGNA), a principled way to incorporate multi-hop context information into attention computation, enabling long-range i nteractions at every layer of the GNN. To compute attention between nodes that

are not directly connected,
MAGNA diffuses the attention scores across the network, which increases the ''receptive field'' for every layer of the GNN.
Unlike previous approaches, MAGNA uses a diffusion prior on attention values, to efficiently account for all paths between the pair of disconnected nodes. This helps MAGNA capture large-scale structural information in every layer, and learn more informative attention. Experimental results on node classification as well as the knowledge graph completion benchmarks show that MAGNA achieves state-of-the-art results: MAGNA achieves up to 5.7% relative error reduction over the previous state-of-the-art on Cora, Citeseer, and Pubmed. MAGNA also obtains the best performance on a large-scale Open Graph Benchmark dataset. On knowledge graph completion MAGNA advances state-of-the-art on WN18RR and FB15k-237 across four different performance metrics.
****
Deep Ensembles with Hierarchical Diversity Pruning
Yanzhao Wu,Ling Liu
Diverse deep ensembles hold the potential for improving accuracy and robustness of deep learning models. Both pairwise and non-pairwise ensemble diversity metrics have been proposed over the past two decades. However, it is also challenging to find the right metrics that can effectively prune those deep ensembles with insufficient ensemble diversity, thus failing to deliver effective ensemble accuracy. In this paper, we first compare six popular diversity metrics in the literature, coined as Q metrics, including both pairwise and non-pairwise representatives. We analyze their inherent limitations in capturing the negative correlation of ensemble member models, and thus inefficient in identifying and pruning low quality ensembles. We next present six HQ ensemble diversity metrics by extending the existing Q-metrics with three novel optimizations: (1) We introduce the concept of focal model and separately measure the ensemble diversity among the deep ensembles of the same team size with the concept of focal model, aiming to better capture the negative correlations of member models of an ensemble. (2) We introduce six HQ-diversity metrics to optimize the corresponding Q-metrics respectively in terms of measuring negative correlation among member models of an ensemble using its ensemble diversity score. (3) We introduce a two phase hierarchical pruning method to effectively identify and prune those deep ensembles with high HQ diversity scores, aiming to increase the lower and upper bounds on ensemble accuracy for the selected ensembles. By combining these three optimizations, deep ensembles selected based on our hierarchical diversity pruning approach significantly outperforms those selected by the corresponding Q-metrics. Comprehensive experimental evaluation over several benchmark datasets shows that our HQ-metrics can effectively select high diversity deep ensembles by pruning out those ensembles with insufficient diversity, and successfully increase the lower bound (worst case) accuracy of the selected deep ensembles, compared to those selected using the state-of-the-art Q-metrics.
****
Near-Optimal Linear Regression under Distribution Shift
Qi Lei,Wei Hu,Jason D. Lee
Transfer learning is an essential technique when sufficient data comes from the source domain, while no or scarce labeled data is from the target domain. We develop estimators that achieve minimax linear risk for linear regression problems under the distribution shift. Our algorithms cover different kinds of settings with covariate shift or model shift. We also consider when data are generating from either linear or general nonlinear models. We show that affine minimax rules are within an absolute constant of the minimax risk even among nonlinear rules, for a variety of source/target distributions.
****
Semantic Hashing with Locality Sensitive Embeddings
Levi Boyles,Aniket Anand Deshmukh,Urun Dogan,Rajesh Koduru,Charles Denis,Eren Manavoglu
Semantic hashing methods have been explored for learning transformations into binary vector spaces. These learned binary representations may then be used in has

hing based retrieval methods, typically by retrieving all neighboring elements i
n the Hamming ball with radius 1 or 2. Prior studies focus on tasks with a few
 dozen to a few hundred semantic categories at most, and it is not currently well
 known how these methods scale to domains with richer semantic structure. In th
is study, we focus on learning embeddings for the use in exact hashing retrieval
, where Approximate Nearest Neighbor search comprises of a simple table lookup.
We propose similarity learning methods in which the optimized similarity is the
 angular similarity (the probability of collision under SimHash.) We demonstrate
 the benefits of these embeddings on a variety of domains, including a coocurren
ce modelling task on a large scale text corpus; a rich structure of which cannot
 be handled by a few hundred semantic groups.
****

Class Normalization for (Continual)? Generalized Zero-Shot Learning
Ivan Skorokhodov,Mohamed Elhoseiny
Normalization techniques have proved to be a crucial ingredient of successful tr
aining in a traditional supervised learning regime. However, in the zero-shot le
arning (ZSL) world, these ideas have received only marginal attention. This work
 studies normalization in ZSL scenario from both theoretical and practical persp
ectives. First, we give a theoretical explanation to two popular tricks used in
zero-shot learning: normalize+scale and attributes normalization and show that t
hey help training by preserving variance during a forward pass. Next, we demonst
rate that they are insufficient to normalize a deep ZSL model and propose Class
Normalization (CN): a normalization scheme, which alleviates this issue both pro
vably and in practice. Third, we show that ZSL models typically have more irregu
lar loss surface compared to traditional classifiers and that the proposed metho
d partially remedies this problem. Then, we test our approach on 4 standard ZSL
datasets and outperform sophisticated modern SotA with a simple MLP optimized wi
thout any bells and whistles and having ~50 times faster training speed. Finally
, we generalize ZSL to a broader problem — continual ZSL, and introduce some pri
ncipled metrics and rigorous baselines for this new setup. The source code is av
ailable at https://github.com/universome/class-norm.
****

Aspect-based Sentiment Classification via Reinforcement Learning
Lichen Wang,Bo Zong,Yunyu Liu,Can Qin,Wei Cheng,Wenchao Yu,Xuchao Zhang,Haifeng
Chen,Yun Fu
Aspect-based sentiment classification aims to predict sentimental polarities of
one or multiple aspects in texts. As texts always contain a large proportion of
task-irrelevant words, accurate alignment between aspects and their sentimental
descriptions is the most crucial and challenging step. State-of-the-art approach
es are mainly based on word-level attention learned from recurrent neural networ
k variants (e.g., LSTM) or graph neural networks. From another view, these metho
ds essentially weight and aggregate all possible alignments. However, this mecha
nism heavily relies on large-scale supervision training: without enough labels,
it could easily overfit with difficulty in generalization. To address this chall
enge, we propose SentRL, a reinforcement learning-based framework for aspect-bas
ed sentiment classification. In this framework, input texts are transformed into
 their dependency graphs. Then, an agent is deployed to walk on the graphs, expl
ores paths from target aspect nodes to their potential sentimental regions, and
differentiates the effectiveness of different paths. By limiting the agent's exp
loration budget, our method encourages the agent to skip task-irrelevant informa
tion and focus on the most effective paths for alignment purpose. Our method con
siderably reduces the impact of task-irrelevant words and improves generalizatio
n performance. Compared with competitive baseline methods, our approach achieves
 the highest performance on public benchmark datasets with up to 3.7% improve
ment.
****

Improved Communication Lower Bounds for Distributed Optimisation
Janne H. Korhonen,Dan Alistarh
Motivated by the interest in communication-efficient methods for distributed mac
hine learning, we consider the communication complexity of minimising a sum of $

d$-dimensional functions $\sum_{i = 1}^N f_i (x)$, where each function f_i is held by one of the N different machines. Such tasks arise naturally in large-scale optimisation, where a standard solution is to apply variants of (stochastic) gradient descent. As our main result, we show that $\Omega(Nd \log d / \varepsilon)$ bits in total need to be communicated between the machines to find an additive ϵ-approximation to the minimum of $\sum_{i = 1}^N f_i (x)$. The results holds for deterministic algorithms, and randomised algorithms under some restrictions on the parameter values. Importantly, our lower bounds require no assumptions on the structure of the algorithm, and are matched within constant factors for strongly convex objectives by a new variant of quantised gradient descent. The lower bounds are obtained by bringing over tools from communication complexity to distributed optimisation, an approach we hope will find further use in future.

****

Adaptive Self-training for Neural Sequence Labeling with Few Labels
Yaqing Wang,Subhabrata Mukherjee,Haoda Chu,Yuancheng Tu,Ming Wu,Jing Gao,Ahmed Hassan Awadallah
Neural sequence labeling is an important technique employed for many Natural Language Processing (NLP) tasks, such as Named Entity Recognition (NER), slot tagging for dialog systems and semantic parsing. Large-scale pre-trained language models obtain very good performance on these tasks when fine-tuned on large amounts of task-specific labeled data. However, such large-scale labeled datasets are difficult to obtain for several tasks and domains due to the high cost of human annotation as well as privacy and data access constraints for sensitive user applications. This is exacerbated for sequence labeling tasks requiring such annotations at token-level. In this work, we develop techniques to address the label scarcity challenge for neural sequence labeling models. Specifically, we develop self-training and meta-learning techniques for training neural sequence taggers with few labels. While self-training serves as an effective mechanism to learn from large amounts of unlabeled data -- meta-learning helps in adaptive sample re-weighting to mitigate error propagation from noisy pseudo-labels. Extensive experiments on six benchmark datasets including two for massive multilingual NER and four slot tagging datasets for task-oriented dialog systems demonstrate the effectiveness of our method. With only 10 labeled examples for each class for each task, our method obtains 10% improvement over state-of-the-art systems demonstrating its effectiveness for the low-resource setting.
****

Learning explanations that are hard to vary
Giambattista Parascandolo,Alexander Neitz,ANTONIO ORVIETO,Luigi Gresele,Bernhard Schölkopf
In this paper, we investigate the principle that good explanations are hard to vary in the context of deep learning.
We show that averaging gradients across examples -- akin to a logical OR of patterns -- can favor memorization and `patchwork' solutions that sew together different strategies, instead of identifying invariances.
To inspect this, we first formalize a notion of consistency for minima of the loss surface, which measures to what extent a minimum appears only when examples are pooled.
We then propose and experimentally validate a simple alternative algorithm based on a logical AND, that focuses on invariances and prevents memorization in a set of real-world tasks.
Finally, using a synthetic dataset with a clear distinction between invariant and spurious mechanisms, we dissect learning signals and compare this approach to well-established regularizers.
****

Transformers for Modeling Physical Systems
Nicholas Geneva,Nicholas Zabaras
Transformers are widely used in neural language processing due to their ability to model longer-term dependencies in text. Although these models achieve state-o

f-the-art performance for many language related tasks, their applicability outsi
de of the neural language processing field has been minimal. In this work, we pr
opose the use of transformer models for the prediction of dynamical systems repr
esentative of physical phenomena. The use of Koopman based embeddings provide a
unique and powerful method for projecting any dynamical system into a vector rep
resentation which can then be predicted by a transformer model. The proposed mod
el is able to accurately predict various dynamical systems and outperform classi
cal methods that are commonly used in the scientific machine learning literature
.
****
Degree-Quant: Quantization-Aware Training for Graph Neural Networks
Shyam Anil Tailor,Javier Fernandez-Marques,Nicholas Donald Lane
Graph neural networks (GNNs) have demonstrated strong performance on a wide vari
ety of tasks due to their ability to model non-uniform structured data. Despite
their promise, there exists little research exploring methods to make them more
efficient at inference time. In this work, we explore the viability of training
quantized GNNs, enabling the usage of low precision integer arithmetic during in
ference. For GNNs seemingly unimportant choices in quantization implementation c
ause dramatic changes in performance. We identify the sources of error that uniq
uely arise when attempting to quantize GNNs, and propose an architecturally-agno
stic and stable method, Degree-Quant, to improve performance over existing quant
ization-aware training baselines commonly used on other architectures, such as C
NNs. We validate our method on six datasets and show, unlike previous quantizati
on attempts, that models generalize to unseen graphs. Models trained with Degree
-Quant for INT8 quantization perform as well as FP32 models in most cases; for I
NT4 models, we obtain up to 26% gains over the baselines. Our work enables up to
 4.7x speedups on CPU when using INT8 arithmetic.
****
Computing Preimages of Deep Neural Networks with Applications to Safety
Kyle Matoba,François Fleuret
To apply an algorithm in a sensitive domain it is important to understand the se
t of input values that result in specific decisions. Deep neural networks suffer
 from an inherent instability that makes this difficult: different outputs can a
rise from very similar inputs.

We present a method to check that the decisions of a deep neural network are as
intended by constructing the exact, analytical preimage of its predictions. Prei
mages generalize verification in the sense that they can be used to verify a wid
e class of properties, and answer much richer questions besides. We examine the
functioning and failures of neural networks used in robotics, including an aircr
aft collision avoidance system, related to sequential decision making and extrap
olation.

Our method iterates backwards through the layers of piecewise linear deep neural
 networks. Uniquely, we compute \emph{all} intermediate values that correspond t
o a prediction, propagating this calculation through layers using analytical for
mulae for layer preimages.

****
Regularization Matters in Policy Optimization - An Empirical Study on Continuous
 Control
Zhuang Liu,Xuanlin Li,Bingyi Kang,Trevor Darrell
Deep Reinforcement Learning (Deep RL) has been receiving increasingly more atten
tion thanks to its encouraging performance on a variety of control tasks. Yet,
conventional regularization techniques in training neural networks (e.g., L_2
regularization, dropout) have been largely ignored in RL methods, possibly becau
se agents are typically trained and evaluated in the same environment, and becau
se the deep RL community focuses more on high-level algorithm designs. In this w
ork, we present the first comprehensive study of regularization techniques with

multiple policy optimization algorithms on continuous control tasks. Interesting ly, we find conventional regularization techniques on the policy networks can of ten bring large improvement, especially on harder tasks. Our findings are shown to be robust against training hyperparameter variations. We also compare these t echniques with the more widely used entropy regularization. In addition, we stud y regularizing different components and find that only regularizing the policy n etwork is typically the best. We further analyze why regularization may help gen eralization in RL from four perspectives - sample complexity, reward distributio n, weight norm, and noise robustness. We hope our study provides guidance for fu ture practices in regularizing policy optimization algorithms. Our code is avail able at https://github.com/xuanlinli17/iclr2021_rlreg .

****

Recurrent Independent Mechanisms

Anirudh Goyal,Alex Lamb,Jordan Hoffmann,Shagun Sodhani,Sergey Levine,Yoshua Beng io,Bernhard Schölkopf

We explore the hypothesis that learning modular structures which reflect the dyn amics of the environment can lead to better generalization and robustness to cha nges that only affect a few of the underlying causes. We propose Recurrent Indep endent Mechanisms (RIMs), a new recurrent architecture in which multiple groups of recurrent cells operate with nearly independent transition dynamics, communic ate only sparingly through the bottleneck of attention, and compete with each ot her so they are updated only at time steps where they are most relevant. We sho w that this leads to specialization amongst the RIMs, which in turn allows for r emarkably improved generalization on tasks where some factors of variation diffe r systematically between training and evaluation.

****

Linear Convergence and Implicit Regularization of Generalized Mirror Descent wit h Time-Dependent Mirrors

Adityanarayanan Radhakrishnan,Mikhail Belkin,Caroline Uhler

The following questions are fundamental to understanding the properties of over -parameterization in modern machine learning: (1) Under what conditions and at w hat rate does training converge to a global minimum? (2) What form of implicit r egularization occurs through training? While significant progress has been made in answering both of these questions for gradient descent, they have yet to be a nswered more completely for general optimization methods. In this work, we estab lish sufficient conditions for linear convergence and obtain approximate implici t regularization results for generalized mirror descent (GMD), a generalization of mirror descent with a possibly time-dependent mirror. GMD subsumes popular fi rst order optimization methods including gradient descent, mirror descent, and p reconditioned gradient descent methods such as Adagrad. By using the Polyak-Loja siewicz inequality, we first present a simple analysis under which non-stochasti c GMD converges linearly to a global minimum. We then present a novel, Taylor-se ries based analysis to establish sufficient conditions for linear convergence of stochastic GMD. As a corollary, our result establishes sufficient conditions an d provides learning rates for linear convergence of stochastic mirror descent an d Adagrad. Lastly, we obtain approximate implicit regularization results for GMD by proving that GMD converges to an interpolating solution that is approximatel y the closest interpolating solution to the initialization in l2-norm in the dua l space, thereby generalizing the result of Azizan, Lale, and Hassibi (2019) in the full batch setting.

****

Byzantine-Resilient Non-Convex Stochastic Gradient Descent

Zeyuan Allen-Zhu,Faeze Ebrahimianghazani,Jerry Li,Dan Alistarh

We study adversary-resilient stochastic distributed optimization, in which m m achines can independently compute stochastic gradients, and cooperate to jointly optimize over their local objective functions. However, an α-fraction of the machines are Byzantine, in that they may behave in arbitrary, adversarial w ays. We consider a variant of this procedure in the challenging non-convex case. Our main result is a new algorithm SafeguardSGD, which can provably escape sadd

le points and find approximate local minima of the non-convex objective. The alg
orithm is based on a new concentration filtering technique, and its sample and t
ime complexity bounds match the best known theoretical bounds in the stochastic,
 distributed setting when no Byzantine machines are present.

Our algorithm is very practical: it improves upon the performance of all prior m
ethods when training deep neural networks, it is relatively lightweight, and it
is the first method to withstand two recently-proposed Byzantine attacks.
****
Learning Active Learning in the Batch-Mode Setup with Ensembles of Active Learni
ng Agents
Malte Ebner,Bernhard Kratzwald,Stefan Feuerriegel
Supervised learning models perform best when trained on a lot of data, but annot
ating training data is very costly in some domains. Active learning aims to chos
e only the most informative subset of unlabelled samples for annotation, thus sa
ving annotation cost. Several heuristics for choosing this subset have been deve
loped, which use fix policies for this choice. They are easily understandable an
d applied. However, there is no heuristic performing optimal in all settings. Th
is lead to the development of agents learning the best selection policy from dat
a. They formulate active learning as a Markov decision process and applying rein
forcement learning (RL) methods to it. Their advantage is that they are able to
use many features and to adapt to the specific task.

Our paper proposes a new approach combining these advantages of learning active
learning and heuristics: We propose to learn active learning using a parametrise
d ensemble of agents, where the parameters are learned using Monte Carlo policy
search. As this approach can incorporate any active learning agent into its ense
mble, it allows to increase the performance of every active learning agent by le
arning how to combine it with others.
****
Divide-and-Conquer Monte Carlo Tree Search
Giambattista Parascandolo,Lars Holger Buesing,Josh Merel,Leonard Hasenclever,Joh
n Aslanides,Jessica B Hamrick,Nicolas Heess,Alexander Neitz,Theophane Weber
Standard planners for sequential decision making (including Monte Carlo planning
, tree search, dynamic programming, etc.) are constrained by an implicit sequent
ial planning assumption: The order in which a plan is constructed is the same in
 which it is executed.
We consider alternatives to this assumption for the class of goal-directed Reinf
orcement Learning (RL) problems.
Instead of an environment transition model, we assume an imperfect, goal-directe
d policy.
This low-level policy can be improved by a plan, consisting of an appropriate se
quence of sub-goals that guide it from the start to the goal state. We propose a
 planning algorithm, Divide-and-Conquer Monte Carlo Tree Search (DC-MCTS), for a
pproximating the optimal plan by means of proposing intermediate sub-goals which
 hierarchically partition the initial tasks into simpler ones that are then solv
ed independently and recursively. The algorithm critically makes use of a learne
d sub-goal proposal for finding appropriate partitions trees of new tasks based
on prior experience.
Different strategies for learning sub-goal proposals give rise to different plan
ning strategies that strictly generalize sequential planning.
We show that this algorithmic flexibility over planning order leads to improved
results in navigation tasks in grid-worlds as well as in challenging continuous
control environments.
****
PettingZoo: Gym for Multi-Agent Reinforcement Learning
J K Terry,Benjamin Black,Mario Jayakumar,Ananth Hari,Luis Santos,Clemens Dieffen
dahl,Niall L Williams,Yashas Lokesh,Ryan Sullivan,Caroline Horsch,Praveen Ravi
OpenAI's Gym library contains a large, diverse set of environments that are use
ful benchmarks in reinforcement learning, under a single elegant Python API (wit

h tools to develop new compliant environments) . The introduction of this librar y has proven a watershed moment for the reinforcement learning community, becaus e it created an accessible set of benchmark environments that everyone could use (including wrapper important existing libraries), and because a standardized AP I let RL learning methods and environments from anywhere be trivially exchanged. This paper similarly introduces PettingZoo, a library of diverse set of multi-a gent environments under a single elegant Python API, with tools to easily make n ew compliant environments.
****

Trans-Caps: Transformer Capsule Networks with Self-attention Routing
Aryan Mobiny,Pietro Antonio Cicalese,Hien Van Nguyen
Capsule Networks (CapsNets) have shown to be a promising alternative to Convolut ional Neural Networks (CNNs) in many computer vision tasks, due to their ability to encode object viewpoint variations. The high computational complexity and nu merical instability of iterative routing mechanisms stem from the challenging na ture of the part-object encoding process. This hinders CapsNets from being utili zed effectively in large-scale image tasks. In this paper, we propose a novel no n-iterative routing strategy named self-attention routing (SAR) that computes th e agreement between the capsules in one forward pass. SAR accomplishes this by u tilizing a learnable inducing mixture of Gaussians (IMoG) to reduce the cost of computing pairwise attention values from quadratic to linear time complexity. Ou r observations show that our Transformer Capsule Network (Trans-Caps) is better suited for complex image tasks including CIFAR-10/100, Tiny-ImageNet, and ImageN et when compared to other prominent CapsNet architectures. We also show that Tra ns-Caps yields a dramatic improvement over its competitors when presented with n ovel viewpoints on the SmallNORB dataset, outperforming EM-Caps by 5.77% and 3.2 5% on the novel azimuth and elevation experiments, respectively. Our observation s suggest that our routing mechanism is able to capture complex part-whole relat ionships which allow Trans-Caps to construct reliable geometrical representation s of the objects.
****

End-to-end Quantized Training via Log-Barrier Extensions
Juncheng B Li,Shuhui Qu,Xinjian Li,Emma Strubell,Florian Metze
Quantization of neural network parameters and activations has emerged as a succe ssful approach to reducing the model size and inference time on hardware that su p-ports native low-precision arithmetic. Fully quantized training would facili tate further computational speed-ups as well as enable model training on embedde d devices, a feature that would alleviate privacy concerns resulting from the tr ansfer of sensitive data and models that is necessitated by off-device training.
  Existing approaches to quantization-aware training (QAT) perform “fake” quanti zation in the forward pass in order to learn model parameters that will perform well when quantized, but rely on higher precision variables to avoid overflow in large matrix multiplications, which is unsuitable for training on fully low-pre cision (e.g. 8-bit)hardware. To enable fully end-to-end quantized training, we p ropose Log BarrierTail-bounded Quantization (LogBTQ). LogBTQ introduces a loss t erm, inspired by the log-barrier for constrained optimization, that enforces sof t constraints on the range of values that model parameters can take on. By const raining and sparsifying model parameters, activations and inputs, our approach e liminates over-flow in practice, allowing for fully quantized 8-bit training of deep neural network models. We show that models trained using our approach achie ve results competitive with state-of-the-art full-precision networks on the MNIS T, CIFAR-10 andImageNet classification benchmarks.
****

Ballroom Dance Movement Recognition Using a Smart Watch and Representation Learn ing
Varun Badrinath Krishna
Smart watches are being increasingly used to detect human gestures and movements . Using a single smart watch, whole body movement recognition remains a hard pro blem because movements may not be adequately captured by the sensors in the watc h. In this paper, we present a whole body movement detection study using a singl

e smart watch in the context of ballroom dancing. Deep learning representations are used to classify well-defined sequences of movements, called \emph{figures}. Those representations are found to outperform ensembles of random forests and hidden Markov models. The classification accuracy of 85.95\% was improved to 92.31\% by modeling a dance as a first-order Markov chain of figures.
****
Manifold Regularization for Locally Stable Deep Neural Networks
Charles Jin,Martin Rinard
We apply concepts from manifold regularization to develop new regularization techniques for training locally stable deep neural networks. Our regularizers encourage functions which are smooth not only in their predictions but also their decision boundaries. Empirically, our networks exhibit stability in a diverse set of perturbation models, including ℓ_2, ℓ_∞, and Wasserstein-based perturbations; in particular, against a state-of-the-art PGD adversary, a single model achieves both ℓ_∞ robustness of 40% at $\epsilon = 8/255$ and ℓ_2 robustness of 48% at $\epsilon = 1.0$ on CIFAR-10. We also obtain state-of-the-art verified accuracy of 21% in the same ℓ_∞ setting. Furthermore, our techniques are efficient, incurring overhead on par with two additional parallel forward passes through the network; in the case of CIFAR-10, we achieve our results after training for only 3 hours, compared to more than 70 hours for standard adversarial training.
****
Spectral Synthesis for Satellite-to-Satellite Translation
Thomas Vandal,Daniel McDuff,Weile Wang,Andrew Michaelis,Ramakrishna Nemani
Earth observing satellites carrying multi-spectral sensors are widely used to monitor the physical and biological states of the atmosphere, land, and oceans. These satellites have different vantage points above the earth and different spectral imaging bands resulting in inconsistent imagery from one to another. This presents challenges in building downstream applications. What if we could generate synthetic bands for existing satellites from the union of all domains? We tackle the problem of generating synthetic spectral imagery for multispectral sensors as an unsupervised image-to-image translation problem with partial labels and introduce a novel shared spectral reconstruction loss. Simulated experiments performed by dropping one or more spectral bands show that cross-domain reconstruction outperforms measurements obtained from a second vantage point. On a downstream cloud detection task, we show that generating synthetic bands with our model improves segmentation performance beyond our baseline. Our proposed approach enables synchronization of multispectral data and provides a basis for more homogeneous remote sensing datasets.
****
Teaching Temporal Logics to Neural Networks
Christopher Hahn,Frederik Schmitt,Jens U. Kreber,Markus Norman Rabe,Bernd Finkbeiner
We study two fundamental questions in neuro-symbolic computing: can deep learning tackle challenging problems in logics end-to-end, and can neural networks learn the semantics of logics. In this work we focus on linear-time temporal logic (LTL), as it is widely used in verification. We train a Transformer on the problem to directly predict a solution, i.e. a trace, to a given LTL formula. The training data is generated with classical solvers, which, however, only provide one of many possible solutions to each formula. We demonstrate that it is sufficient to train on those particular solutions to formulas, and that Transformers can predict solutions even to formulas from benchmarks from the literature on which the classical solver timed out. Transformers also generalize to the semantics of the logics: while they often deviate from the solutions found by the classical solvers, they still predict correct solutions to most formulas.
****
Neuro-algorithmic Policies for Discrete Planning
Marin Vlastelica Pogan■i■,Michal Rolinek,Georg Martius
Although model-based and model-free approaches to learning the control of systems have achieved impressive results on standard benchmarks, generalization to var

iations in the task are still unsatisfactory. Recent results suggest that generalization of standard architectures improves only after obtaining exhaustive amounts of data.We give evidence that the generalization capabilities are in many cases bottlenecked by the inability to generalize on the combinatorial aspects. Further, we show that for a certain subclass of the MDP framework, this can be alleviated by neuro-algorithmic architectures.

Many control problems require long-term planning that is hard to solve generically with neural networks alone. We introduce a neuro-algorithmic policy architecture consisting of a neural network and an embedded time-depended shortest path solver. These policies can be trained end-to-end by blackbox differentiation. We show that this type of architecture generalizes well to unseen variations in the environment already after seeing a few examples.

https://sites.google.com/view/neuro-algorithmic
****
A Critical Analysis of Distribution Shift
Dan Hendrycks,Steven Basart,Norman Mu,Saurav Kadavath,Frank Wang,Evan Dorundo,Rahul Desai,Tyler Zhu,Samyak Parajuli,Mike Guo,Dawn Song,Jacob Steinhardt,Justin Gilmer
We introduce three new robustness benchmarks consisting of naturally occurring distribution changes in image style, geographic location, camera operation, and more. Using our benchmarks, we take stock of previously proposed hypotheses for out-of-distribution robustness and put them to the test. We find that using larger models and synthetic data augmentation can improve robustness on real-world distribution shifts, contrary to claims in prior work. Motivated by this, we introduce a new data augmentation method which advances the state-of-the-art and outperforms models pretrained with 1000x more labeled data. We find that some methods consistently help with distribution shifts in texture and local image statistics, but these methods do not help with some other distribution shifts like geographic changes. Hence no evaluated method consistently improves robustness. We conclude that future research should study multiple distribution shifts simultaneously.
****
Attainability and Optimality: The Equalized-Odds Fairness Revisited
Zeyu Tang,Kun Zhang
Fairness of machine learning algorithms has been of increasing interest. In order to suppress or eliminate discrimination in prediction, various notions as well as approaches to impose fairness have been proposed. However, in different scenarios, whether or not the chosen notion of fairness can always be attained, even if with unlimited amount of data, is not well addressed. In this paper, focusing on the Equalized Odds notion of fairness, we consider the attainability of this criterion, and furthermore, if attainable, the optimality of the prediction performance under various settings. In particular, for classification with a deterministic prediction function of the input, we give the condition under which Equalized Odds can hold true; if randomized prediction is acceptable, we show that under mild assumptions, fair classifiers can always be derived. Moreover, we prove that compared to enforcing fairness by post-processing, one can always benefit from exploiting all available features during training and get better prediction performance while remaining fair. However, for regression tasks, Equalized Odds is not always attainable if certain conditions on the joint distribution of the features and the target variable are not met. This indicates the inherent difficulty in achieving fairness in certain cases and suggests a broader class of prediction methods might be needed for fairness.
****
Bypassing the Ambient Dimension: Private SGD with Gradient Subspace Identification
Yingxue Zhou,Steven Wu,Arindam Banerjee
Differentially private SGD (DP-SGD) is one of the most popular methods for solvi

ng differentially private empirical risk minimization (ERM). Due to its noisy pe rturbation on each gradient update, the error rate of DP-SGD scales with the amb ient dimension p, the number of parameters in the model. Such dependence can b e problematic for over-parameterized models where $p \gg n$, the number of train ing samples. Existing lower bounds on private ERM show that such dependence on $ p$ is inevitable in the worst case. In this paper, we circumvent the dependence on the ambient dimension by leveraging a low-dimensional structure of gradient s pace in deep networks---that is, the stochastic gradients for deep nets usually stay in a low dimensional subspace in the training process. We propose Projected DP-SGD that performs noise reduction by projecting the noisy gradients to a low -dimensional subspace, which is given by the top gradient eigenspace on a small public dataset. We provide a general sample complexity analysis on the public da taset for the gradient subspace identification problem and demonstrate that unde r certain low-dimensional assumptions the public sample complexity only grows lo garithmically in p. Finally, we provide a theoretical analysis and empirical e valuations to show that our method can substantially improve the accuracy of DP-SGD in the high privacy regime (corresponding to low privacy loss ϵ).

****

Enforcing Predictive Invariance across Structured Biomedical Domains

Wengong Jin,Regina Barzilay,Tommi S. Jaakkola

Many biochemical applications such as molecular property prediction require mode ls to generalize beyond their training domains (environments). Moreover, natural environments in these tasks are structured, defined by complex descriptors such as molecular scaffolds or protein families. Therefore, most environments are ei ther never seen during training, or contain only a single training example. To a ddress these challenges, we propose a new regret minimization (RGM) algorithm an d its extension for structured environments. RGM builds from invariant risk mini mization (IRM) by recasting simultaneous optimality condition in terms of predic tive regret, finding a representation that enables the predictor to compete agai nst an oracle with hindsight access to held-out environments. The structured ext ension adaptively highlights variation due to complex environments via specializ ed domain perturbations. We evaluate our method on multiple applications: molecu lar property prediction, protein homology and stability prediction and show that RGM significantly outperforms previous state-of-the-art baselines.

****

Hamiltonian Q-Learning: Leveraging Importance-sampling for Data Efficient RL

Udari Madhushani,Biswadip Dey,Naomi Leonard,Amit Chakraborty

Model-free reinforcement learning (RL), in particular Q-learning is widely use d to learn optimal policies for a variety of planning and control problems. Howe ver, when the underlying state-transition dynamics are stochastic and high-dimen sional, Q-learning requires a large amount of data and incurs a prohibitively high computational cost. In this paper, we introduce Hamiltonian Q-Learning, a data efficient modification of the Q-learning approach, which adopts an impor tance-sampling based technique for computing the Q function. To exploit stocha stic structure of the state-transition dynamics, we employ Hamiltonian Monte Car lo to update Q function estimates by approximating the expected future rewards using Q values associated with a subset of next states. Further, to exploit t he latent low-rank structure of the dynamic system, Hamiltonian Q-Learning use s a matrix completion algorithm to reconstruct the updated Q function from Q value updates over a much smaller subset of state-action pairs. By providing an efficient way to apply Q-learning in stochastic, high-dimensional problems, t he proposed approach broadens the scope of RL algorithms for real-world applicat ions, including classical control tasks and environmental monitoring.

****

Adversarial and Natural Perturbations for General Robustness

Sadaf Gulshad,Jan Hendrik Metzen,Arnold W.M. Smeulders

In this paper we aim to explore the general robustness of neural network classif iers by utilizing adversarial as well as natural perturbations. Different from p

revious works which mainly focus on studying the robustness of neural networks a gainst adversarial perturbations, we also evaluate their robustness on natural p erturbations before and after robustification. After standardizing the compariso n between adversarial and natural perturbations, we demonstrate that although ad versarial training improves the performance of the networks against adversarial perturbations, it leads to drop in the performance for naturally perturbed sampl es besides clean samples. In contrast, natural perturbations like elastic deform ations, occlusions and wave does not only improve the performance against natura l perturbations, but also lead to improvement in the performance for the adversa rial perturbations. Additionally they do not drop the accuracy on the clean imag es.
****

Exploring Zero-Shot Emergent Communication in Embodied Multi-Agent Populations
Kalesha Bullard,Franziska Meier,Douwe Kiela,Joelle Pineau,Jakob Nicolaus Foerste r

Effective communication is an important skill for enabling information exchange and cooperation in multi-agent settings. Indeed, emergent communication is now a vibrant field of research, with common settings involving discrete cheap-talk c hannels. One limitation of this setting is that it does not allow for the emerge nt protocols to generalize beyond the training partners.
Furthermore, so far emergent communication has primarily focused on the use of symbolic channels. In this work, we extend this line of work to a new modality, by studying agents that learn to communicate via actuating their joints in a 3D environment. We show that under realistic assumptions, a non-uniform distributio n of intents and a common-knowledge energy cost, these agents can find protocols that generalize to novel partners. We also explore and analyze specific difficu lties associated with finding these solutions in practice. Finally, we propose a nd evaluate initial training improvements to address these challenges, involving both specific training curricula and providing the latent feature that can be c oordinated on during training.
****

Generative Time-series Modeling with Fourier Flows
Ahmed Alaa,Alex James Chan,Mihaela van der Schaar

Generating synthetic time-series data is crucial in various application domains, such as medical prognosis, wherein research is hamstrung by the lack of access to data due to concerns over privacy. Most of the recently proposed methods for generating synthetic time-series rely on implicit likelihood modeling using gene rative adversarial networks (GANs)—but such models can be difficult to train, an d may jeopardize privacy by “memorizing” temporal patterns in training data. In this paper, we propose an explicit likelihood model based on a novel class of no rmalizing flows that view time-series data in the frequency-domain rather than t he time-domain. The proposed flow, dubbed a Fourier flow, uses a discrete Fourie r transform (DFT) to convert variable-length time-series with arbitrary sampling periods into fixed-length spectral representations, then applies a (data-depend ent) spectral filter to the frequency-transformed time-series. We show that, by virtue of the DFT analytic properties, the Jacobian determinants and inverse map ping for the Fourier flow can be computed efficiently in linearithmic time, with out imposing explicit structural constraints as in existing flows such as NICE (Dinh et al. (2014)), RealNVP (Dinh et al. (2016)) and GLOW (Kingma & Dhariwal (2 018)). Experiments show that Fourier flows perform competitively compared to sta te-of-the-art baselines.
****

Optimal allocation of data across training tasks in meta-learning
Georgios Batzolis,Alberto Bernacchia,Da-shan Shiu,Michael Bromberg,Alexandru Cio ba

Meta-learning models transfer the knowledge acquired from previous tasks to quic kly learn new ones. They are tested on benchmarks with a fixed number of data-po ints for each training task, and this number is usually arbitrary, for example, 5 instances per class in few-shot classification. It is unknown how the performa nce of meta-learning is affected by the distribution of data across training tas

ks. Since labelling of data is expensive, finding the optimal allocation of labe
ls across training tasks may reduce costs.
Given a fixed budget b of labels to distribute across tasks, should we use a sma
ll number of highly labelled tasks, or many tasks with few labels each? In MAML
applied to mixed linear regression, we prove that the optimal number of tasks fo
llows the scaling law sqrt{b}. We develop an online algorithm for data allocatio
n across tasks, and show that the same scaling law applies to nonlinear regressi
on. We also show preliminary experiments on few-shot image classification. Our w
ork provides a theoretical guide for allocating labels across tasks in meta-lear
ning, which we believe will prove useful in a large number of applications.

****
Training Federated GANs with Theoretical Guarantees: A Universal Aggregation Ap
proach
Yikai Zhang,Hui Qu,Huidong Liu,Qi Chang,Dimitris N. Metaxas,Chao Chen
Recently, Generative Adversarial Networks (GANs) have demonstrated their potenti
al in federated learning, i.e., learning a centralized model from data privately
 hosted by multiple sites. A federated GAN jointly trains a centralized generato
r and multiple private discriminators hosted at different sites. A major theoret
ical challenge for the federated GAN is the heterogeneity of the local data dist
ributions. Traditional approaches cannot guarantee to learn the target distribut
ion, which is a mixture of the highly different local distributions. This paper
tackles this theoretical challenge, and for the first time, provides a provably
correct framework for federated GAN. We propose a new approach called Universal
Aggregation, which simulates a centralized discriminator via carefully aggregati
ng the mixture of all private discriminators. We prove that a generator trained
with this simulated centralized discriminator can learn the desired target distr
ibution. Through synthetic and real datasets, we show that our method can learn
the mixture of largely different distributions, when existing federated GAN meth
ods fail to.
****
Overparameterisation and worst-case generalisation: friend or foe?
Aditya Krishna Menon,Ankit Singh Rawat,Sanjiv Kumar
Overparameterised neural networks have demonstrated the remarkable ability to pe
rfectly fit training samples, while still generalising to unseen test samples. H
owever, several recent works have revealed that such models' good average perfor
mance does not always translate to good worst-case performance: in particular, t
hey may perform poorly on subgroups that are under-represented in the training s
et. In this paper, we show that in certain settings, overparameterised models' p
erformance on under-represented subgroups may be improved via post-hoc processin
g. Specifically, such models' bias can be restricted to their classification lay
ers, and manifest as structured prediction shifts for rare subgroups. We detail
two post-hoc correction techniques to mitigate this bias, which operate purely o
n the outputs of standard model training. We empirically verify that with such p
ost-hoc correction, overparameterisation can improve average and worst-case perf
ormance.
****
Model agnostic meta-learning on trees
Jezabel Garcia,Federica Freddi,Jamie McGowan,Tim Nieradzik,Da-shan Shiu,Ye Tian,
Alberto Bernacchia
In meta-learning, the knowledge learned from previous tasks is transferred to ne
w ones, but this transfer only works if tasks are related, and sharing informati
on between unrelated tasks might hurt performance. A fruitful approach is to sha
re gradients across similar tasks during training, and recent work suggests that
 the gradients themselves can be used as a measure of task similarity.
We study the case in which datasets associated to different tasks have a hierarc
hical, tree structure. While a few methods have been proposed for hierarchical m
eta-learning in the past, we propose the first algorithm that is model-agnostic,
 a simple extension of MAML. As in MAML, our algorithm adapts the model to each
task with a few gradient steps, but the adaptation follows the tree structure: i

n each step, gradients are pooled across task clusters, and subsequent steps fol
low down the tree. We test the algorithm on linear and non-linear regression on
synthetic data, and show that the algorithm significantly improves over MAML. In
terestingly, the algorithm performs best when it does not know in advance the tr
ee structure of the data.
****

AdaDGS: An adaptive black-box optimization method with a nonlocal directional Ga
ussian smoothing gradient
Hoang A Tran,Guannan Zhang
The local gradient points to the direction of the steepest slope in an infinites
imal neighborhood. An optimizer guided by the local gradient is often trapped in
 local optima when the loss landscape is multi-modal. A directional Gaussian smo
othing (DGS) approach was recently proposed in (Zhang et al., 2020) and used to
define a truly nonlocal gradient, referred to as the DGS gradient, for high-dime
nsional black-box optimization. Promising results show that replacing the tradi
tional local gradient with the DGS gradient can significantly improve the perfor
mance of gradient-based methods in optimizing highly multi-modal loss functions.
  However, the optimal performance of the DGS gradient may rely on fine tuning o
f two important hyper-parameters, i.e., the smoothing radius and the learning ra
te. In this paper, we present a simple, yet ingenious and efficient adaptive ap
proach for optimization with the DGS gradient, which removes the need of hyper-p
arameter fine tuning. Since the DGS gradient generally points to a good search
direction, we perform a line search along the DGS direction to determine the ste
p size at each iteration. The learned step size in turn will inform us of the sc
ale of function landscape in the surrounding area, based on which we adjust the
smoothing radius accordingly for the next iteration. We present experimental res
ults on high-dimensional benchmark functions, an airfoil design problem and a ga
me content generation problem. The AdaDGS method has shown superior performance
over several the state-of-the-art black-box optimization methods.
****

Near-Optimal Regret Bounds for Model-Free RL in Non-Stationary Episodic MDPs
Weichao Mao,Kaiqing Zhang,Ruihao Zhu,David Simchi-Levi,Tamer Basar
We consider model-free reinforcement learning (RL) in non-stationary Markov deci
sion processes (MDPs). Both the reward functions and the state transition distri
butions are allowed to vary over time, either gradually or abruptly, as long as
their cumulative variation magnitude does not exceed certain budgets. We propose
 an algorithm, named Restarted Q-Learning with Upper Confidence Bounds (RestartQ
-UCB), for this setting, which adopts a simple restarting strategy and an extra
optimism term. Our algorithm outperforms the state-of-the-art (model-based) solu
tion in terms of dynamic regret. Specifically, RestartQ-UCB with Freedman-type b
onus terms achieves a dynamic regret of $\widetilde{O}(S^{\frac{1}{3}} A^{\frac{
1}{3}} \Delta^{\frac{1}{3}} H T^{\frac{2}{3}})$, where S and A are the numbe
rs of states and actions, respectively, $\Delta>0$ is the variation budget, H
is the number of steps per episode, and T is the total number of steps. We fur
ther show that our algorithm is near-optimal by establishing an information-theo
retical lower bound of $\Omega(S^{\frac{1}{3}} A^{\frac{1}{3}} \Delta^{\frac{1}{
3}} H^{\frac{2}{3}} T^{\frac{2}{3}})$, which to the best of our knowledge is the
 first impossibility result in non-stationary RL in general.
****

VECoDeR - Variational Embeddings for Community Detection and Node Representation
Rayyan Ahmad Khan,Muhammad Umer Anwaar,Omran Kaddah,Martin Kleinsteuber
In this paper, we study how to simultaneously learn two highly correlated tasks
of graph analysis, i.e., community detection and node representation learning. W
e propose an efficient generative model called VECoDeR for jointly learning Vari
ational Embeddings for COmmunity DEtection and node Representation. VECoDeR assu
mes that every node can be a member of one or more communities. The node embeddi
ngs are learned in such a way that connected nodes are not only ``closer" to eac
h other but also share similar community assignments. A joint learning framework
 leverages community-aware node embeddings for better community detection. We de
monstrate on several graph datasets that VECoDeR effectively outperforms many co

mpetitive baselines on all three tasks i.e. node classification, overlapping com
munity detection and non-overlapping community detection. We also show that VECo
DeR is computationally efficient and has quite robust performance with varying h
yperparameters.
****

Improving Zero-Shot Voice Style Transfer via Disentangled Representation Learnin
g

Siyang Yuan,Pengyu Cheng,Ruiyi Zhang,Weituo Hao,Zhe Gan,Lawrence Carin

Voice style transfer, also called voice conversion, seeks to modify one speaker'
s voice to generate speech as if it came from another (target) speaker. Previous
 works have made progress on voice conversion with parallel training data and pr
e-known speakers. However, zero-shot voice style transfer, which learns from non
-parallel data and generates voices for previously unseen speakers, remains a ch
allenging problem. In this paper we propose a novel zero-shot voice transfer met
hod via disentangled representation learning. The proposed method first encodes
speaker-related style and voice content of each input voice into separate low-di
mensional embedding spaces, and then transfers to a new voice by combining the s
ource content embedding and target style embedding through a decoder. With infor
mation-theoretic guidance, the style and content embedding spaces are representa
tive and (ideally) independent of each other. On real-world datasets, our method
 outperforms other baselines and obtains state-of-the-art results in terms of tr
ansfer accuracy and voice naturalness.
****

Meta-learning with negative learning rates

Alberto Bernacchia

Deep learning models require a large amount of data to perform well. When data i
s scarce for a target task, we can transfer the knowledge gained by training on
similar tasks to quickly learn the target. A successful approach is meta-learnin
g, or "learning to learn" a distribution of tasks, where "learning" is represent
ed by an outer loop, and "to learn" by an inner loop of gradient descent. Howeve
r, a number of recent empirical studies argue that the inner loop is unnecessary
 and more simple models work equally well or even better. We study the performan
ce of MAML as a function of the learning rate of the inner loop, where zero lear
ning rate implies that there is no inner loop. Using random matrix theory and ex
act solutions of linear models, we calculate an algebraic expression for the tes
t loss of MAML applied to mixed linear regression and nonlinear regression with
overparameterized models. Surprisingly, while the optimal learning rate for adap
tation is positive, we find that the optimal learning rate for training is alway
s negative, a setting that has never been considered before. Therefore, not only
 does the performance increase by decreasing the learning rate to zero, as sugge
sted by recent work, but it can be increased even further by decreasing the lear
ning rate to negative
values. These results help clarify under what circumstances meta-learning perfor
ms best.
****

Learning to Recombine and Resample Data For Compositional Generalization

Ekin Akyürek,Afra Feyza Akyürek,Jacob Andreas

Flexible neural sequence models outperform grammar- and automaton-based counterp
arts on a variety of tasks. However, neural models perform poorly in settings re
quiring compositional generalization beyond the training data—particularly to ra
re or unseen subsequences. Past work has found symbolic scaffolding (e.g. gramma
rs or automata) essential in these settings. We describe R&R, a learned data aug
mentation scheme that enables a large category of compositional generalizations
without appeal to latent symbolic structure. R&R has two components: recombinati
on of original training examples via a prototype-based generative model and resa
mpling of generated examples to encourage extrapolation. Training an ordinary ne
ural sequence model on a dataset augmented with recombined and resampled example
s significantly improves generalization in two language processing problems—inst
ruction following (SCAN) and morphological analysis (SIGMORPHON 2018)—where R&R
enables learning of new constructions and tenses from as few as eight initial ex

amples.
****
Revisiting the Stability of Stochastic Gradient Descent: A Tightness Analysis
Yikai Zhang,Samuel Bald,wenjia Zhang,Vamsi Pritham Pingali,Chao Chen,Mayank Goswami
The technique of algorithmic stability has been used to capture the generalization power of several learning models, especially those trained with stochastic gradient descent (SGD). This paper investigates the tightness of the algorithmic stability bounds for SGD given by~\cite{hardt2016train}. We show that the analysis of~\cite{hardt2016train} is tight for convex objective functions, but loose for non-convex objective functions. In the non-convex case we provide a tighter upper bound on the stability (and hence generalization error), and provide evidence that it is asymptotically tight up to a constant factor.
However, deep neural networks trained with SGD exhibit much better stability and generalization in practice than what is suggested by these (tight) bounds, namely, linear or exponential degradation with time for SGD with constant step size. We aim towards characterizing deep learning loss functions with good generalization guarantees, despite being trained using SGD with constant step size. We propose the Hessian Contractive (HC) condition, which specifies the contractivity of regions containing local minima in the neural network loss landscape. We provide empirical evidence that this condition holds for several loss functions, and provide theoretical evidence that the known tight SGD stability bounds for convex and non-convex loss functions can be circumvented by HC loss functions, thus partially explaining the generalization of deep neural networks.
****
Dataset Inference: Ownership Resolution in Machine Learning
Pratyush Maini,Mohammad Yaghini,Nicolas Papernot
With increasingly more data and computation involved in their training, machine learning models constitute valuable intellectual property. This has spurred interest in model stealing, which is made more practical by advances in learning with partial, little, or no supervision. Existing defenses focus on inserting unique watermarks in a model's decision surface, but this is insufficient: the watermarks are not sampled from the training distribution and thus are not always preserved during model stealing. In this paper, we make the key observation that knowledge contained in the stolen model's training set is what is common to all stolen copies. The adversary's goal, irrespective of the attack employed, is always to extract this knowledge or its by-products. This gives the original model's owner a strong advantage over the adversary: model owners have access to the original training data. We thus introduce $\textit{dataset inference}$, the process of identifying whether a suspected model copy has private knowledge from the original model's dataset, as a defense against model stealing. We develop an approach for dataset inference that combines statistical testing with the ability to estimate the distance of multiple data points to the decision boundary. Our experiments on CIFAR10, SVHN, CIFAR100 and ImageNet show that model owners can claim with confidence greater than 99% that their model (or dataset as a matter of fact) was stolen, despite only exposing 50 of the stolen model's training points. Dataset inference defends against state-of-the-art attacks even when the adversary is adaptive. Unlike prior work, it does not require retraining or overfitting the defended model.
****
Fooling a Complete Neural Network Verifier
Dániel Zombori,Balázs Bánhelyi,Tibor Csendes,István Megyeri,Márk Jelasity
The efficient and accurate characterization of the robustness of neural networks to input perturbation is an important open problem. Many approaches exist including heuristic and exact (or complete) methods. Complete methods are expensive but their mathematical formulation guarantees that they provide exact robustness metrics. However, this guarantee is valid only if we assume that the verified network applies arbitrary-precision arithmetic and the verifier is reliable. In practice, however, both the networks and the verifiers apply limited-precision floating point arithmetic. In this paper, we show that numerical roundoff errors ca

n be exploited to craft adversarial networks, in which the actual robustness and the robustness computed by a state-of-the-art complete verifier radically differ. We also show that such adversarial networks can be used to insert a backdoor into any network in such a way that the backdoor is completely missed by the verifier. The attack is easy to detect in its naive form but, as we show, the adversarial network can be transformed to make its detection less trivial. We offer a simple defense against our particular attack based on adding a very small perturbation to the network weights. However, our conjecture is that other numerical attacks are possible, and exact verification has to take into account all the details of the computation executed by the verified networks, which makes the problem significantly harder.

****

UMEC: Unified model and embedding compression for efficient recommendation systems

Jiayi Shen,Haotao Wang,Shupeng Gui,Jianchao Tan,Zhangyang Wang,Ji Liu

The recommendation system (RS) plays an important role in the content recommendation and retrieval scenarios. The core part of the system is the Ranking neural network, which is usually a bottleneck of whole system performance during online inference. In this work, we propose a unified model and embedding compression (UMEC) framework to hammer an efficient neural network-based recommendation system. Our framework jointly learns input feature selection and neural network compression together, and solve them as an end-to-end resource-constrained optimization problem using ADMM. Our method outperforms other baselines in terms of neural network Flops, sparse embedding feature size and the number of sparse embedding features. We evaluate our method on the public benchmark of DLRM, trained over the Kaggle Criteo dataset. The codes can be found at https://github.com/VITA-Group/UMEC.

****

Evaluations and Methods for Explanation through Robustness Analysis

Cheng-Yu Hsieh,Chih-Kuan Yeh,Xuanqing Liu,Pradeep Kumar Ravikumar,Seungyeon Kim,Sanjiv Kumar,Cho-Jui Hsieh

Feature based explanations, that provide importance of each feature towards the model prediction, is arguably one of the most intuitive ways to explain a model. In this paper, we establish a novel set of evaluation criteria for such feature based explanations by robustness analysis. In contrast to existing evaluations which require us to specify some way to "remove" features that could inevitably introduces biases and artifacts, we make use of the subtler notion of smaller adversarial perturbations. By optimizing towards our proposed evaluation criteria, we obtain new explanations that are loosely necessary and sufficient for a prediction. We further extend the explanation to extract the set of features that would move the current prediction to a target class by adopting targeted adversarial attack for the robustness analysis. Through experiments across multiple domains and a user study, we validate the usefulness of our evaluation criteria and our derived explanations.

****

Learning One-hidden-layer Neural Networks on Gaussian Mixture Models with Guaranteed Generalizability

Hongkang Li,Shuai Zhang,Meng Wang

We analyze the learning problem of fully connected neural networks with the sigmoid activation function for binary classification in the teacher-student setup, where the outputs are assumed to be generated by a ground-truth teacher neural network with unknown parameters, and the learning objective is to estimate the teacher network model by minimizing a non-convex cross-entropy risk function of the training data over a student neural network. This paper analyzes a general and practical scenario that the input features follow a Gaussian mixture model of a finite number of Gaussian distributions of various mean and variance. We propose a gradient descent algorithm with a tensor initialization approach and show that our algorithm converges linearly to a critical point that has a diminishing

distance to the ground-truth model with guaranteed generalizability. We characterize the required number of samples for successful convergence, referred to as the sample complexity, as a function of the parameters of the Gaussian mixture model. We prove analytically that when any mean or variance in the mixture model is large, or when all variances are close to zero, the sample complexity increases, and the convergence slows down, indicating a more challenging learning problem. Although focusing on one-hidden-layer neural networks, to the best of our knowledge, this paper provides the first explicit characterization of the impact of the parameters of the input distributions on the sample complexity and learning rate.

****

Provable Fictitious Play for General Mean-Field Games

Qiaomin Xie,Zhuoran Yang,Zhaoran Wang,Andreea Minca

We propose a reinforcement learning algorithm for stationary mean-field games, where the goal is to learn a pair of mean-field state and stationary policy that constitutes the Nash equilibrium. When viewing the mean-field state and the policy as two players, we propose a fictitious play algorithm which alternatively updates the mean-field state and the policy via gradient-descent and proximal policy optimization, respectively. Our algorithm is in stark contrast with previous literature which solves each single-agent reinforcement learning problem induced by the iterates mean-field states to the optimum. Furthermore, we prove that our fictitious play algorithm converges to the Nash equilibrium at a sublinear rate. To the best of our knowledge, this seems the first provably convergent reinforcement learning algorithm for mean-field games based on iterative updates of both mean-field state and policy.

****

Learning and Evaluating Representations for Deep One-Class Classification

Kihyuk Sohn,Chun-Liang Li,Jinsung Yoon,Minho Jin,Tomas Pfister

We present a two-stage framework for deep one-class classification. We first learn self-supervised representations from one-class data, and then build one-class classifiers on learned representations. The framework not only allows to learn better representations, but also permits building one-class classifiers that are faithful to the target task. We argue that classifiers inspired by the statistical perspective in generative or discriminative models are more effective than existing approaches, such as a normality score from a surrogate classifier. We thoroughly evaluate different self-supervised representation learning algorithms under the proposed framework for one-class classification. Moreover, we present a novel distribution-augmented contrastive learning that extends training distributions via data augmentation to obstruct the uniformity of contrastive representations. In experiments, we demonstrate state-of-the-art performance on visual domain one-class classification benchmarks, including novelty and anomaly detection. Finally, we present visual explanations, confirming that the decision-making process of deep one-class classifiers is intuitive to humans. The code is available at https://github.com/google-research/deep_representation_one_class.

****

Learning from Protein Structure with Geometric Vector Perceptrons

Bowen Jing,Stephan Eismann,Patricia Suriana,Raphael John Lamarre Townshend,Ron Dror

Learning on 3D structures of large biomolecules is emerging as a distinct area in machine learning, but there has yet to emerge a unifying network architecture that simultaneously leverages the geometric and relational aspects of the problem domain. To address this gap, we introduce geometric vector perceptrons, which extend standard dense layers to operate on collections of Euclidean vectors. Graph neural networks equipped with such layers are able to perform both geometric and relational reasoning on efficient representations of macromolecules. We demonstrate our approach on two important problems in learning from protein structure: model quality assessment and computational protein design. Our approach improves over existing classes of architectures on both problems, including state-of-the-art convolutional neural networks and graph neural networks. We release our

code at https://github.com/drorlab/gvp.
****
Unsupervised Object Keypoint Learning using Local Spatial Predictability
Anand Gopalakrishnan,Sjoerd van Steenkiste,Jürgen Schmidhuber
We propose PermaKey, a novel approach to representation learning based on object keypoints. It leverages the predictability of local image regions from spatial neighborhoods to identify salient regions that correspond to object parts, which are then converted to keypoints. Unlike prior approaches, it utilizes predictability to discover object keypoints, an intrinsic property of objects. This ensures that it does not overly bias keypoints to focus on characteristics that are not unique to objects, such as movement, shape, colour etc. We demonstrate the efficacy of PermaKey on Atari where it learns keypoints corresponding to the most salient object parts and is robust to certain visual distractors. Further, on downstream RL tasks in the Atari domain we demonstrate how agents equipped with our keypoints outperform those using competing alternatives, even on challenging environments with moving backgrounds or distractor objects.

****
On Representing (Anti)Symmetric Functions
Marcus Hutter
Permutation-invariant, -equivariant, and -covariant functions and anti-symmetric functions are important in quantum physics, computer vision, and other disciplines. Applications often require most or all of the following properties: (a) a large class of such functions can be approximated, e.g. all continuous function (b) only the (anti)symmetric functions can be represented (c) a fast algorithm for computing the approximation (d) the representation itself is continuous or differentiable (e) the architecture is suitable for learning the function from data (Anti)symmetric neural networks have recently been developed and applied with great success. A few theoretical approximation results have been proven, but many questions are still open, especially for particles in more than one dimension and the anti-symmetric case, which this work focuses on. More concretely, we derive natural polynomial approximations in the symmetric case, and approximations based on a single generalized Slater determinant in the anti-symmetric case. Unlike some previous super-exponential and discontinuous approximations, these seem a more promising basis for future tighter bounds.
****
Towards Understanding Label Smoothing
Yi Xu,Yuanhong Xu,Qi Qian,Li Hao,Rong Jin
Label smoothing regularization (LSR) has a great success in training deep neural networks by stochastic algorithms such as stochastic gradient descent and its variants. However, the theoretical understanding of its power from the view of optimization is still rare. This study opens the door to a deep understanding of LSR by initiating the analysis. In this paper, we analyze the convergence behaviors of stochastic gradient descent with label smoothing regularization for solving non-convex problems and show that an appropriate LSR can help to speed up the convergence by reducing the variance. More interestingly, we proposed a simple yet effective strategy, namely Two-Stage LAbel smoothing algorithm (TSLA), that uses LSR in the early training epochs and drops it off in the later training epochs. We observe from the improved convergence result of TSLA that it benefits from LSR in the first stage and essentially converges faster in the second stage. To the best of our knowledge, this is the first work for understanding the power of LSR via establishing convergence complexity of stochastic methods with LSR in non-convex optimization. We empirically demonstrate the effectiveness of the proposed method in comparison with baselines on training ResNet models over benchmark data sets.
****
Importance-based Multimodal Autoencoder
Sayan Ghosh,Eugene Laksana,Louis-Philippe Morency,Stefan Scherer
Integrating information from multiple modalities (e.g., verbal, acoustic and visual data) into meaningful representations has seen great progress in recent ye

ars. However, two challenges are not sufficiently addressed by current approache s: (1) computationally efficient training of multimodal autoencoder networks whi ch are robust in the absence of modalities, and (2) unsupervised learning of imp ortant subspaces in each modality which are correlated with other modalities.
In this paper we propose the IMA (Importance-based Multimodal Autoencoder) model , a scalable model that learns modality importances and robust multimodal repres entations through a novel cross-covariance based loss function. We conduct expe riments on MNIST-TIDIGITS a multimodal dataset of spoken and image digits,and on IEMOCAP, a multimodal emotion corpus. The IMA model is able to distinguish dig its from uncorrelated noise, and word-level importances learnt that corr espond to the separation between function and emotional words. The multimodal r epresentations learnt by IMA are also competitive with state-of-the-art baseline approaches on downstream tasks.
****

Conditional Negative Sampling for Contrastive Learning of Visual Representations
Mike Wu,Milan Mosse,Chengxu Zhuang,Daniel Yamins,Noah Goodman
Recent methods for learning unsupervised visual representations, dubbed contrast ive learning, optimize the noise-contrastive estimation (NCE) bound on mutual in formation between two transformations of an image. NCE typically uses randomly s ampled negative examples to normalize the objective, but this may often include many uninformative examples either because they are too easy or too hard to disc riminate. Taking inspiration from metric learning, we show that choosing semi-h ard negatives can yield stronger contrastive representations. To do this, we int roduce a family of mutual information estimators that sample negatives condition ally -- in a "ring" around each positive. We prove that these estimators remain lower-bounds of mutual information, with higher bias but lower variance than NCE . Experimentally, we find our approach, applied on top of existing models (IR, C MC, and MoCo) improves accuracy by 2-5% absolute points in each case, measured b y linear evaluation on four standard image benchmarks. Moreover, we find continu ed benefits when transferring features to a variety of new image distributions f rom the Meta-Dataset collection and to a variety of downstream tasks such as obj ect detection, instance segmentation, and key-point detection.
****

Communication in Multi-Agent Reinforcement Learning: Intention Sharing
Woojun Kim,Jongeui Park,Youngchul Sung
Communication is one of the core components for learning coordinated behavior in multi-agent systems.
In this paper, we propose a new communication scheme named Intention Sharing (I S) for multi-agent reinforcement learning in order to enhance the coordination a mong agents. In the proposed IS scheme, each agent generates an imagined traject ory by modeling the environment dynamics and other agents' actions. The imagined trajectory is the simulated future trajectory of each agent based on the learne d model of the environment dynamics and other agents and represents each agent's future action plan. Each agent compresses this imagined trajectory capturing it s future action plan to generate its intention message for communication by appl ying an attention mechanism to learn the relative importance of the components i n the imagined trajectory based on the received message from other agents. Numer al results show that the proposed IS scheme outperforms other communication sche mes in multi-agent reinforcement learning.
****

Neural Thompson Sampling
Weitong ZHANG,Dongruo Zhou,Lihong Li,Quanquan Gu
Thompson Sampling (TS) is one of the most effective algorithms for solving conte xtual multi-armed bandit problems. In this paper, we propose a new algorithm, ca lled Neural Thompson Sampling, which adapts deep neural networks for both explor ation and exploitation. At the core of our algorithm is a novel posterior distri bution of the reward, where its mean is the neural network approximator, and its variance is built upon the neural tangent features of the corresponding neural network. We prove that, provided the underlying reward function is bounded, the proposed algorithm is guaranteed to achieve a cumulative regret of $O(T^{1/2})$,

which matches the regret of other contextual bandit algorithms in terms of total round number T. Experimental comparisons with other benchmark bandit algorithms on various data sets corroborate our theory.
****

A Maximum Mutual Information Framework for Multi-Agent Reinforcement Learning
Woojun Kim,Whiyoung Jung,Myungsik Cho,Youngchul Sung
In this paper, we propose a maximum mutual information (MMI) framework for multi-agent reinforcement learning (MARL) to enable multiple agents to learn coordinated behaviors by regularizing the accumulated return with the mutual information between actions. By introducing a latent variable to induce nonzero mutual information between actions and applying a variational bound, we derive a tractable lower bound on the considered MMI-regularized objective function. Applying policy iteration to maximize the derived lower bound, we propose a practical algorithm named variational maximum mutual information multi-agent actor-critic (VM3-AC), which follows centralized learning with decentralized execution (CTDE). We evaluated VM3-AC for several games requiring coordination, and numerical results show that VM3-AC outperforms MADDPG and other MARL algorithms in multi-agent tasks requiring coordination.
****

WeMix: How to Better Utilize Data Augmentation
Yi Xu,Asaf Noy,Ming Lin,Qi Qian,Li Hao,Rong Jin
Data augmentation is a widely used training trick in deep learning to improve the network generalization ability. Despite many encouraging results, several recent studies did point out limitations of the conventional data augmentation scheme in certain scenarios, calling for a better theoretical understanding of data augmentation. In this work, we develop a comprehensive analysis that reveals pros and cons of data augmentation. The main limitation of data augmentation arises from the data bias, i.e. the augmented data distribution can be quite different from the original one. This data bias leads to a suboptimal performance of existing data augmentation methods. To this end, we develop two novel algorithms, termed "AugDrop" and "MixLoss", to correct the data bias in the data augmentation. Our theoretical analysis shows that both algorithms are guaranteed to improve the effect of data augmentation through the bias correction, which is further validated by our empirical studies. Finally, we propose a generic algorithm "WeMix" by combining AugDrop and MixLoss, whose effectiveness is observed from extensive empirical evaluations.
****

Anomaly detection in dynamical systems from measured time series
Andrei Ivanov,Anna Golovkina
The paper addresses a problem of abnormalities detection in nonlinear processes represented by measured time series. Anomaly detection problem is usually formulated as finding outlier data points relative to some usual signals such as unexpected spikes, drops, or trend changes. In nonlinear dynamical systems, there are cases where a time series does not contain statistical outliers while the process corresponds to an abnormal configuration of the dynamical system. Since the polynomial neural architecture has a strong connection with the theory of differential equations, we use it for the feature extraction that describes the dynamical system itself. The paper discusses in both simulations and a practical example with real measurements the applicability of the proposed approach and it's benchmarking with existing methods.
****

Improving Generalizability of Protein Sequence Models via Data Augmentations
Hongyu Shen,Layne C. Price,Mohammad Taha Bahadori,Franziska Seeger
While protein sequence data is an emerging application domain for machine learning methods, small modifications to protein sequences can result in difficult-to-predict changes to the protein's function. Consequently, protein machine learning models typically do not use randomized data augmentation procedures analogous to those used in computer vision or natural language, e.g., cropping or synonym substitution. In this paper, we empirically explore a set of simple string manipulations, which we use to augment protein sequence data when fine-tuning semi-su

pervised protein models. We provide 276 different comparisons to the Tasks Asses sing Protein Embeddings (TAPE) baseline models, with Transformer-based models an d training datasets that vary from the baseline methods only in the data augment ations and representation learning procedure. For each TAPE validation task, we demonstrate improvements to the baseline scores when the learned protein represe ntation is fixed between tasks. We also show that contrastive learning fine-tun ing methods typically outperform masked-token prediction in these models, with i ncreasing amounts of data augmentation generally improving performance for contr astive learning protein methods. We find the most consistent results across TAP E tasks when using domain-motivated transformations, such as amino acid replacem ent, as well as restricting the Transformer attention to randomly sampled sub-re gions of the protein sequence. In rarer cases, we even find that information-de stroying augmentations, such as randomly shuffling entire protein sequences, can improve downstream performance.
****

Randomized Entity-wise Factorization for Multi-Agent Reinforcement Learning
Shariq Iqbal,Christian Schroeder de Witt,Bei Peng,Wendelin Boehmer,Shimon Whites on,Fei Sha
Real world multi-agent tasks often involve varying types and quantities of agent s and non-agent entities; however, agents within these tasks rarely need to cons ider all others at all times in order to act effectively. Factored value functio n approaches have historically leveraged such independences to improve learning efficiency, but these approaches typically rely on domain knowledge to select fi xed subsets of state features to include in each factor. We propose to utilize v alue function factoring with random subsets of entities in each factor as an aux iliary objective in order to disentangle value predictions from irrelevant entit ies. This factoring approach is instantiated through a simple attention mechanis m masking procedure. We hypothesize that such an approach helps agents learn mor e effectively in multi-agent settings by discovering common trajectories across episodes within sub-groups of agents/entities. Our approach, Randomized Entity-w ise Factorization for Imagined Learning (REFIL), outperforms all strong baseline s by a significant margin in challenging StarCraft micromanagement tasks.
****

Optimal Regularization can Mitigate Double Descent
Preetum Nakkiran,Prayaag Venkat,Sham M. Kakade,Tengyu Ma
Recent empirical and theoretical studies have shown that many learning algorithm s -- from linear regression to neural networks -- can have test performance that is non-monotonic in quantities such the sample size and model size. This striki ng phenomenon, often referred to as "double descent", has raised questions of if we need to re-think our current understanding of generalization. In this work, we study whether the double-descent phenomenon can be avoided by using optimal r egularization. Theoretically, we prove that for certain linear regression models with isotropic data distribution, optimally-tuned ℓ_2 regularization achie ves monotonic test performance as we grow either the sample size or the model si ze.
We also demonstrate empirically that optimally-tuned ℓ_2 regularization can mitigate double descent for more general models, including neural networks.
Our results suggest that it may also be informative to study the test risk scali ngs of various algorithms in the context of appropriately tuned regularization.
****

Separation and Concentration in Deep Networks
John Zarka,Florentin Guth,Stéphane Mallat
Numerical experiments demonstrate that deep neural network classifiers progressi vely separate class distributions around their mean, achieving linear separabili ty on the training set, and increasing the Fisher discriminant ratio. We explain this mechanism with two types of operators. We prove that a rectifier without b iases applied to sign-invariant tight frames can separate class means and increa se Fisher ratios. On the opposite, a soft-thresholding on tight frames can reduc e within-class variabilities while preserving class means. Variance reduction bo unds are proved for Gaussian mixture models. For image classification, we show t

hat separation of class means can be achieved with rectified wavelet tight frame
s that are not learned. It defines a scattering transform. Learning $1 \times 1
$ convolutional tight frames along scattering channels and applying a soft-thres
holding reduces within-class variabilities. The resulting scattering network rea
ches the classification accuracy of ResNet-18 on CIFAR-10 and ImageNet, with few
er layers and no learned biases.
****

Variance Reduction in Hierarchical Variational Autoencoders
Adeel Pervez,Efstratios Gavves
Variational autoencoders with deep hierarchies of stochastic layers have been kn
own to suffer from the problem of posterior collapse, where the top layers fall
back to the prior and become independent of input.
We suggest that the hierarchical VAE objective explicitly includes the variance
of the function parameterizing the mean and variance of the latent Gaussian dist
ribution which itself is often a high variance function.
Building on this we generalize VAE neural networks by incorporating a smoothing
parameter motivated by Gaussian analysis to reduce variance in parameterizing fu
nctions and show that this can help to solve the problem of posterior collapse.
We further show that under such smoothing the VAE loss exhibits a phase transiti
on, where the top layer KL divergence sharply drops to zero at a critical value
of the smoothing parameter.
We validate the phenomenon across model configurations and datasets.
****

GANMEX: Class-Targeted One-vs-One Attributions using GAN-based Model Explainabil
ity
Sheng-Min Shih,Pin-Ju Tien,Zohar Karnin
Attribution methods have been shown as promising approaches for identifying key
features that led to learned model predictions. While most existing attribution
methods rely on a baseline input for performing feature perturbations, limited r
esearch has been conducted to address the baseline selection issues. Poor choice
s of baselines can lead to unfair attributions as well as limited ability of one
-vs-one explanations for multi-class classifiers, which means explaining why the
 input belongs to its original class but not the other specified target class. A
chieving one-vs-one explanation is crucial when certain classes are more similar
 than others, e.g. two bird types among multiple animals. One-vs-one explanation
s focus on key differentiating features rather than features shared across the o
riginal and the target classes. In this paper, we present GANMEX, a novel algori
thm applying Generative Adversarial Networks (GAN) by incorporating the to-be-ex
plained classifier as part of the adversarial networks. Our approach effectively
 selects the baseline as the closest realistic sample belong to the target class
, which allows attribution methods to provide true one-vs-one explanations. We s
howed that GANMEX baselines improved the saliency maps visually and led to stron
ger performance on perturbation-based evaluation metrics over the existing basel
ines. Attribution results with the existing baselines are known to be insensitiv
e to model randomization, and we demonstrated that GANMEX baselines led to bette
r outcome under the randomization sanity checks.
****

Fast Geometric Projections for Local Robustness Certification
Aymeric Fromherz,Klas Leino,Matt Fredrikson,Bryan Parno,Corina Pasareanu
Local robustness ensures that a model classifies all inputs within an ℓ_p-b
all consistently, which precludes various forms of adversarial inputs.
In this paper, we present a fast procedure for checking local robustness in feed
-forward neural networks with piecewise-linear activation functions.
Such networks partition the input space into a set of convex polyhedral regions
in which the network’s behavior is linear;
hence, a systematic search for decision boundaries within the regions around a g
iven input is sufficient for assessing robustness.
Crucially, we show how the regions around a point can be analyzed using simple g
eometric projections, thus admitting an efficient, highly-parallel GPU implement
ation that excels particularly for the ℓ_2 norm, where previous work has be

en less effective.
Empirically we find this approach to be far more precise than many approximate v erification approaches, while at the same time performing multiple orders of mag nitude faster than complete verifiers, and scaling to much deeper networks.
****
Transferable Unsupervised Robust Representation Learning
De-An Huang,Zhiding Yu,Anima Anandkumar
Robustness is an important, and yet, under-explored aspect of unsupervised repre sentation learning, which has seen a lot of recent developments. In this work, w e address this gap by developing a novel framework: Unsupervised Robust Represen tation Learning (URRL), which combines unsupervised representation learning's pr etext task and robust supervised learning (e.g., AugMix). Moreover, it is commo nly assumed that there needs to be a trade-off between natural accuracy (on clea n data) and robust accuracy (on corrupted data). We upend this view and show tha t URRL improves both the natural accuracy of unsupervised representation learnin g and its robustness to corruptions and adversarial noise. A further challenge i s that the robustness of a representation might not be preserved in the transfer learning process after fine-tuning on downstream tasks. We develop transferable robustness by proposing a task-agnostic similarity regularization during the fi ne-tuning process. We show that this improves the robustness of the resulting mo del without the need for any adversarial training or further data augmentation d uring fine-tuning.
****
Distributional Generalization: A New Kind of Generalization
Preetum Nakkiran,Yamini Bansal
We introduce a new notion of generalization--- Distributional Generalization--- which roughly states that outputs of a classifier at train and test time are clo se as distributions, as opposed to close in just their average error. For exampl e, if we mislabel 30% of dogs as cats in the train set of CIFAR-10, then a ResNe t 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 behavior is n ot captured by classical generalization, which would only consider the average e rror and not the distribution of errors over the input domain. Our formal conjec tures, which are much more general than this example, characterize the form of d istributional generalization that can be expected in terms of problem parameters : model architecture, training procedure, number of samples, and data distributi on. We give empirical evidence for these conjectures across a variety of domains in machine learning, including neural networks, kernel machines, and decision t rees. Our results thus advance our understanding of interpolating classifiers.
****
Differentiable Learning of Graph-like Logical Rules from Knowledge Graphs
Hongzhi Shi,quanming yao,Yong Li
Logical rules inside a knowledge graph (KG) are essential for reasoning, logical inference, and rule mining. However, existing works can only handle simple, i.e ., chain-like and tree-like, rules and cannot capture KG's complex semantics, wh ich can be better captured by graph-like rules. Besides, learning graph-like rul es is very difficult because the graph structure exhibits a huge discrete search space. To address these issues, observing that the plausibility of logical rule s can be explained by how frequently it appears in a KG, we propose a score func tion that represents graph-like rules with learnable parameters. The score also helps relax the discrete space into a continuous one and can be uniformly transf ormed into matrix form by the Einstein summation convention. Thus, it allows us to learn graph-like rules in an efficient, differentiable, and end-to-end traini ng manner by optimizing the normalized score. We conduct extensive experiments o n real-world datasets to show that our method outperforms previous works due to logical rules' better expressive ability. Furthermore, we demonstrate that our m ethod can learn high-quality and interpretable graph-like logical rules.
****
GINN: Fast GPU-TEE Based Integrity for Neural Network Training
Aref Asvadishirehjini,Murat Kantarcioglu,Bradley A. Malin

Machine learning models based on Deep Neural Networks (DNNs) are increasingly being deployed in a wide range of applications ranging from self-driving cars to Covid-19 diagnostics. The computational power necessary to learn a DNN is non-trivial. So, as a result, cloud environments with dedicated hardware support emerged as important infrastructure. However, outsourcing computation to the cloud raises security, privacy, and integrity challenges. To address these challenges, previous works tried to leverage homomorphic encryption, secure multi-party computation, and trusted execution environments (TEE). Yet, none of these approaches can scale up to support realistic DNN model training workloads with deep architectures and millions of training examples without sustaining a significant performance hit. In this work, we focus on the setting where the integrity of the outsourced Deep Learning (DL) model training is ensured by TEE. We choose the TEE based approach because it has been shown to be more efficient compared to the pure cryptographic solutions, and the availability of TEEs on cloud environments. To mitigate the loss in performance, we combine random verification of selected computation steps with careful adjustments of DNN used for training. Our experimental results show that the proposed approach may achieve 2X to 20X performance improvement compared to the pure TEE based solution while guaranteeing the integrity of the computation with high probability (e.g., 0.999) against the state-of-the-art DNN backdoor attacks.

****

Group Equivariant Generative Adversarial Networks

Neel Dey,Antong Chen,Soheil Ghafurian

Recent improvements in generative adversarial visual synthesis incorporate real and fake image transformation in a self-supervised setting, leading to increased stability and perceptual fidelity. However, these approaches typically involve image augmentations via additional regularizers in the GAN objective and thus spend valuable network capacity towards approximating transformation equivariance instead of their desired task. In this work, we explicitly incorporate inductive symmetry priors into the network architectures via group-equivariant convolutional networks. Group-convolutions have higher expressive power with fewer samples and lead to better gradient feedback between generator and discriminator. We show that group-equivariance integrates seamlessly with recent techniques for GAN training across regularizers, architectures, and loss functions. We demonstrate the utility of our methods for conditional synthesis by improving generation in the limited data regime across symmetric imaging datasets and even find benefits for natural images with preferred orientation.

****

InfoBERT: Improving Robustness of Language Models from An Information Theoretic Perspective

Boxin Wang,Shuohang Wang,Yu Cheng,Zhe Gan,Ruoxi Jia,Bo Li,Jingjing Liu

Large-scale language models such as BERT have achieved state-of-the-art performance across a wide range of NLP tasks. Recent studies, however, show that such BERT-based models are vulnerable facing the threats of textual adversarial attacks. We aim to address this problem from an information-theoretic perspective, and propose InfoBERT, a novel learning framework for robust ■ne-tuning of pre-trained language models. InfoBERT contains two mutual-information-based regularizers for model training: (i) an Information Bottleneck regularizer, which suppresses noisy mutual information between the input and the feature representation; and (ii) a Robust Feature regularizer, which increases the mutual information between local robust features and global features. We provide a principled way to theoretically analyze and improve the robustness of representation learning for language models in both standard and adversarial training. Extensive experiments demonstrate that InfoBERT achieves state-of-the-art robust accuracy over several adversarial datasets on Natural Language Inference (NLI) and Question Answering (QA) tasks.

Our code is available at https://github.com/AI-secure/InfoBERT.

****

Sparse Linear Networks with a Fixed Butterfly Structure: Theory and Practice

Nir Ailon,Omer Leibovitch,Vineet Sreedharan Nair

A butterfly network consists of logarithmically many layers, each with a linear number of non-zero weights (pre-specified). The fast Johnson-Lindenstrauss transform (FJLT) can be represented as a butterfly network followed by a random projection to a subset of the coordinates. Moreover, a random matrix based on FJLT with high probability approximates the action of any matrix on a vector. Motivated by these facts, we propose to replace a dense linear layer in any neural network by an architecture based on the butterfly network. The proposed architecture significantly improves upon the quadratic number of weights required in a standard dense layer to nearly linear with little compromise in expressibility of the resulting operator. In a collection of wide variety of experiments, including supervised prediction on both the NLP and vision data, we show that this not only produces results that match and often outperform existing well-known architectures, but it also offers faster training and prediction in deployment. To understand the optimization problems posed by neural networks with a butterfly network, we study the optimization landscape of the encoder-decoder network, where the encoder is replaced by a butterfly network followed by a dense linear layer in smaller dimension. Theoretical result presented in the paper explain why the training speed and outcome are not compromised by our proposed approach. Empirically we demonstrate that the network performs as well as the encoder-decoder network.
****

Random Feature Attention
Hao Peng,Nikolaos Pappas,Dani Yogatama,Roy Schwartz,Noah Smith,Lingpeng Kong
Transformers are state-of-the-art models for a variety of sequence modeling tasks. At their core is an attention function which models pairwise interactions between the inputs at every timestep. While attention is powerful, it does not scale efficiently to long sequences due to its quadratic time and space complexity in the sequence length. We propose RFA, a linear time and space attention that uses random feature methods to approximate the softmax function, and explore its application in transformers. RFA can be used as a drop-in replacement for conventional softmax attention and offers a straightforward way of learning with recency bias through an optional gating mechanism. Experiments on language modeling and machine translation demonstrate that RFA achieves similar or better performance compared to strong transformer baselines. In the machine translation experiment, RFA decodes twice as fast as a vanilla transformer. Compared to existing efficient transformer variants, RFA is competitive in terms of both accuracy and efficiency on three long text classification datasets. Our analysis shows that RFA's efficiency gains are especially notable on long sequences, suggesting that RFA will be particularly useful in tasks that require working with large inputs, fast decoding speed, or low memory footprints.
****

A Spectral Perspective on Deep Supervised Community Detection
Nathan Grinsztajn,Philippe Preux,Edouard Oyallon
In this work, we study the behavior of standard models for community detection under spectral manipulations. Through various ablation experiments, we evaluate the impact of bandpass filtering on the numerical performances of a GCN: we empirically show that most of the necessary and used information for nodes classification is contained in the low-frequency domain, and thus contrary to Euclidean graph (e.g., images), high-frequencies are less crucial to community detection. In particular, it is possible to obtain accuracies at a state-of-the-art level with simple classifiers that rely only on a few low frequencies: this is surprising because contrary to GCNs, no cascade of filtering along the graph structure is involved and it indicates that the important spectral components for the supervised community detection task are essentially in the low-frequency domain.
****

Dropout's Dream Land: Generalization from Learned Simulators to Reality
Zac Wellmer,James Kwok
A World Model is a generative model used to simulate an environment. World Models have proven capable of learning spatial and temporal representations of Reinforcement Learning environments. In some cases, a World Model offers an agent the opportunity to learn entirely inside of its own dream environment. In this work

we explore improving the generalization capabilities from dream environments to reality (Dream2Real). We present a general approach to improve a controller's ab ility to transfer from a neural network dream environment to reality at little a dditional cost. These improvements are gained by drawing on inspiration from dom ain randomization, where the basic idea is to randomize as much of a simulator a s possible without fundamentally changing the task at hand. Generally, domain ra ndomization assumes access to a pre-built simulator with configurable parameters but oftentimes this is not available. By training the World Model using dropout , the dream environment is capable of creating a nearly infinite number of \text it{different} dream environments. Our experimental results show that Dropout's D ream Land is an effective technique to bridge the reality gap between dream envi ronments and reality. Furthermore, we additionally perform an extensive set of a blation studies.
****

Attention-driven Robotic Manipulation
Stephen James,Andrew Davison
Despite the success of reinforcement learning methods, they have yet to have the ir breakthrough moment when applied to a broad range of robotic manipulation tas ks. This is partly due to the fact that reinforcement learning algorithms are no toriously difficult and time consuming to train, which is exacerbated when train ing from images rather than full-state inputs. As humans perform manipulation ta sks, our eyes closely monitor every step of the process with our gaze focusing s equentially on the objects being manipulated. With this in mind, we present our Attention-driven Robotic Manipulation (ARM) algorithm, which is a general manipu lation algorithm that can be applied to a range of real-world sparse-rewarded ta sks without any prior task knowledge. ARM splits the complex task of manipulatio n into a 3 stage pipeline: (1) a Q-attention agent extracts interesting pixel lo cations from RGB and point cloud inputs, (2) a next-best pose agent that accepts crops from the Q-attention agent and outputs poses, and (3) a control agent tha t takes the goal pose and outputs joint actions. We show that current state-of-t he-art reinforcement learning algorithms catastrophically fail on a range of RLB ench tasks, whilst ARM is successful within a few hours.
****

Perfect density models cannot guarantee anomaly detection
Charline Le Lan,Laurent Dinh
Thanks to the tractability of their likelihood, some deep generative models show promise for seemingly straightforward but important applications like anomaly d etection, uncertainty estimation, and active learning. However, the likelihood v alues empirically attributed to anomalies conflict with the expectations these p roposed applications suggest. In this paper, we take a closer look at the behavi or of distribution densities and show that these quantities carry less meaningfu l information than previously thought, beyond estimation issues or the curse of dimensionality. We conclude that the use of these likelihoods for out-of-distrib ution detection relies on strong and implicit hypotheses and highlight the neces sity of explicitly formulating these assumptions for reliable anomaly detection.
****

Using latent space regression to analyze and leverage compositionality in GANs
Lucy Chai,Jonas Wulff,Phillip Isola
In recent years, Generative Adversarial Networks have become ubiquitous in both research and public perception, but how GANs convert an unstructured latent code to a high quality output is still an open question. In this work, we investigat e regression into the latent space as a probe to understand the compositional pr operties of GANs. We find that combining the regressor and a pretrained generat or provides a strong image prior, allowing us to create composite images from a collage of random image parts at inference time while maintaining global consist ency. To compare compositional properties across different generators, we measur e the trade-offs between reconstruction of the unrealistic input and image quali ty of the regenerated samples. We find that the regression approach enables more localized editing of individual image parts compared to direct editing in the l atent space, and we conduct experiments to quantify this independence effect. Ou

r method is agnostic to the semantics of edits, and does not require labels or p
redefined concepts during training. Beyond image composition, our method extends
 to a number of related applications, such as image inpainting or example-based
image editing, which we demonstrate on several GANs and datasets, and
because it uses only a single forward pass, it can operate in real-time. Code is
 available on our project page: https://chail.github.io/latent-composition/.
****

Learning to Dynamically Select Between Reward Shaping Signals
Alexander Politowicz,Bing Liu
Reinforcement learning (RL) algorithms often have the limitation of sample compl
exity. Previous research has shown that the reliance on large amounts of experie
nce can be mitigated through the presence of additional feedback. Automatic rewa
rd shaping is one approach to solving this problem, using automatic identificati
on and modulation of shaping reward signals that are more informative about how
agents should behave in any given scenario to learn and adapt faster. However, a
utomatic reward shaping is still very challenging. To better study it, we break
it down into two separate sub-problems: learning shaping reward signals in an ap
plication and learning how the signals can be adaptively used to provide a singl
e reward feedback in the RL learning process. This paper focuses on the latter s
ub-problem. Unlike existing research, which tries to learn one shaping reward fu
nction from shaping signals, the proposed method learns to dynamically select th
e right reward signal to apply at each state, which is considerably more flexibl
e. We further show that using an online strategy that seeks to match the learned
 shaping feedback with optimal value differences can lead to effective reward sh
aping and accelerated learning. The proposed ideas are verified through experime
nts in a variety of environments using different shaping reward paradigms.
****

Go with the flow: Adaptive control for Neural ODEs
Mathieu Chalvidal,Matthew Ricci,Rufin VanRullen,Thomas Serre
Despite their elegant formulation and lightweight memory cost, neural ordinary d
ifferential equations (NODEs) suffer from known representational limitations. In
 particular, the single flow learned by NODEs cannot express all homeomorphisms
from a given data space to itself, and their static weight parameterization rest
ricts the type of functions they can learn compared to discrete architectures wi
th layer-dependent weights. Here, we describe a new module called neurally-contr
olled ODE (N-CODE) designed to improve the expressivity of NODEs. The parameters
 of N-CODE modules are dynamic variables governed by a trainable map from initia
l or current activation state, resulting in forms of open-loop and closed-loop c
ontrol, respectively. A single module is sufficient for learning a distribution
on non-autonomous flows that adaptively drive neural representations. We provide
 theoretical and empirical evidence that N-CODE circumvents limitations of previ
ous NODEs models and show how increased model expressivity manifests in several
supervised and unsupervised learning problems. These favorable empirical results
 indicate the potential of using data- and activity-dependent plasticity in neur
al networks across numerous domains.
****

A Learning Theoretic Perspective on Local Explainability
Jeffrey Li,Vaishnavh Nagarajan,Gregory Plumb,Ameet Talwalkar
In this paper, we explore connections between interpretable machine learning and
 learning theory through the lens of local approximation explanations. First, we
 tackle the traditional problem of performance generalization and bound the test
-time predictive accuracy of a model using a notion of how locally explainable i
t is. Second, we explore the novel problem of explanation generalization which
is an important concern for a growing class of finite sample-based local approxi
mation explanations. Finally, we validate our theoretical results empirically an
d show that they reflect what can be seen in practice.
****

The Bures Metric for Taming Mode Collapse in Generative Adversarial Networks
Hannes De Meulemeester,Joachim Schreurs,Michaël Fanuel,Bart De Moor,Johan Suyken
s

Generative Adversarial Networks (GANs) are performant generative methods yieldin
g high-quality samples. However, under certain circumstances, the training of GA
Ns can lead to mode collapse or mode dropping, i.e. the generative models not be
ing able to sample from the entire probability distribution. To address this pro
blem, we use the last layer of the discriminator as a feature map to study the d
istribution of the real and the fake data. During training, we propose to match
the real batch diversity to the fake batch diversity by using the Bures distance
 between covariance matrices in feature space. The computation of the Bures dist
ance can be conveniently done in either feature space or kernel space in terms o
f the covariance and kernel matrix respectively. We observe that diversity match
ing reduces mode collapse substantially and has a positive effect on the sample
quality. On the practical side, a very simple training procedure, that does not
require additional hyperparameter tuning, is proposed and assessed on several da
tasets.
****

FERMI: Fair Empirical Risk Minimization via Exponential Rényi Mutual Information
Rakesh Pavan,Andrew Lowy,Sina Baharlouei,Meisam Razaviyayn,Ahmad Beirami
Several notions of fairness, such as demographic parity and equal opportunity, a
re defined based on statistical independence between a predicted target and a se
nsitive attribute. In machine learning applications, however, the data distribut
ion is unknown to the learner and statistical independence is not verifiable. He
nce, the learner could only resort to empirical evaluation of the degree of fair
ness violation. Many fairness violation notions are defined as a divergence/dist
ance between the joint distribution of the target and sensitive attributes and
 the Kronecker product of their marginals, such as \Renyi correlation, mutual i
nformation, L_∞ distance, to name a few.
 In this paper, we propose another notion of fairness violation, called Exponen
tial R\'enyi Mutual Information (ERMI) between sensitive attributes and the pred
icted target. We show that ERMI is a strong fairness violation notion in the sen
se that it provides an upper bound guarantee on all of the aforementioned notion
s of fairness violation. We also propose the Fair Empirical Risk Minimization vi
a ERMI regularization framework, called FERMI. Whereas existing in-processing fa
irness algorithms are deterministic, we provide a stochastic optimization method
 for solving FERMI that is amenable to large-scale problems. In addition, we pro
vide a batch (deterministic) method to solve FERMI. Both of our proposed algorit
hms come with theoretical convergence guarantees. Our experiments show that FERM
I achieves the most favorable tradeoffs between fairness violation and accuracy
on test data across different problem setups, even when fairness violation is me
asured in notions other than ERMI.
****

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
Alexey Dosovitskiy,Lucas Beyer,Alexander Kolesnikov,Dirk Weissenborn,Xiaohua Zha
i,Thomas Unterthiner,Mostafa Dehghani,Matthias Minderer,Georg Heigold,Sylvain Ge
lly,Jakob Uszkoreit,Neil Houlsby
While the Transformer architecture has become the de-facto standard for natural
language processing tasks, its applications to computer vision remain limited. I
n vision, attention is either applied in conjunction with convolutional networks
, or used to replace certain components of convolutional networks while keeping
their overall structure in place. We show that this reliance on CNNs is not nece
ssary and a pure transformer applied directly to sequences of image patches can
perform very well on image classification tasks. When pre-trained on large amoun
ts of data and transferred to multiple mid-sized or small image recognition benc
hmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excel
lent results compared to state-of-the-art convolutional networks while requiring
 substantially fewer computational resources to train.
****

Jointly-Trained State-Action Embedding for Efficient Reinforcement Learning
Paul Julian Pritz,Liang Ma,Kin Leung
While reinforcement learning has achieved considerable successes in recent years
, state-of-the-art models are often still limited by the size of state and actio

n spaces. Model-free reinforcement learning approaches use some form of state representations and the latest work has explored embedding techniques for actions, both with the aim of achieving better generalization and applicability. However, these approaches consider only states or actions, ignoring the interaction between them when generating embedded representations. In this work, we propose a new approach for jointly learning embeddings for states and actions that combines aspects of model-free and model-based reinforcement learning, which can be applied in both discrete and continuous domains. Specifically, we use a model of the environment to obtain embeddings for states and actions and present a generic architecture that uses these to learn a policy. In this way, the embedded representations obtained via our approach enable better generalization over both states and actions by capturing similarities in the embedding spaces. Evaluations of our approach on several gaming, robotic control, and recommender systems show it significantly outperforms state-of-the-art models in both discrete/continuous domains with large state/action spaces, thus confirming its efficacy and the overall superior performance.
****

A Robust Fuel Optimization Strategy For Hybrid Electric Vehicles: A Deep Reinforcement Learning Based Continuous Time Design Approach
Nilanjan Mukherjee,Sudeshna Sarkar
This paper deals with the fuel optimization problem for hybrid electric vehicles in reinforcement learning framework. Firstly, considering the hybrid electric vehicle as a completely observable non-linear system with uncertain dynamics, we solve an open-loop deterministic optimization problem. This is followed by the design of a deep reinforcement learning based optimal controller for the non-linear system using concurrent learning based system identifier such that the actual states and the control policy are able to track the optimal trajectory and optimal policy, autonomously even in the presence of external disturbances, modeling errors, uncertainties and noise and signigicantly reducing the computational complexity at the same time, which is in sharp contrast to the conventional methods like PID and Model Predictive Control (MPC) as well as traditional RL approaches like ADP, DDP and DQN that mostly depend on a set of pre-defined rules and provide sub-optimal solutions under similar conditions. The low value of the H-infinity (H_{∞}) performance index of the proposed optimization algorithm addresses the robustness issue. The optimization technique thus proposed is compared with the traditional fuel optimization strategies for hybrid electric vehicles to illustate the efficacy of the proposed method.
****

CopulaGNN: Towards Integrating Representational and Correlational Roles of Graphs in Graph Neural Networks
Jiaqi Ma,Bo Chang,Xuefei Zhang,Qiaozhu Mei
Graph-structured data are ubiquitous. However, graphs encode diverse types of information and thus play different roles in data representation. In this paper, we distinguish the \textit{representational} and the \textit{correlational} roles played by the graphs in node-level prediction tasks, and we investigate how Graph Neural Network (GNN) models can effectively leverage both types of information. Conceptually, the representational information provides guidance for the model to construct better node features; while the correlational information indicates the correlation between node outcomes conditional on node features. Through a simulation study, we find that many popular GNN models are incapable of effectively utilizing the correlational information. By leveraging the idea of the copula, a principled way to describe the dependence among multivariate random variables, we offer a general solution. The proposed Copula Graph Neural Network (CopulaGNN) can take a wide range of GNN models as base models and utilize both representational and correlational information stored in the graphs. Experimental results on two types of regression tasks verify the effectiveness of the proposed method.
****

Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability
Jeremy Cohen,Simran Kaur,Yuanzhi Li,J Zico Kolter,Ameet Talwalkar

We empirically demonstrate that full-batch gradient descent on neural network tr
aining objectives typically operates in a regime we call the Edge of Stability.
In this regime, the maximum eigenvalue of the training loss Hessian hovers just
above the value $2 / \text{(step size)}$, and the training loss behaves non-mono
tonically over short timescales, yet consistently decreases over long timescales
. Since this behavior is inconsistent with several widespread presumptions in th
e field of optimization, our findings raise questions as to whether these presum
ptions are relevant to neural network training. We hope that our findings will i
nspire future efforts aimed at rigorously understanding optimization at the Edge
 of Stability.
****

AutoLRS: Automatic Learning-Rate Schedule by Bayesian Optimization on the Fly
Yuchen Jin,Tianyi Zhou,Liangyu Zhao,Yibo Zhu,Chuanxiong Guo,Marco Canini,Arvind
Krishnamurthy
The learning rate (LR) schedule is one of the most important hyper-parameters ne
eding careful tuning in training DNNs. However, it is also one of the least auto
mated parts of machine learning systems and usually costs significant manual eff
ort and computing. Though there are pre-defined LR schedules and optimizers with
 adaptive LR, they introduce new hyperparameters that need to be tuned separatel
y for different tasks/datasets. In this paper, we consider the question: Can we
automatically tune the LR over the course of training without human involvement?
 We propose an efficient method, AutoLRS, which automatically optimizes the LR f
or each training stage by modeling training dynamics. AutoLRS aims to find an LR
 that minimizes the validation loss, every τ steps. We formulate it as blac
k-box optimization and solve it by Bayesian optimization (BO). However, collecti
ng training instances for BO requires a system to evaluate each LR queried by BO
's acquisition function for τ steps, which is prohibitively expensive in pr
actice. Instead, we apply each candidate LR for only $\tau'\ll\tau$ steps and tr
ain an exponential model to predict the validation loss after τ steps. This
 mutual-training process between BO and the exponential model allows us to bound
 the number of training steps invested in the BO search. We demonstrate the adva
ntages and the generality of AutoLRS through extensive experiments of training D
NNs from diverse domains and using different optimizers. The LR schedules auto-g
enerated by AutoLRS leads to a speedup of $1.22\times$, $1.43\times$, and $1.5\t
imes$ when training ResNet-50, Transformer, and BERT, respectively, compared to
the LR schedules in their original papers, and an average speedup of $1.31\times
$ over state-of-the-art highly tuned LR schedules.
****

Synthesizer: Rethinking Self-Attention for Transformer Models
Yi Tay,Dara Bahri,Donald Metzler,Da-Cheng Juan,Zhe Zhao,Che Zheng
The dot product self-attention is known to be central and indispensable to state
-of-the-art Transformer models. But is it really required? This paper investigat
es the true importance and contribution of the dot product-based self-attention
mechanism on the performance of Transformer models. Via extensive experiments, w
e find that (1) random alignment matrices surprisingly perform quite competitive
ly and (2) learning attention weights from token-token (query-key) interactions
is useful but not that important after all. To this end, we propose \textsc{Synt
hesizer}, a model that learns synthetic attention weights without token-token in
teractions. In our experiments, we first show that simple Synthesizers achieve h
ighly competitive performance when compared against vanilla Transformer models a
cross a range of tasks, including machine translation, language modeling, text g
eneration and GLUE/SuperGLUE benchmarks. When composed with dot product attentio
n, we find that Synthesizers consistently outperform Transformers. Moreover, we
conduct additional comparisons of Synthesizers against Dynamic Convolutions, sho
wing that simple Random Synthesizer is not only 60% faster but also improves
perplexity by a relative 3.5%. Finally, we show that simple factorized Synthe
sizers can outperform Linformers on encoding only tasks.
****

On Episodes, Prototypical Networks, and Few-Shot Learning
Steinar Laenen,Luca Bertinetto

Episodic learning is a popular practice among researchers and practitioners interested in few-shot learning. It consists of organising training in a series of learning problems, each relying on small “support” and “query” sets to mimic the few-shot circumstances encountered during evaluation.
In this paper, we investigate the usefulness of episodic learning in Prototypical Networks, one of the most popular algorithms making use of this practice.
Surprisingly, in our experiments we found that, for Prototypical Networks, it is detrimental to use the episodic learning strategy of separating training samples between support and query set, as it is
a data-inefficient way to exploit training batches. This “non-episodic” version of Prototypical Networks, which corresponds to the classic Neighbourhood Component Analysis, reliably improves over its episodic counterpart in multiple datasets, achieving an accuracy that is competitive with the state-of-the-art, despite being extremely simple.
****
Autoregressive Dynamics Models for Offline Policy Evaluation and Optimization
Michael R Zhang,Thomas Paine,Ofir Nachum,Cosmin Paduraru,George Tucker,ziyu wang,Mohammad Norouzi
Standard dynamics models for continuous control make use of feedforward computation to predict the conditional distribution of next state and reward given current state and action using a multivariate Gaussian with a diagonal covariance structure. This modeling choice assumes that different dimensions of the next state and reward are conditionally independent given the current state and action and may be driven by the fact that fully observable physics-based simulation environments entail deterministic transition dynamics. In this paper, we challenge this conditional independence assumption and propose a family of expressive autoregressive dynamics models that generate different dimensions of the next state and reward sequentially conditioned on previous dimensions. We demonstrate that autoregressive dynamics models indeed outperform standard feedforward models in log-likelihood on heldout transitions. Furthermore, we compare different model-based and model-free off-policy evaluation (OPE) methods on RL Unplugged, a suite of offline MuJoCo datasets, and find that autoregressive dynamics models consistently outperform all baselines, achieving a new state-of-the-art. Finally, we show that autoregressive dynamics models are useful for offline policy optimization by serving as a way to enrich the replay buffer through data augmentation and improving performance using model-based planning.

****
On the role of planning in model-based deep reinforcement learning
Jessica B Hamrick,Abram L. Friesen,Feryal Behbahani,Arthur Guez,Fabio Viola,Sims Witherspoon,Thomas Anthony,Lars Holger Buesing,Petar Veli■kovi■,Theophane Weber
Model-based planning is often thought to be necessary for deep, careful reasoning and generalization in artificial agents. While recent successes of model-based reinforcement learning (MBRL) with deep function approximation have strengthened this hypothesis, the resulting diversity of model-based methods has also made it difficult to track which components drive success and why. In this paper, we seek to disentangle the contributions of recent methods by focusing on three questions: (1) How does planning benefit MBRL agents? (2) Within planning, what choices drive performance? (3) To what extent does planning improve generalization? To answer these questions, we study the performance of MuZero (Schrittwieser et al., 2019), a state-of-the-art MBRL algorithm with strong connections and overlapping components with many other MBRL algorithms. We perform a number of interventions and ablations of MuZero across a wide range of environments, including control tasks, Atari, and 9x9 Go. Our results suggest the following: (1) Planning is most useful in the learning process, both for policy updates and for providing a more useful data distribution. (2) Using shallow trees with simple Monte-Carlo rollouts is as performant as more complex methods, except in the most difficult reasoning tasks. (3) Planning alone is insufficient to drive strong generalization. These results indicate where and how to utilize planning in reinforcement learning settings, and highlight a number of open questions for future MBRL re

search.
****
A Flexible Framework for Discovering Novel Categories with Contrastive Learning
Xuhui Jia,Kai Han,Yukun Zhu,Bradley Green
This paper studies the problem of novel category discovery on single- and multi-modal data with labels from different but relevant categories. We present a generic, end-to-end framework to jointly learn a reliable representation and assign clusters to unlabelled data. To avoid over-fitting the learnt embedding to labelled data, we take inspiration from self-supervised representation learning by noise-contrastive estimation and extend it to jointly handle labelled and unlabelled data. In particular, we proposed using category discrimination on labelled data and cross-modal discrimination on multi-modal data to augment instance discrimination used in conventional contrastive learning approaches. We further introduce Winner-Take-All (WTA) hashing algorithm on the shared representation space to generate pairwise pseudo labels for unlabelled data to better predict cluster assignments. We thoroughly evaluate our framework on large-scale multi-modal video benchmarks Kinetics-400 and VGG-Sound, and image benchmarks CIFAR10, CIFAR100 and ImageNet, obtaining state-of-the-art results.
****
K-PLUG: KNOWLEDGE-INJECTED PRE-TRAINED LANGUAGE MODEL FOR NATURAL LANGUAGE UNDERSTANDING AND GENERATION
Song Xu,Haoran Li,Peng Yuan,Yujia Wang,Youzheng Wu,Xiaodong He,Ying Liu,Bowen Zhou
Existing pre-trained language models (PLMs) have demonstrated the effectiveness of self-supervised learning for a broad range of natural language processing (NLP) tasks. However, most of them are not explicitly aware of domain-specific knowledge, which is essential for downstream tasks in many domains, such as tasks in e-commerce scenarios. In this paper, we propose K-PLUG, a knowledge-injected pre-trained language model based on the encoder-decoder transformer that can be transferred to both natural language understanding and generation tasks. We verify our method in a diverse range of e-commerce scenarios that require domain-specific knowledge. Specifically, we propose five knowledge-aware self-supervised pre-training objectives to formulate the learning of domain-specific knowledge, including e-commerce domain-specific knowledge-bases, aspects of product entities, categories of product entities, and unique selling propositions of product entities. K-PLUG achieves new state-of-the-art results on a suite of domain-specific NLP tasks, including product knowledge base completion, abstractive product summarization, and multi-turn dialogue, significantly outperforms baselines across the board, which demonstrates that the proposed method effectively learns a diverse set of domain-specific knowledge for both language understanding and generation tasks. The code, data, and models will be publicly available.

****
Reinforcement Learning Based Asymmetrical DNN Modularization for Optimal Loading
Brijraj Singh,Yash Jain,Mayukh Das,Praveen Doreswamy Naidu
Latency of DNN (Deep Neural Network) based prediction is the summation of model loading latency and inference latency. Model loading latency affects the first response from the applications, whereas inference latency affects the subsequent responses. As model loading latency is directly proportional to the model size, this work aims at improving the response time of an intelligent app by reducing the loading latency. The speedup is gained by asymmetrically modularizing the given DNN model among several small child models and loading them in parallel. The decision about number of feasible child models and their corresponding split positions are taken care by reinforcement learning unit (RLU). RLU takes into account the available hardware resources on-device and provides the best splitting index k and their positions $\vec{p}$ specific to the DNN model and device, where $\vec{p}=(p_1, p_2, ..., p_k)$ and p_i is the end position of i^{th} child: M_i. The proposed method has shown significant loading improvement (up to 7X) on popular DNNs, used for camera use-case. The proposed method can be used to speed up the app response. Along with that RLU driven approach facilitates

for On-device personalization by separating one module only with trainable layer
s and loading that particular module while training on-device.
****
Adversarial Privacy Preservation in MRI Scans of the Brain
Lennart Alexander Van der Goten,Tobias Hepp,Zeynep Akata,Kevin Smith
De-identification of magnetic resonance imagery (MRI) is intrinsically difficult
since, even with all metadata removed, a person's face can easily be rendered a
nd matched against a database. Existing de-identification methods tackle this ta
sk by obfuscating or removing parts of the face, but they either fail to reliabl
y hide the patient's identity or they remove so much information that they adver
sely affect further analyses in the 3D space surrounding the face. In this work,
we describe a new class of MRI de-identification techniques that remodel privac
y-sensitive facial features as opposed to removing them.To accomplish this, we p
ropose a conditional, multi-scale, 3D GAN architecture that takes a patient's MR
I scan as input and generates a 3D volume in which the brain is not modified but
the face has been de-identified. Compared to the classical removal-based techni
ques, our deep learning framework preserves privacy more reliably without advers
ely affecting downstream medical analyses on the brain, including segmentation a
nd age prediction.
****
Model-centric data manifold: the data through the eyes of the model
Luca Grementieri,Rita Fioresi
We discover that deep ReLU neural network classifiers can see a low-dimensional
Riemannian manifold structure on data. Such structure comes via the local data m
atrix, a variation of the Fisher information matrix, where the role of the model
parameters is taken by the data variables. We obtain a foliation of the data do
main and we show that the dataset on which the model is trained lies on a leaf,
the data leaf, whose dimension is bounded by the number of classification labels
. We validate our results with some experiments with the MNIST dataset: paths on
the data leaf connect valid images, while other leaves cover noisy images.

****
How to Avoid Being Eaten by a Grue: Structured Exploration Strategies for Textua
l Worlds
Prithviraj Ammanabrolu,Ethan Tien,Matthew Hausknecht,Mark Riedl
Text-based games are long puzzles or quests, characterized by a sequence of spar
se and potentially deceptive rewards. They provide an ideal platform to develop
agents that perceive and act upon the world using a combinatorially sized natura
l language state-action space. Standard Reinforcement Learning agents are poorly
equipped to effectively explore such spaces and often struggle to overcome bott
lenecks---states that agents are unable to pass through simply because they do n
ot see the right action sequence enough times to be sufficiently reinforced. We
introduce Q*BERT, an agent that learns to build a knowledge graph of the world b
y answering questions, which leads to greater sample efficiency. To overcome bot
tlenecks, we further introduce MC!Q*BERT an agent that uses an knowledge-graph-b
ased intrinsic motivation to detect bottlenecks and a novel exploration strategy
to efficiently learn a chain of policy modules to overcome them. We present an
ablation study and results demonstrating how our method outperforms the current
state-of-the-art on nine text games, including the popular game, Zork, where, fo
r the first time, a learning agent gets past the bottleneck where the player is
eaten by a Grue.
****
Zero-Cost Proxies for Lightweight NAS
Mohamed S Abdelfattah,Abhinav Mehrotra,■ukasz Dudziak,Nicholas Donald Lane
Neural Architecture Search (NAS) is quickly becoming the standard methodology to
design neural network models. However, NAS is typically compute-intensive becau
se multiple models need to be evaluated before choosing the best one. To reduce
the computational power and time needed, a proxy task is often used for evaluati
ng each model instead of full training. In this paper, we evaluate conventional
reduced-training proxies and quantify how well they preserve ranking between neu

ral network models during search when compared with the rankings produced by final trained accuracy. We propose a series of zero-cost proxies, based on recent pruning literature, that use just a single minibatch of training data to compute a model's score. Our zero-cost proxies use 3 orders of magnitude less computation but can match and even outperform conventional proxies. For example, Spearman's rank correlation coefficient between final validation accuracy and our best zero-cost proxy on NAS-Bench-201 is 0.82, compared to 0.61 for EcoNAS (a recently proposed reduced-training proxy). Finally, we use these zero-cost proxies to enhance existing NAS search algorithms such as random search, reinforcement learning, evolutionary search and predictor-based search. For all search methodologies and across three different NAS datasets, we are able to significantly improve sample efficiency, and thereby decrease computation, by using our zero-cost proxies. For example on NAS-Bench-101, we achieved the same accuracy 4$\times$ quicker than the best previous result. Our code is made public at: https://github.com/mohsaied/zero-cost-nas.
****

Revisiting Graph Neural Networks for Link Prediction
Muhan Zhang,Pan Li,Yinglong Xia,Kai Wang,Long Jin
Graph neural networks (GNNs) have achieved great success in recent years. Three most common applications include node classification, link prediction, and graph classification. While there is rich literature on node classification and graph classification, GNNs for link prediction is relatively less studied and less understood. Two representative classes of methods exist: GAE and SEAL. GAE (Graph Autoencoder) first uses a GNN to learn node embeddings for all nodes, and then aggregates the embeddings of the source and target nodes as their link representation. SEAL extracts a subgraph around the source and target nodes, labels the nodes in the subgraph, and then uses a GNN to learn a link representation from the labeled subgraph. In this paper, we thoroughly discuss the differences between these two classes of methods, and conclude that simply aggregating \textit{node} embeddings does not lead to effective \textit{link} representations, while learning from \textit{properly labeled subgraphs} around links provides highly expressive and generalizable link representations. Experiments on the recent large-scale OGB link prediction datasets show that SEAL has up to 195\% performance gains over GAE methods, achieving new state-of-the-art results on 3 out of 4 datasets.
****

Simplifying Models with Unlabeled Output Data
Sang Michael Xie,Tengyu Ma,Percy Liang
We focus on prediction problems with high-dimensional outputs that are subject to output validity constraints, e.g. a pseudocode-to-code translation task where the code must compile. For these problems, labeled input-output pairs are expensive to obtain, but "unlabeled" outputs, i.e. outputs without corresponding inputs, are freely available and provide information about output validity (e.g. code on GitHub). In this paper, we present predict-and-denoise, a framework that can leverage unlabeled outputs. Specifically, we first train a denoiser to map possibly invalid outputs to valid outputs using synthetic perturbations of the unlabeled outputs. Second, we train a predictor composed with this fixed denoiser. We show theoretically that for a family of functions with a high-dimensional discrete valid output space, composing with a denoiser reduces the complexity of a 2-layer ReLU network needed to represent the function and that this complexity gap can be arbitrarily large. We evaluate the framework empirically on several datasets, including image generation from attributes and pseudocode-to-code translation. On the SPoC pseudocode-to-code dataset, our framework improves the proportion of code outputs that pass all test cases by 3-5% over a baseline Transformer.
****

How to Design Sample and Computationally Efficient VQA Models
Karan Samel,Zelin Zhao,Kuan Wang,Robin Luo,Binghong Chen,Le Song
In multi-modal reasoning tasks, such as visual question answering (VQA), there have been many modeling and training paradigms tested. Previous models propose different methods for the vision and language tasks, but which ones perform the be

st while being sample and computationally efficient? Based on our experiments, w
e find that representing the text as probabilistic programs and images as object
-level scene graphs best satisfy these desiderata. We extend existing models to
leverage these soft programs and scene graphs to train on question answer pairs
in an end-to-end manner. Empirical results demonstrate that this differentiable
end-to-end program executor is able to maintain state-of-the-art accuracy while
being sample and computationally efficient.
****

Contextual HyperNetworks for Novel Feature Adaptation
Angus Lamb,Evgeny Saveliev,Yingzhen Li,Sebastian Tschiatschek,Camilla Longden,Si
mon Woodhead,José Miguel Hernández-Lobato,Richard E Turner,Pashmina Cameron,Chen
g Zhang
While deep learning has obtained state-of-the-art results in many applications,
the adaptation of neural network architectures to incorporate new features remai
ns a research challenge. This issue is particularly severe in online learning se
ttings, where new features are added continually with few or no associated obser
vations. As such, methods for adapting neural networks to novel features which a
re both time and data-efficient are desired. To address this, we propose the Con
textual HyperNetwork (CHN), which predicts the network weights associated with n
ew features by incorporating information from both existing data as well as the
few observations for the new feature and any associated feature metadata. At pre
diction time, the CHN requires only a single forward pass through a small neural
 network, yielding a significant speed-up when compared to re-training and fine-
tuning approaches. In order to showcase the performance of CHNs, in this work we
 use a CHN to augment a partial variational autoencoder (P-VAE), a flexible deep
 generative model which can impute the values of missing features in sparsely-ob
served data. We show that this system obtains significantly improved performance
 for novel feature adaptation over existing imputation and meta-learning baselin
es across recommender systems, e-learning, and healthcare tasks.
****

Personalized Federated Learning with First Order Model Optimization
Michael Zhang,Karan Sapra,Sanja Fidler,Serena Yeung,Jose M. Alvarez
While federated learning traditionally aims to train a single global model acros
s decentralized local datasets, one model may not always be ideal for all partic
ipating clients. Here we propose an alternative, where each client only federate
s with other relevant clients to obtain a stronger model per client-specific obj
ectives. To achieve this personalization, rather than computing a single model a
verage with constant weights for the entire federation as in traditional FL, we
efficiently calculate optimal weighted model combinations for each client, based
 on figuring out how much a client can benefit from another's model. We do not a
ssume knowledge of any underlying data distributions or client similarities, and
 allow each client to optimize for arbitrary target distributions of interest, e
nabling greater flexibility for personalization. We evaluate and characterize ou
r method on a variety of federated settings, datasets, and degrees of local data
 heterogeneity. Our method outperforms existing alternatives, while also enablin
g new features for personalized FL such as transfer outside of local data distri
butions.
****

Revisiting Explicit Regularization in Neural Networks for Reliable Predictive Pr
obability
Taejong Joo,Uijung Chung
From the statistical learning perspective, complexity control via explicit regul
arization is a necessity for improving the generalization of over-parameterized
models, which deters the memorization of intricate patterns existing only in the
 training data. However, the impressive generalization performance of over-param
eterized neural networks with only implicit regularization challenges the import
ance of explicit regularization. Furthermore, explicit regularization does not p
revent neural networks from memorizing unnatural patterns, such as random labels
. In this work, we revisit the role and importance of explicit regularization me
thods for generalization of the predictive probability, not just the generalizat

ion of the 0-1 loss. Specifically, we analyze the possible cause of the poor predictive probability and identify that regularization of predictive confidence is required during training. We then empirically show that explicit regularization significantly improves the reliability of the predictive probability, which enables better predictive uncertainty representation and prevents the overconfidence problem. Our findings present a new direction to improve the predictive probability quality of deterministic neural networks, which can be an efficient and scalable alternative to Bayesian neural networks and ensemble methods.

****

Deconstructing the Regularization of BatchNorm

Yann Dauphin,Ekin Dogus Cubuk

Batch normalization (BatchNorm) has become a standard technique in deep learning. Its popularity is in no small part due to its often positive effect on generalization. Despite this success, the regularization effect of the technique is still poorly understood. This study aims to decompose BatchNorm into separate mechanisms that are much simpler. We identify three effects of BatchNorm and assess their impact directly with ablations and interventions. Our experiments show that preventing explosive growth at the final layer at initialization and during training can recover a large part of BatchNorm's generalization boost. This regularization mechanism can lift accuracy by 2.9% for Resnet-50 on Imagenet without BatchNorm. We show it is linked to other methods like Dropout and recent initializations like Fixup. Surprisingly, this simple mechanism matches the improvement of 0.9% of the more complex Dropout regularization for the state-of-the-art Efficientnet-B8 model on Imagenet. This demonstrates the underrated effectiveness of simple regularizations and sheds light on directions to further improve generalization for deep nets.

****

DarKnight: A Data Privacy Scheme for Training and Inference of Deep Neural Networks

Hanieh Hashemi,Yongqin Wang,Murali Annavaram

Protecting the privacy of input data is of growing importance as machine learning methods reach new application domains.

In this paper, we provide a unified training and inference framework for large DNNs while protecting input privacy and computation integrity. Our approach called DarKnight uses a novel data blinding strategy using matrix masking to create input obfuscation within a trusted execution environment (TEE). Our rigorous mathematical proof demonstrates that our blinding process provides an information-theoretic privacy guarantee by bounding information leakage. The obfuscated data can then be offloaded to any GPU for accelerating linear operations on blinded data. The results from linear operations on blinded data are decoded before performing non-linear operations within the TEE. This cooperative execution allows DarKnight to exploit the computational power of GPUs to perform linear operations while exploiting TEEs to protect input privacy. We implement DarKnight on an Intel SGX TEE augmented with a GPU to evaluate its performance.

****

No Cost Likelihood Manipulation at Test Time for Making Better Mistakes in Deep Networks

Shyamgopal Karthik,Ameya Prabhu,Puneet K. Dokania,Vineet Gandhi

There has been increasing interest in building deep hierarchy-aware classifiers that aim to quantify and reduce the severity of mistakes, and not just reduce the number of errors. The idea is to exploit the label hierarchy (e.g., the WordNet ontology) and consider graph distances as a proxy for mistake severity. Surprisingly, on examining mistake-severity distributions of the top-1 prediction, we find that current state-of-the-art hierarchy-aware deep classifiers do not always show practical improvement over the standard cross-entropy baseline in making better mistakes. The reason for the reduction in average mistake-severity can be attributed to the increase in low-severity mistakes, which may also explain the noticeable drop in their accuracy. To this end, we use the classical Conditional Risk Minimization (CRM) framework for hierarchy-aware classification. Given a

cost matrix and a reliable estimate of likelihoods (obtained from a trained network), CRM simply amends mistakes at inference time; it needs no extra hyperparameters and requires adding just a few lines of code to the standard cross-entropy baseline. It significantly outperforms the state-of-the-art and consistently obtains large reductions in the average hierarchical distance of top-k predictions across datasets, with very little loss in accuracy. CRM, because of its simplicity, can be used with any off-the-shelf trained model that provides reliable likelihood estimates.
****

Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design

Yue Cao,Payel Das,Pin-Yu Chen,Vijil Chenthamarakshan,Igor Melnyk,Yang Shen

Designing novel protein sequences consistent with a desired 3D structure or fold, often referred to as the inverse protein folding problem, is a central, but non-trivial, task in protein engineering. It has a wide range of applications in energy, biomedicine, and materials science. However, challenges exist due to the complex sequence-fold relationship and difficulties associated with modeling 3D folds. To overcome these challenges, we propose Fold2Seq, a novel transformer-based generative framework for designing protein sequences conditioned on a specific fold. Our model learns a fold embedding from the density of the secondary structural elements in 3D voxels, and then models the complex sequence-structure relationship by learning a joint sequence-fold embedding. Experiments on high-resolution, complete, and single-structure test set demonstrate improved performance of Fold2Seq in terms of speed and reliability for sequence design, compared to existing baselines including the state-of-the-art RosettaDesign and other neural net-based approaches. The unique advantages of fold-based Fold2Seq becomes more evident on diverse real-world test sets comprised of low-resolution, incomplete, or ensemble structures, in comparison to a structure-based model.
****

Intrinsic-Extrinsic Convolution and Pooling for Learning on 3D Protein Structures

Pedro Hermosilla,Marco Schäfer,Matej Lang,Gloria Fackelmann,Pere-Pau Vázquez,Barbora Kozlikova,Michael Krone,Tobias Ritschel,Timo Ropinski

Proteins perform a large variety of functions in living organisms and thus play a key role in biology. However, commonly used algorithms in protein representation learning were not specifically designed for protein data, and are therefore not able to capture all relevant structural levels of a protein during learning. To fill this gap, we propose two new learning operators, specifically designed to process protein structures. First, we introduce a novel convolution operator that considers the primary, secondary, and tertiary structure of a protein by using n-D convolutions defined on both the Euclidean distance, as well as multiple geodesic distances between the atoms in a multi-graph. Second, we introduce a set of hierarchical pooling operators that enable multi-scale protein analysis. We further evaluate the accuracy of our algorithms on common downstream tasks, where we outperform state-of-the-art protein learning algorithms.
****

Action Guidance: Getting the Best of Sparse Rewards and Shaped Rewards for Real-time Strategy Games

Shengyi Huang,Santiago Ontanon

Training agents using Reinforcement Learning in games with sparse rewards is a challenging problem, since large amounts of exploration are required to retrieve even the first reward. To tackle this problem, a common approach is to use reward shaping to help exploration. However, an important drawback of reward shaping is that agents sometimes learn to optimize the shaped reward instead of the true objective. In this paper, we present a novel technique that we call action guidance that successfully trains agents to eventually optimize the true objective in games with sparse rewards while maintaining most of the sample efficiency that comes with reward shaping. We evaluate our approach in a simplified real-time strategy (RTS) game simulator called μRTS.
****

Generative Language-Grounded Policy in Vision-and-Language Navigation with Bayes ' Rule

Shuhei Kurita,Kyunghyun Cho

Vision-and-language navigation (VLN) is a task in which an agent is embodied in a realistic 3D environment and follows an instruction to reach the goal node. While most of the previous studies have built and investigated a discriminative approach, we notice that there are in fact two possible approaches to building such a VLN agent: discriminative and generative. In this paper, we design and investigate a generative language-grounded policy which uses a language model to compute the distribution over all possible instructions i.e. all possible sequences of vocabulary tokens given action and the transition history. In experiments, we show that the proposed generative approach outperforms the discriminative approach in the Room-2-Room (R2R) and Room-4-Room (R4R) datasets, especially in the unseen environments. We further show that the combination of the generative and discriminative policies achieves close to the state-of-the art results in the R2R dataset, demonstrating that the generative and discriminative policies capture the different aspects of VLN.

****

Evaluating Agents Without Rewards

Brendon Matusch,Jimmy Ba,Danijar Hafner

Reinforcement learning has enabled agents to solve challenging control tasks from raw image inputs. However, manually crafting reward functions can be time consuming, expensive, and prone to human error. Competing objectives have been proposed for agents to learn without external supervision, such as artificial input entropy, information gain, and empowerment. Estimating these objectives can be challenging and it remains unclear how well they reflect task rewards or human behavior. We study these objectives across seven agents and three Atari games. Retrospectively computing the objectives from the agent's lifetime of experience simplifies accurate estimation. We find that all three objectives correlate more strongly with a human behavior similarity metric than with task reward. Moreover, input entropy and information gain both correlate more strongly with human similarity than task reward does.

****

Learning a Latent Search Space for Routing Problems using Variational Autoencoders

André Hottung,Bhanu Bhandari,Kevin Tierney

Methods for automatically learning to solve routing problems are rapidly improving in performance. While most of these methods excel at generating solutions quickly, they are unable to effectively utilize longer run times because they lack a sophisticated search component. We present a learning-based optimization approach that allows a guided search in the distribution of high-quality solutions for a problem instance. More precisely, our method uses a conditional variational autoencoder that learns to map points in a continuous (latent) search space to high-quality, instance-specific routing problem solutions. The learned space can then be searched by any unconstrained continuous optimization method. We show that even using a standard differential evolution search strategy our approach is able to outperform existing purely machine learning based approaches.

****

Parallel Training of Deep Networks with Local Updates

Michael Laskin,Luke Metz,Seth Nabarro,Mark Saroufim,Badreddine Noune,Carlo Luschi,Jascha Sohl-Dickstein,Pieter Abbeel

Deep learning models trained on large data sets have been widely successful in both vision and language domains. As state-of-the-art deep learning architectures have continued to grow in parameter count so have the compute budgets and times required to train them, increasing the need for compute-efficient methods that parallelize training. Two common approaches to parallelize the training of deep networks have been data and model parallelism. While useful, data and model parallelism suffer from diminishing returns in terms of compute efficiency for large batch sizes. In this paper, we investigate how to continue scaling compute efficiently beyond the point of diminishing returns for large batches through local

parallelism, a framework which parallelizes training of individual layers in dee
p networks by replacing global backpropagation with truncated layer-wise backpro
pagation. Local parallelism enables fully asynchronous layer-wise parallelism wi
th a low memory footprint, and requires little communication overhead compared w
ith model parallelism. We show results in both vision and language domains acros
s a diverse set of architectures, and find that local parallelism is particularl
y effective in the high-compute regime.
****
Efficient Long-Range Convolutions for Point Clouds
Yifan Peng,Lin Lin,Lexing Ying,Leonardo Zepeda-Nunez
The efficient treatment of long-range interactions for point clouds is a challen
ging problem in many scientific machine learning applications. To extract global
 information, one usually needs a large window size, a large number of layers, a
nd/or a large number of channels. This can often significantly increase the comp
utational cost. In this work, we present a novel neural network layer that direc
tly incorporates long-range information for a point cloud. This layer, dubbed th
e long-range convolutional (LRC)-layer, leverages the convolutional theorem coup
led with the non-uniform Fourier transform. In a nutshell, the LRC-layer mollifi
es the point cloud to an adequately sized regular grid, computes its Fourier tra
nsform, multiplies the result by a set of trainable Fourier multipliers, compute
s the inverse Fourier transform, and finally interpolates the result back to the
 point cloud. The resulting global all-to-all convolution operation can be perfo
rmed in nearly-linear time asymptotically with respect to the number of input po
ints. The LRC-layer is a particularly powerful tool when combined with local con
volution as together they offer efficient and seamless treatment of both short a
nd long range interactions. We showcase this framework by introducing a neural n
etwork architecture that combines LRC-layers with short-range convolutional laye
rs to accurately learn the energy and force associated with a N-body potential
. We also exploit the induced two-level decomposition and propose an efficient
strategy to train the combined architecture with a reduced number of samples.
****
Continual Prototype Evolution: Learning Online from Non-Stationary Data Streams
Matthias De Lange,Tinne Tuytelaars
Attaining prototypical features to represent class distributions is well establi
shed in representation learning. However, learning prototypes online from stream
s of data proves a challenging endeavor as they rapidly become outdated, caused
by an ever-changing parameter space in the learning process. Additionally, conti
nual learning assumes a non-stationary nature of the data stream, typically resu
lting in catastrophic forgetting of previous knowledge. As a first, we introduce
 a system addressing both problems, where prototypes evolve continually in a sha
red latent space, enabling learning and prediction at any point in time. In cont
rast to the major body of work in continual learning, data streams are processed
 in an online fashion, without additional task-information, and an efficient mem
ory scheme provides robustness to imbalanced data streams. Besides nearest neigh
bor based prediction, learning is facilitated by a novel objective function, enc
ouraging cluster density about the class prototype and increased inter-class var
iance. Furthermore, the latent space quality is elevated by pseudo-prototypes in
 each batch, constituted by replay of exemplars from memory. We generalize the e
xisting paradigms in continual learning to incorporate data incremental learning
 from data streams by formalizing a two-agent learner-evaluator framework, and o
btain state-of-the-art performance by a significant margin on eight benchmarks,
including three highly imbalanced data streams.
****
Energy-based Out-of-distribution Detection for Multi-label Classification
Haoran Wang,Weitang Liu,Alex Bocchieri,Yixuan Li
Out-of-distribution (OOD) detection is essential to prevent anomalous inputs fro
m causing a model to fail during deployment. Improved methods for OOD detection
in multi-class classification have emerged, while OOD detection methods for mult
i-label classification remain underexplored and use rudimentary techniques. We p
ropose SumEnergy, a simple and effective method, which estimates the OOD indicat

or scores by aggregating energy scores from multiple labels. We show that SumEne
rgy can be mathematically interpreted from a joint likelihood perspective. Our r
esults show consistent improvement over previous methods that are based on the m
aximum-valued scores, which fail to capture joint information from multiple labe
ls. We demonstrate the effectiveness of our method on three common multi-label c
lassification benchmarks, including MS-COCO, PASCAL-VOC, and NUS-WIDE. We show t
hat SumEnergy reduces the FPR95 by up to 10.05% compared to the previous best ba
seline, establishing state-of-the-art performance.
****

Action and Perception as Divergence Minimization
Danijar Hafner,Pedro A Ortega,Jimmy Ba,Thomas Parr,Karl Friston,Nicolas Heess
We introduce a unified objective for action and perception of intelligent agents
. Extending representation learning and control, we minimize the joint divergenc
e between the combined system of agent and environment and a target distribution
. Intuitively, such agents use perception to align their beliefs with the world,
 and use actions to align the world with their beliefs. Minimizing the joint div
ergence to an expressive target maximizes the mutual information between the age
nt's representations and inputs, thus inferring representations that are informa
tive of past inputs and exploring future inputs that are informative of the repr
esentations. This lets us explain intrinsic objectives, such as representation l
earning, information gain, empowerment, and skill discovery from minimal assumpt
ions. Moreover, interpreting the target distribution as a latent variable model
suggests powerful world models as a path toward highly adaptive agents that seek
 large niches in their environments, rendering task rewards optional. The framew
ork provides a common language for comparing a wide range of objectives, advance
s the understanding of latent variables for decision making, and offers a recipe
 for designing novel objectives. We recommend deriving future agent objectives t
he joint divergence to facilitate comparison, to point out the agent's target di
stribution, and to identify the intrinsic objective terms needed to reach that d
istribution.
****

Mastering Atari with Discrete World Models
Danijar Hafner,Timothy P Lillicrap,Mohammad Norouzi,Jimmy Ba
Intelligent agents need to generalize from past experience to achieve goals in c
omplex environments. World models facilitate such generalization and allow learn
ing behaviors from imagined outcomes to increase sample-efficiency. While learni
ng world models from image inputs has recently become feasible for some tasks, m
odeling Atari games accurately enough to derive successful behaviors has remaine
d an open challenge for many years. We introduce DreamerV2, a reinforcement lear
ning agent that learns behaviors purely from predictions in the compact latent s
pace of a powerful world model. The world model uses discrete representations an
d is trained separately from the policy. DreamerV2 constitutes the first agent t
hat achieves human-level performance on the Atari benchmark of 55 tasks by learn
ing behaviors inside a separately trained world model. With the same computation
al budget and wall-clock time, Dreamer V2 reaches 200M frames and surpasses the
final performance of the top single-GPU agents IQN and Rainbow. DreamerV2 is als
o applicable to tasks with continuous actions, where it learns an accurate world
 model of a complex humanoid robot and solves stand-up and walking from only pix
el inputs.
****

Spatial Dependency Networks: Neural Layers for Improved Generative Image Modelin
g
■or■e Miladinovi■,Aleksandar Stani■,Stefan Bauer,Jürgen Schmidhuber,Joachim M. B
uhmann
How to improve generative modeling by better exploiting spatial regularities and
 coherence in images? We introduce a novel neural network for building image gen
erators (decoders) and apply it to variational autoencoders (VAEs). In our spati
al dependency networks (SDNs), feature maps at each level of a deep neural net a
re computed in a spatially coherent way, using a sequential gating-based mechani
sm that distributes contextual information across 2-D space. We show that augmen

ting the decoder of a hierarchical VAE by spatial dependency layers considerably improves density estimation over baseline convolutional architectures and the state-of-the-art among the models within the same class. Furthermore, we demonstrate that SDN can be applied to large images by synthesizing samples of high quality and coherence. In a vanilla VAE setting, we find that a powerful SDN decoder also improves learning disentangled representations, indicating that neural architectures play an important role in this task. Our results suggest favoring spatial dependency over convolutional layers in various VAE settings. The accompanying source code is given at https://github.com/djordjemila/sdn.

****

Approximating Pareto Frontier through Bayesian-optimization-directed Robust Multi-objective Reinforcement Learning

Xiangkun He,Jianye HAO,Dong Li,Bin Wang,Wulong Liu

Many real-word decision or control problems involve multiple conflicting objectives and uncertainties, which requires learned policies are not only Pareto optimal but also robust. In this paper, we proposed a novel algorithm to approximate a representation for robust Pareto frontier through Bayesian-optimization-directed robust multi-objective reinforcement learning (BRMORL). Firstly, environmental uncertainty is modeled as an adversarial agent over the entire space of preferences by incorporating zero-sum game into multi-objective reinforcement learning (MORL). Secondly, a comprehensive metric based on hypervolume and information entropy is presented to evaluate convergence, diversity and evenness of the distribution for Pareto solutions. Thirdly, the agent’s learning process is regarded as a black-box, and the comprehensive metric we proposed is computed after each episode of training, then a Bayesian optimization (BO) algorithm is adopted to guide the agent to evolve towards improving the quality of the approximated Pareto frontier. Finally, we demonstrate the effectiveness of proposed approach on challenging multi-objective tasks across four environments, and show our scheme can produce robust policies under environmental uncertainty.

****

For interpolating kernel machines, minimizing the norm of the ERM solution minimizes stability

Akshay Rangamani,Lorenzo Rosasco,Tomaso Poggio

We study the average CV Leave One Out stability of kernel ridge-less regression and derive corresponding risk bounds. We show that the interpolating solution with minimum norm minimizes a bound on CV Leave One Out stability, which in turn is controlled by the condition number of the empirical kernel matrix. The latter can be characterized in the asymptotic regime where both the dimension and cardinality of the data go to infinity. Under the assumption of random kernel matrices, the corresponding test error should be expected to follow a double descent curve.

****

Reinforcement Learning with Latent Flow

Wenling Shang,Xiaofei Wang,Aravind Rajeswaran,Aravind Srinivas,Yang Gao,Pieter Abbeel,Michael Laskin

Temporal information is essential to learning effective policies with Reinforcement Learning (RL). However, current state-of-the-art RL algorithms either assume that such information is given as part of the state space or, when learning from pixels, use the simple heuristic of frame-stacking to implicitly capture temporal information present in the image observations. This heuristic is in contrast to the current paradigm in video classification architectures, which utilize explicit encodings of temporal information through methods such as optical flow and two-stream architectures to achieve state-of-the-art performance. Inspired by leading video classification architectures, we introduce the Flow of Latents for Reinforcement Learning Flare, a network architecture for RL that explicitly encodes temporal information through latent vector differences. We show that Flare (i) recovers optimal performance in state-based RL without explicit access to the state velocity, solely with positional state information, (ii) achieves state-of-the-art performance on pixel-based continuous control tasks within the DeepMind control benchmark suite, (iii) is the most sample efficient model-free pixel-

based RL algorithm on challenging environments in the DeepMind control suite such as quadruped walk, hopper hop, finger turn hard, pendulum swing, and walker run, outperforming the prior model-free state-of-the-art by 1.9 and 1.5 on the 500k and 1M step benchmarks, respectively, and (iv), when augmented over rainbow DQN, outperforms or matches the baseline on a diversity of challenging Atari games at 50M time step benchmark.
****
Efficient Transformers in Reinforcement Learning using Actor-Learner Distillation
Emilio Parisotto,Russ Salakhutdinov
Many real-world applications such as robotics provide hard constraints on power and compute that limit the viable model complexity of Reinforcement Learning (RL) agents. Similarly, in many distributed RL settings, acting is done on un-accelerated hardware such as CPUs, which likewise restricts model size to prevent intractable experiment run times. These "actor-latency" constrained settings present a major obstruction to the scaling up of model complexity that has recently been extremely successful in supervised learning. To be able to utilize large model capacity while still operating within the limits imposed by the system during acting, we develop an "Actor-Learner Distillation" (ALD) procedure that leverages a continual form of distillation that transfers learning progress from a large capacity learner model to a small capacity actor model. As a case study, we develop this procedure in the context of partially-observable environments, where transformer models have had large improvements over LSTMs recently, at the cost of significantly higher computational complexity. With transformer models as the learner and LSTMs as the actor, we demonstrate in several challenging memory environments that using Actor-Learner Distillation largely recovers the clear sample-efficiency gains of the transformer learner model while maintaining the fast inference and reduced total training time of the LSTM actor model.
****
IOT: Instance-wise Layer Reordering for Transformer Structures
Jinhua Zhu,Lijun Wu,Yingce Xia,Shufang Xie,Tao Qin,Wengang Zhou,Houqiang Li,Tie-Yan Liu
With sequentially stacked self-attention, (optional) encoder-decoder attention, and feed-forward layers, Transformer achieves big success in natural language processing (NLP), and many variants have been proposed. Currently, almost all these models assume that the \emph{layer order} is fixed and kept the same across data samples. We observe that different data samples actually favor different orders of the layers. Based on this observation, in this work, we break the assumption of the fixed layer order in Transformer and introduce instance-wise layer reordering into model structure. Our Instance-wise Ordered Transformer (IOT) can model variant functions by reordered layers, which enables each sample to select the better one to improve the model performance under the constraint of almost same number of parameters. To achieve this, we introduce a light predictor with negligible parameter and inference cost to decide the most capable and favorable layer order for any input sequence. Experiments on 3 tasks (neural machine translation, abstractive summarization, and code generation) and 9 datasets demonstrate consistent improvements of our method. We further show that our method can also be applied to other architectures beyond Transformer. Our code is released at Github\footnote{\url{https://github.com/instance-wise-ordered-transformer/IOT}}.
****
Federated Learning via Posterior Averaging: A New Perspective and Practical Algorithms
Maruan Al-Shedivat,Jennifer Gillenwater,Eric Xing,Afshin Rostamizadeh
Federated learning is typically approached as an optimization problem, where the goal is to minimize a global loss function by distributing computation across client devices that possess local data and specify different parts of the global objective. We present an alternative perspective and formulate federated learning as a posterior inference problem, where the goal is to infer a global posterior distribution by having client devices each infer the posterior of their local

data. While exact inference is often intractable, this perspective provides a principled way to search for global optima in federated settings. Further, starting with the analysis of federated quadratic objectives, we develop a computation- and communication-efficient approximate posterior inference algorithm—federated posterior averaging (FedPA). Our algorithm uses MCMC for approximate inference of local posteriors on the clients and efficiently communicates their statistics to the server, where the latter uses them to refine a global estimate of the posterior mode. Finally, we show that FedPA generalizes federated averaging (FedAvg), can similarly benefit from adaptive optimizers, and yields state-of-the-art results on four realistic and challenging benchmarks, converging faster, to better optima.

****

Adam$^+$: A Stochastic Method with Adaptive Variance Reduction

Mingrui Liu,Wei Zhang,Francesco Orabona,Tianbao Yang

Adam is a widely used stochastic optimization method for deep learning applications. While practitioners prefer Adam because it requires less parameter tuning, its use is problematic from a theoretical point of view since it may not converge. Variants of Adam have been proposed with provable convergence guarantee, but they tend not be competitive with Adam on the practical performance. In this paper, we propose a new method named Adam$^+$ (pronounced as Adam-plus). Adam$^+$ retains some of the key components of Adam but it also has several noticeable differences: (i) it does not maintain the moving average of second moment estimate but instead computes the moving average of first moment estimate at extrapolated points; (ii) its adaptive step size is formed not by dividing the square root of second moment estimate but instead by dividing the root of the norm of first moment estimate. As a result, Adam$^+$ requires few parameter tuning, as Adam, but it enjoys a provable convergence guarantee. Our analysis further shows that Adam$^+$ enjoys adaptive variance reduction, i.e., the variance of the stochastic gradient estimator reduces as the algorithm converges, hence enjoying an adaptive convergence. We also propose a more general variant of Adam$^+$ with different adaptive step sizes and establish their fast convergence rate. Our empirical studies on various deep learning tasks, including image classification, language modeling, and automatic speech recognition, demonstrate that Adam$^+$ significantly outperforms Adam and achieves comparable performance with best-tuned SGD and momentum SGD.

****

Getting a CLUE: A Method for Explaining Uncertainty Estimates

Javier Antoran,Umang Bhatt,Tameem Adel,Adrian Weller,José Miguel Hernández-Lobato

Both uncertainty estimation and interpretability are important factors for trustworthy machine learning systems. However, there is little work at the intersection of these two areas. We address this gap by proposing a novel method for interpreting uncertainty estimates from differentiable probabilistic models, like Bayesian Neural Networks (BNNs). Our method, Counterfactual Latent Uncertainty Explanations (CLUE), indicates how to change an input, while keeping it on the data manifold, such that a BNN becomes more confident about the input's prediction. We validate CLUE through 1) a novel framework for evaluating counterfactual explanations of uncertainty, 2) a series of ablation experiments, and 3) a user study. Our experiments show that CLUE outperforms baselines and enables practitioners to better understand which input patterns are responsible for predictive uncertainty.

****

The Logical Options Framework

Brandon Araki,Xiao Li,Kiran Vodrahalli,Jonathan DeCastro,J Micah Fry,Daniela Rus

Learning composable policies for environments with complex rules and tasks is a challenging problem. We introduce a hierarchical reinforcement learning framework called the Logical Options Framework (LOF) that learns policies that are satisfying, optimal, and composable. LOF efficiently learns policies that satisfy tasks by representing the task as an automaton and integrating it into learning and planning. We provide and prove conditions under which LOF will learn satisfying

, optimal policies. And lastly, we show how LOF's learned policies can be compos ed to satisfy unseen tasks with only 10-50 retraining steps. We evaluate LOF on four tasks in discrete and continuous domains.
****

Learning Visual Representations for Transfer Learning by Suppressing Texture
Shlok Kumar Mishra,Anshul Shah,Ankan Bansal,Jonghyun Choi,Abhinav Shrivastava,Ab hishek Sharma,David Jacobs
Recent works have shown that features obtained from supervised training of CNNs may over-emphasize texture rather than encoding high-level information. In self -supervised learning, in particular, texture as a low-level cue may provide shor tcuts that prevent the network from learning higher-level representations. To a ddress these problems we propose to use classic methods based on anisotropic dif fusion to augment training using images with suppressed texture. This simple met hod helps retain important edge information and suppress texture at the same tim e.
We report our observations for fully supervised and self-supervised learning tas ks like MoCoV2 and Jigsaw and achieve state-of-the-art results on object detecti on and image classification with eight diverse datasets.
Our method is particularly effective for transfer learning tasks and we observed improved performance on five standard transfer learning datasets.
The large improvements on the Sketch-ImageNet dataset, DTD dataset and additional visual analyses of saliency maps suggest that our approach helps in l earning better representations that transfer well.
****

Extracting Strong Policies for Robotics Tasks from Zero-Order Trajectory Optimiz ers
Cristina Pinneri,Shambhuraj Sawant,Sebastian Blaes,Georg Martius
Solving high-dimensional, continuous robotic tasks is a challenging optimization problem. Model-based methods that rely on zero-order optimizers like the cross- entropy method (CEM) have so far shown strong performance and are considered sta te-of-the-art in the model-based reinforcement learning community. However, this success comes at the cost of high computational complexity, being therefore not suitable for real-time control. In this paper, we propose a technique to jointl y optimize the trajectory and distill a policy, which is essential for fast exec ution in real robotic systems. Our method builds upon standard approaches, like guidance cost and dataset aggregation, and introduces a novel adaptive factor wh ich prevents the optimizer from collapsing to the learner's behavior at the begi nning of the training. The extracted policies reach unprecedented performance on challenging tasks as making a humanoid stand up and opening a door without rewa rd shaping
****

Deep Convolution for Irregularly Sampled Temporal Point Clouds
Erich Merrill III,Stefan Lee,Li Fuxin,Thomas G Dietterich,Alan Fern
We consider the problem of modeling the dynamics of continuous spatial-temporal processes represented by irregular samples through both space and time. Such pro cesses occur in sensor networks, citizen science, multi-robot systems, and many others.
We propose a new deep model that is able to directly learn and predict over this irregularly sampled data, without voxelization, by leveraging a recent convolut ional architecture for static point clouds. The model also easily incorporates t he notion of multiple entities in the process. In particular, the model can flex ibly answer prediction queries about arbitrary space-time points for different e ntities regardless of the distribution of the training or test-time data. We pre sent experiments on real-world weather station data and battles between large ar mies in StarCraft II. The results demonstrate the model's flexibility in answeri ng a variety of query types and demonstrate improved performance and efficiency compared to state-of-the-art baselines.
****

Parameter-Efficient Transfer Learning with Diff Pruning
Demi Guo,Alexander M Rush,Yoon Kim

While task-specific finetuning of deep networks pretrained with self-supervision has led to significant empirical advances in NLP, their large size makes the standard finetuning approach difficult to apply to multi-task, memory-constrained settings, as storing the full model parameters for each task become prohibitively expensive. We propose $\textit{diff pruning}$ as a simple approach to enable parameter-efficient transfer learning within the pretrain-finetune framework. This approach views finetuning as learning a task-specific diff vector that is applied on top of the pretrained parameter vector, which remains fixed and is shared across different tasks. The diff vector is adaptively pruned during training with a differentiable approximation to the L_0-norm penalty to encourage sparsity. Diff pruning becomes parameter-efficient as the number of tasks increases, as it requires storing only the nonzero positions and weights of the diff vector for each task, while the cost of storing the shared pretrained model remains constant. We find that models finetuned with diff pruning can match the performance of fully finetuned baselines on the GLUE benchmark while only modifying 0.5$\%$ of the pretrained model's parameters per task.
****
Global Attention Improves Graph Networks Generalization
Omri Puny,Heli Ben-Hamu,Yaron Lipman
This paper advocates incorporating a Low-Rank Global Attention (LRGA) module, a Computation and memory efficient variant of the dot-product attention (Vaswani et al., 2017), to Graph Neural Networks (GNNs) for improving their generalization power.
To theoretically quantify the generalization properties granted by adding the LRGA module to GNNs, we focus on a specific family of expressive GNNs and show that augmenting it with LRGA provides algorithmic alignment to a powerful graph isomorphism test, namely the 2-Folklore Weisfeiler-Lehman (2-FWL) algorithm. In more detail we: (i) consider the recent Random Graph Neural Network (RGNN) (Sato et al., 2020) framework and prove that it is universal in probability; (ii) show that RGNN augmented with LRGA aligns with 2-FWL update step via polynomial kernels; and (iii) bound the sample complexity of the kernel's feature map when learned with a randomly initialized two-layer MLP.
From a practical point of view, augmenting existing GNN layers with LRGA produces state of the art results in current GNN benchmarks. Lastly, we observe that augmenting various GNN architectures with LRGA often closes the performance gap across different models.
****
FAST GRAPH ATTENTION NETWORKS USING EFFECTIVE RESISTANCE BASED GRAPH SPARSIFICATION
Rakshith Sharma Srinivasa,Cao Xiao,Lucas Glass,Justin Romberg,Jimeng Sun
The attention mechanism has demonstrated superior performance for inference over nodes in graph neural networks (GNNs), however, they result in a high computational burden during both training and inference. We propose FastGAT, a method to make attention based GNNs lightweight by using spectral sparsification to generate an optimal pruning of the input graph. This results in a per-epoch time that is almost linear in the number of graph nodes as opposed to quadratic. Further, we provide a re-formulation of a specific attention based GNN, Graph Attention Network (GAT) that interprets it as a graph convolution method using the random walk normalized graph Laplacian. Using this framework, we theoretically prove that spectral sparsification preserves the features computed by the GAT model, thereby justifying our FastGAT algorithm. We experimentally evaluate FastGAT on several large real world graph datasets for node classification tasks, FastGAT can dramatically reduce (up to 10x) the computational time and memory requirements, allowing the usage of attention based GNNs on large graphs.
****
Sharpness-aware Minimization for Efficiently Improving Generalization
Pierre Foret,Ariel Kleiner,Hossein Mobahi,Behnam Neyshabur
In today's heavily overparameterized models, the value of the training loss provides few guarantees on model generalization ability. Indeed, optimizing only the training loss value, as is commonly done, can easily lead to suboptimal model q

uality. Motivated by the connection between geometry of the loss landscape and g eneralization---including a generalization bound that we prove here---we introdu ce a novel, effective procedure for instead simultaneously minimizing loss value and loss sharpness. In particular, our procedure, Sharpness-Aware Minimization (SAM), seeks parameters that lie in neighborhoods having uniformly low loss; th is formulation results in a min-max optimization problem on which gradient desce nt can be performed efficiently. We present empirical results showing that SAM i mproves model generalization across a variety of benchmark datasets (e.g., CIFAR -{10, 100}, ImageNet, finetuning tasks) and models, yielding novel state-of-the- art performance for several. Additionally, we find that SAM natively provides r obustness to label noise on par with that provided by state-of-the-art procedure s that specifically target learning with noisy labels.

****

FTSO: Effective NAS via First Topology Second Operator

Likang Wang,Lei Chen

Existing one-shot neural architecture search (NAS) methods generally contain a g iant supernet, which leads to heavy computational cost. Our method, named FTSO, separates the whole architecture search into two sub-steps. In the first step, w e only search for the topology, and in the second step, we only search for the o perators. FTSO not only reduces NAS’s search time from days to 0.68 seconds, but also significantly improves the accuracy. Specifically, our experiments on Imag eNet show that within merely 18 seconds, FTSO can achieve 76.4% testing accuracy , 1.5% higher than the baseline, PC-DARTS. In addition, FTSO can reach 97.77% te sting accuracy, 0.27% higher than the baseline, with 99.8% of search time saved on CIFAR10.

****

Provably Faster Algorithms for Bilevel Optimization and Applications to Meta-Lea rning

Kaiyi Ji,Junjie Yang,Yingbin Liang

Bilevel optimization has arisen as a powerful tool for many machine learning pro blems such as meta-learning, hyperparameter optimization, and reinforcement lear ning. In this paper, we investigate the nonconvex-strongly-convex bilevel optimi zation problem. For deterministic bilevel optimization, we provide a comprehensi ve finite-time convergence analysis for two popular algorithms respectively base d on approximate implicit differentiation (AID) and iterative differentiation (I TD). For the AID-based method, we orderwisely improve the previous finite-time c onvergence analysis due to a more practical parameter selection as well as a war m start strategy, and for the ITD-based method we establish the first theoretica l convergence rate. Our analysis also provides a quantitative comparison between ITD and AID based approaches. For stochastic bilevel optimization, we propose a novel algorithm named stocBiO, which features a sample-efficient hypergradient estimator using efficient Jacobian- and Hessian-vector product computations. We provide the finite-time convergence guarantee for stocBiO, and show that stocBiO outperforms the best known computational complexities orderwisely with respect to the condition number κ and the target accuracy ϵ. We further validate our theoretical results and demonstrate the efficiency of bilevel optim ization algorithms by the experiments on meta-learning and hyperparameter optim ization.

****

BREEDS: Benchmarks for Subpopulation Shift

Shibani Santurkar,Dimitris Tsipras,Aleksander Madry

We develop a methodology for assessing the robustness of models to subpopulation shift---specifically, their ability to generalize to novel data subpopulations that were not observed during training. Our approach leverages the class structu re underlying existing datasets to control the data subpopulations that comprise the training and test distributions. This enables us to synthesize realistic di stribution shifts whose sources can be precisely controlled and characterized, w ithin existing

large-scale datasets. Applying this methodology to the ImageNet dataset, we crea te a suite of subpopulation shift benchmarks of varying granularity. We then val

idate that the corresponding shifts are tractable by obtaining human baselines. Finally, we utilize these benchmarks to measure the sensitivity of standard model architectures as well as the effectiveness of existing train-time robustness interventions.

****

Quantifying Task Complexity Through Generalized Information Measures
Aditya Chattopadhyay,Benjamin David Haeffele,Donald Geman,Rene Vidal
How can we measure the “complexity” of a learning task so that we can compare one task to another? From classical information theory, we know that entropy is a useful measure of the complexity of a random variable and provides a lower bound on the minimum expected number of bits needed for transmitting its state. In this paper, we propose to measure the complexity of a learning task by the minimum expected number of questions that need to be answered to solve the task. For example, the minimum expected number of patches that need to be observed to classify FashionMNIST images. We prove several properties of the proposed complexity measure, including connections with classical entropy and sub-additivity for multiple tasks. As the computation of the minimum expected number of questions is generally intractable, we propose a greedy procedure called “information pursuit” (IP), which selects one question at a time depending on previous questions and their answers. This requires learning a probabilistic generative model relating data and questions to the task, for which we employ variational autoencoders and normalizing flows. We illustrate the usefulness of the proposed measure on various binary image classification tasks using image patches as the query set. Our results indicate that the complexity of a classification task increases as signal-to-noise ratio decreases, and that classification of the KMNIST dataset is more complex than classification of the FashionMNIST dataset. As a byproduct of choosing patches as queries, our approach also provides a principled way of determining which pixels in an image are most informative for a task.

****

Status-Quo Policy Gradient in Multi-agent Reinforcement Learning
Pinkesh Badjatiya,Mausoom Sarkar,Abhishek Sinha,Nikaash Puri,Jayakumar Subramanian,Siddharth Singh,Balaji Krishnamurthy
Individual rationality, which involves maximizing expected individual return, does not always lead to optimal individual or group outcomes in multi-agent problems. For instance, in social dilemma situations, Reinforcement Learning (RL) agents trained to maximize individual rewards converge to mutual defection that is individually and socially sub-optimal. In contrast, humans evolve individual and socially optimal strategies in such social dilemmas. Inspired by ideas from human psychology that attribute this behavior in humans to the status-quo bias, we present a status-quo loss (SQLoss) and the corresponding policy gradient algorithm that incorporates this bias in an RL agent. We demonstrate that agents trained with SQLoss evolve individually as well as socially optimal behavior in several social dilemma matrix games. To apply SQLoss to games where cooperation and defection are determined by a sequence of non-trivial actions, we present GameDistill, an algorithm that reduces a multi-step game with visual input to a matrix game. We empirically show how agents trained with SQLoss on a GameDistill reduced version of the Coin Game evolve optimal policies.

****

On the Impossibility of Global Convergence in Multi-Loss Optimization
Alistair Letcher
Under mild regularity conditions, gradient-based methods converge globally to a critical point in the single-loss setting. This is known to break down for vanilla gradient descent when moving to multi-loss optimization, but can we hope to build some algorithm with global guarantees? We negatively resolve this open problem by proving that desirable convergence properties cannot simultaneously hold for any algorithm. Our result has more to do with the existence of games with no satisfactory outcomes, than with algorithms per se. More explicitly we construct a two-player game with zero-sum interactions whose losses are both coercive and analytic, but whose only simultaneous critical point is a strict maximum. Any 'reasonable' algorithm, defined to avoid strict maxima, will therefore fail to c

onverge. This is fundamentally different from single losses, where coercivity im
plies existence of a global minimum. Moreover, we prove that a wide range of ex
isting gradient-based methods almost surely have bounded but non-convergent iter
ates in a constructed zero-sum game for suitably small learning rates. It noneth
eless remains an open question whether such behavior can arise in high-dimension
al games of interest to ML practitioners, such as GANs or multi-agent RL.
****
Efficient Wasserstein Natural Gradients for Reinforcement Learning
Ted Moskovitz,Michael Arbel,Ferenc Huszar,Arthur Gretton
A novel optimization approach is proposed for application to policy gradient met
hods and evolution strategies for reinforcement learning (RL). The procedure use
s a computationally efficient \emph{Wasserstein natural gradient} (WNG) descent
that takes advantage of the geometry induced by a Wasserstein penalty to speed o
ptimization. This method follows the recent theme in RL of including divergence
penalties in the objective to establish trust regions. Experiments on challengin
g tasks demonstrate improvements in both computational cost and performance over
 advanced baselines.

****
Neural networks behave as hash encoders: An empirical study
Fengxiang He,Shiye Lei,Jianmin Ji,Dacheng Tao
The input space of a neural network with ReLU-like activations is partitioned in
to multiple linear regions, each corresponding to a specific activation pattern
of the included ReLU-like activations. We demonstrate that this partition exhibi
ts the following encoding properties across a variety of deep learning models: (
1) {\it determinism}: almost every linear region contains at most one training e
xample. We can therefore represent almost every training example by a unique act
ivation pattern, which is parameterized by a {\it neural code}; and (2) {\it cat
egorization}: according to the neural code, simple algorithms, such as K-Means
, K-NN, and logistic regression, can achieve fairly good performance on both t
raining and test data. These encoding properties surprisingly suggest that {\it
normal neural networks well-trained for classification behave as hash encoders w
ithout any extra efforts.} In addition, the encoding properties exhibit variabil
ity in different scenarios. {Further experiments demonstrate that {\it model siz
e}, {\it training time}, {\it training sample size}, {\it regularization}, and {
\it label noise} contribute in shaping the encoding properties, while the impact
s of the first three are dominant.} We then define an {\it activation hash phase
 chart} to represent the space expanded by {model size}, training time, training
 sample size, and the encoding properties, which is divided into three canonical
 regions: {\it under-expressive regime}, {\it critically-expressive regime}, and
 {\it sufficiently-expressive regime}.
****
Visualizing High-Dimensional Trajectories on the Loss-Landscape of ANNs
Stefan Horoi,Jessie Huang,Guy Wolf,Smita Krishnaswamy
Training artificial neural networks requires the optimization of highly non-conv
ex loss functions. Throughout the years, the scientific community has developed
an extensive set of tools and architectures that render this optimization task t
ractable and a general intuition has been developed for choosing hyper parameter
s that help the models reach minima that generalize well to unseen data. However
, for the most part, the difference in trainability in between architectures, ta
sks and even the gap in network generalization abilities still remain unexplaine
d. Visualization tools have played a key role in uncovering key geometric charac
teristics of the loss-landscape of ANNs and how they impact trainability and gen
eralization capabilities. However, most visualizations methods proposed so far h
ave been relatively limited in their capabilities since they are of linear natur
e and only capture features in a limited number of dimensions. We propose the us
e of the modern dimensionality reduction method PHATE which represents the SOTA
in terms of capturing both global and local structures of high-dimensional data.
 We apply this method to visualize the loss landscape during and after training.
 Our visualizations reveal differences in training trajectories and generalizati

on capabilities when used to make comparisons between optimization methods, init
ializations, architectures, and datasets. Given this success we anticipate this
method to be used in making informed choices about these aspects of neural netwo
rks.
****
Fundamental Limits and Tradeoffs in Invariant Representation Learning
Han Zhao,Chen Dan,Bryon Aragam,Tommi S. Jaakkola,Geoff Gordon,Pradeep Kumar Ravi
kumar
Many machine learning applications involve learning representations that achieve
 two competing goals: To maximize information or accuracy with respect to a targ
et while simultaneously maximizing invariance or independence with respect to a
subset of features. Typical examples include privacy-preserving learning, domain
 adaptation, and algorithmic fairness, just to name a few. In fact, all of the a
bove problems admit a common minimax game-theoretic formulation, whose equilibri
um represents a fundamental tradeoff between accuracy and invariance. In this pa
per, we provide an information-theoretic analysis of this general and important
problem under both classification and regression settings. In both cases, we ana
lyze the inherent tradeoffs between accuracy and invariance by providing a geome
tric characterization of the feasible region in the information plane, where we
connect the geometric properties of this feasible region to the fundamental limi
tations of the tradeoff problem. In the regression setting, we also derive a tig
ht lower bound on the Lagrangian objective that quantifies the tradeoff between
accuracy and invariance. Our results shed new light on this fundamental problem
by providing insights on the interplay between accuracy and invariance. These re
sults deepen our understanding of this fundamental problem and may be useful in
guiding the design of adversarial representation learning algorithms.

****
Block Skim Transformer for Efficient Question Answering
Yue Guan,Jingwen Leng,Yuhao Zhu,Minyi Guo
  Transformer-based encoder models have achieved promising results on natural la
nguage processing (NLP) tasks including question answering (QA). Different from
 sequence classification or language modeling tasks, hidden states at all positi
ons are used for the final classification in QA. However, we do not always need
 all the context to answer the raised question. Following this idea, we propose
d Block Skim Transformer (BST) to improve and accelerate the processing of trans
former QA models. The key idea of BST is to identify the context that must be f
urther processed and the blocks that could be safely discarded early on during i
nference. Critically, we learn such information from self-attention weights. As
 a result, the model hidden states are pruned at the sequence dimension, achievi
ng significant inference speedup. We also show that such extra training optimiz
ation objection also improves model accuracy. As a plugin to the transformer-bas
ed QA models, BST is compatible with other model compression methods without cha
nging existing network architectures. BST improves QA models' accuracies on dif
ferent datasets and achieves $1.6\times$ speedup on $BERT_{large}$ model.

****
Learning Self-Similarity in Space and Time as a Generalized Motion for Action Re
cognition
Heeseung Kwon,Manjin Kim,Suha Kwak,Minsu Cho
Spatio-temporal convolution often fails to learn motion dynamics in videos and t
hus an effective motion representation is required for video understanding in th
e wild. In this paper, we propose a rich and robust motion representation method
 based on spatio-temporal self-similarity (STSS). Given a sequence of frames, ST
SS represents each local region as similarities to its neighbors in space and ti
me. By converting appearance features into relational values, it enables the lea
rner to better recognize structural patterns in space and time. We leverage the
whole volume of STSS and let our model learn to extract an effective motion repr
esentation from it.
The proposed method is implemented as a neural block, dubbed SELFY, that can be

easily inserted into neural architectures and learned end-to-end without additio
nal supervision. With a sufficient volume of the neighborhood in space and time,
 it effectively captures long-term interaction and fast motion in the video, lea
ding to robust action recognition.
Our experimental analysis demonstrates its superiority over previous methods for
 motion modeling as well as its complementarity to spatio-temporal features from
 direct convolution. On the standard action recognition benchmarks, Something-So
mething-V1 & V2, Diving-48, and FineGym, the proposed method achieves the state-
of-the-art results.
****

Optimal Rates for Averaged Stochastic Gradient Descent under Neural Tangent Kern
el Regime
Atsushi Nitanda,Taiji Suzuki
We analyze the convergence of the averaged stochastic gradient descent for overp
arameterized two-layer neural networks for regression problems. It was recently
found that a neural tangent kernel (NTK) plays an important role in showing the
global convergence of gradient-based methods under the NTK regime, where the lea
rning dynamics for overparameterized neural networks can be almost characterized
 by that for the associated reproducing kernel Hilbert space (RKHS). However, th
ere is still room for a convergence rate analysis in the NTK regime. In this stu
dy, we show that the averaged stochastic gradient descent can achieve the minima
x optimal convergence rate, with the global convergence guarantee, by exploiting
 the complexities of the target function and the RKHS associated with the NTK. M
oreover, we show that the target function specified by the NTK of a ReLU network
 can be learned at the optimal convergence rate through a smooth approximation o
f a ReLU network under certain conditions.
****

BasisNet: Two-stage Model Synthesis for Efficient Inference
Mingda Zhang,Andrey Zhmoginov,Andrew G. Howard,Brendan Jou,Yukun Zhu,Li Zhang,Re
becca Hwa,Adriana Kovashka
We present BasisNet which combines recent advancements in efficient neural netwo
rk architectures, conditional computation, and early termination in a simple new
 form. Our approach uses a lightweight model to preview an image and generate in
put-dependent combination coefficients, which are later used to control the synt
hesis of a specialist model for making more accurate final prediction. The two-s
tage model synthesis strategy can be used with any network architectures and bot
h stages can be jointly trained end to end. We validated BasisNet on ImageNet cl
assification with MobileNets as backbone, and demonstrated clear advantage on ac
curacy-efficiency trade-off over strong baselines such as EfficientNet (Tan & Le
, 2019), FBNetV3 (Dai et al., 2020) and OFA (Cai et al., 2019). Specifically, Ba
sisNet-MobileNetV3 obtained 80.3% top-1 accuracy with only 290M Multiply-Add ope
rations (MAdds), halving the computational cost of previous state-of-the-art wit
hout sacrificing accuracy. Besides, since the first-stage lightweight model can
independently make predictions, inference can be terminated early if the predict
ion is sufficiently confident. With early termination, the average cost can be f
urther reduced to 198M MAdds while maintaining accuracy of 80.0%.
****

Extrapolatable Relational Reasoning With Comparators in Low-Dimensional Manifold
s
Duo Wang,Mateja Jamnik,Pietro Liò
While modern deep neural architectures generalise well when test data is sampled
 from the same distribution as training data, they fail badly for cases when the
 test data distribution differs from the training distribution even along a few
dimensions. This lack of out-of-distribution generalisation is increasingly mani
fested when the tasks become more abstract and complex, such as in relational re
asoning. In this paper we propose a neuroscience-inspired inductive-biased modul
e that can be readily amalgamated with current neural network architectures to i
mprove out-of-distribution (o.o.d) generalisation performance on relational reas
oning tasks. This module learns to project high-dimensional object representatio
ns to low-dimensional manifolds for more efficient and generalisable relational

comparisons. We show that neural nets with this inductive bias achieve considera bly better o.o.d generalisation performance for a range of relational reasoning tasks. We finally analyse the proposed inductive bias module to understand the i mportance of lower dimension projection, and propose an augmentation to the algo rithmic alignment theory to better measure algorithmic alignment with generalisa tion.
****

A Probabilistic Approach to Constrained Deep Clustering
Laura Manduchi,Kieran Chin-Cheong,Holger Michel,Sven Wellmann,Julia E Vogt
Clustering with constraints has gained significant attention in the field of sem i-supervised machine learning as it can leverage partial prior information on a growing amount of unlabelled data. Following recent advances in deep generative models, we derive a novel probabilistic approach to constrained clustering that can be trained efficiently in the framework of stochastic gradient variational B ayes. In contrast to existing approaches, our model (CVaDE) uncovers the underly ing distribution of the data conditioned on prior clustering preferences, expres sed as pairwise constraints. The inclusion of such constraints allows the user t o guide the clustering process towards a desirable partition of the data by indi cating which samples should or should not belong to the same class. We provide e xtensive experiments to demonstrate that CVaDE shows superior clustering perform ances and robustness compared to state-of-the-art deep constrained clustering me thods in a variety of data sets. We further demonstrate the usefulness of our ap proach on challenging real-world medical applications and face image generation.
****

Bayesian Metric Learning for Robust Training of Deep Models under Noisy Labels
Toan Tran,Hieu Vu,Gustavo Carneiro,Hung Bui
Label noise is a natural event of data collection and annotation and has been sh own to have significant impact on the performance of deep learning models regard ing accuracy reduction and sample complexity increase. This paper aims to develo p a novel theoretically sound Bayesian deep metric learning that is robust again st noisy labels.
Our proposed approach is inspired by a linear Bayesian large margin nearest neig hbor classification, and is a combination of Bayesian learning, triplet loss-bas ed deep metric learning and variational inference frameworks. We theoretically s how the robustness under label noise of our proposed method. The experimental re sults on benchmark data sets that contain both synthetic and realistic label noi se show a considerable improvement in the classification accuracy of our method compared to the linear Bayesian metric learning and the point estimate deep metr ic learning.
****

PMI-Masking: Principled masking of correlated spans
Yoav Levine,Barak Lenz,Opher Lieber,Omri Abend,Kevin Leyton-Brown,Moshe Tennenho ltz,Yoav Shoham
Masking tokens uniformly at random constitutes a common flaw in the pretraining of Masked Language Models (MLMs) such as BERT. We show that such uniform masking allows an MLM to minimize its training objective by latching onto shallow local signals, leading to pretraining inefficiency and suboptimal downstream performa nce. To address this flaw, we propose PMI-Masking, a principled masking strategy based on the concept of Pointwise Mutual Information (PMI), which jointly masks a token n-gram if it exhibits high collocation over the corpus. PMI-Masking mot ivates, unifies, and improves upon prior more heuristic approaches that attempt to address the drawback of random uniform token masking, such as whole-word mask ing, entity/phrase masking, and random-span masking. Specifically, we show exper imentally that PMI-Masking reaches the performance of prior masking approaches i n half the training time, and consistently improves performance at the end of pr etraining.
****

The Surprising Power of Graph Neural Networks with Random Node Initialization
Ralph Abboud,Ismail Ilkan Ceylan,Martin Grohe,Thomas Lukasiewicz
Graph neural networks (GNNs) are effective models for representation learning on

graph-structured data. However, standard GNNs are limited in their expressive power, as they cannot distinguish graphs beyond the capability of the Weisfeiler-Leman (1-WL) graph isomorphism heuristic. This limitation motivated a large body of work, including higher-order GNNs, which are provably more powerful models. To date, higher-order invariant and equivariant networks are the only models with known universality results, but these results are practically hindered by prohibitive computational complexity. Thus, despite their limitations, standard GNNs are commonly used, due to their strong practical performance. In practice, GNNs have shown a promising performance when enhanced with random node initialization (RNI), where the idea is to train and run the models with randomized initial node features. In this paper, we analyze the expressive power of GNNs with RNI, and pose the following question: are GNNs with RNI more expressive than GNNs? We prove that this is indeed the case, by showing that GNNs with RNI are universal, a first such result for GNNs not relying on computationally demanding higher-order properties. We then empirically analyze the effect of RNI on GNNs, based on carefully constructed datasets. Our empirical findings support the superior performance of GNNs with RNI over standard GNNs. In fact, we demonstrate that the performance of GNNs with RNI is often comparable with or better than that of higher-order GNNs, while keeping the much lower memory requirements of standard GNNs. However, this improvement typically comes at the cost of slower model convergence. Somewhat surprisingly, we found that the convergence rate and the accuracy of the models can be improved by using only a partial random initialization regime.
****
Ask Question with Double Hints: Visual Question Generation with Answer-awareness and Region-reference
Shen Kai,Lingfei Wu,Siliang Tang,Fangli Xu,Zhu Zhang,Yu Qiang,Yueting Zhuang
The task of visual question generation~(VQG) aims to generate human-like questions from an image and potentially other side information (e.g. answer type or the answer itself). Despite promising results have been achieved, previous works on VQG either i) suffer from one image to many questions mapping problem rendering the failure of generating referential and meaningful questions from an image, or ii) ignore rich correlations among the visual objects in an image and potential interactions between the side information and image. To address these limitations, we first propose a novel learning paradigm to generate visual questions with answer-awareness and region-reference. In particular, we aim to ask the right visual questions with \emph{Double Hints - textual answers and visual regions of interests}, effectively mitigating the existing one-to-many mapping issue. To this end, we develop a simple methodology to self-learn the visual hints without introducing any additional human annotations. Furthermore, to capture these sophisticated relationships, we propose a new double-hints guided Graph-to-Sequence learning framework that first models them as a dynamic graph and learns the implicit topology end-to-end, and then utilize a graph-to-sequence model to generate the questions with double hints. Our experiments on VQA2.0 and COCO-QA datasets demonstrate that our proposed model on this new setting can significantly outperform existing state-of-the-art baselines by a large margin.
****
Faster Training of Word Embeddings
Eliza Wszola,Martin Jaggi,Markus Püschel
Word embeddings have gained increasing popularity in the recent years due to the Word2vec library and its extension fastText that uses subword information. In this paper, we aim at improving the execution speed of fastText training on homogeneous multi- and manycore CPUs while maintaining accuracy. We present a novel open-source implementation that flexibly incorporates various algorithmic variants including negative sample sharing, batched updates, and a byte-pair encoding-based alternative for subword units. We build these novel variants over a fastText implementation that we carefully optimized for the architecture, memory hierarchy, and parallelism of current manycore CPUs. Our experiments on three languages demonstrate 3-20x speed-up in training time at competitive semantic and syntactic accuracy.

****
Multi-Level Local SGD: Distributed SGD for Heterogeneous Hierarchical Networks
Timothy Castiglia,Anirban Das,Stacy Patterson
We propose Multi-Level Local SGD, a distributed stochastic gradient method for learning a smooth, non-convex objective in a multi-level communication network with heterogeneous workers. Our network model consists of a set of disjoint sub-networks, with a single hub and multiple workers; further, workers may have different operating rates. The hubs exchange information with one another via a connected, but not necessarily complete communication network. In our algorithm, sub-networks execute a distributed SGD algorithm, using a hub-and-spoke paradigm, and the hubs periodically average their models with neighboring hubs. We first provide a unified mathematical framework that describes the Multi-Level Local SGD algorithm. We then present a theoretical analysis of the algorithm; our analysis shows the dependence of the convergence error on the worker node heterogeneity, hub network topology, and the number of local, sub-network, and global iterations. We illustrate the effectiveness of our algorithm in a multi-level network with slow workers via simulation-based experiments.
****
Kanerva++: Extending the Kanerva Machine With Differentiable, Locally Block Allocated Latent Memory
Jason Ramapuram,Yan Wu,Alexandros Kalousis
Episodic and semantic memory are critical components of the human memory model. The theory of complementary learning systems (McClelland et al., 1995) suggests that the compressed representation produced by a serial event (episodic memory) is later restructured to build a more generalized form of reusable knowledge (semantic memory). In this work, we develop a new principled Bayesian memory allocation scheme that bridges the gap between episodic and semantic memory via a hierarchical latent variable model. We take inspiration from traditional heap allocation and extend the idea of locally contiguous memory to the Kanerva Machine, enabling a novel differentiable block allocated latent memory. In contrast to the Kanerva Machine, we simplify the process of memory writing by treating it as a fully feed forward deterministic process, relying on the stochasticity of the read key distribution to disperse information within the memory. We demonstrate that this allocation scheme improves performance in memory conditional image generation, resulting in new state-of-the-art conditional likelihood values on binarized MNIST (≤41.58 nats/image) , binarized Omniglot (≤66.24 nats/image), as well as presenting competitive performance on CIFAR10, DMLab Mazes, Celeb-A and ImageNet32×32.
****
EVALUATION OF NEURAL ARCHITECTURES TRAINED WITH SQUARE LOSS VS CROSS-ENTROPY IN CLASSIFICATION TASKS
Like Hui,Mikhail Belkin
Modern neural architectures for classification tasks are trained using the cross-entropy loss, which is widely believed to be empirically superior to the square loss. In this work we provide evidence indicating that this belief may not be well-founded.

We explore several major neural architectures and a range of standard benchmark datasets for NLP, automatic speech recognition (ASR) and computer vision tasks to show that these architectures, with the same hyper-parameter settings as reported in the literature, perform comparably or better when trained with the square loss, even after equalizing computational resources.

Indeed, we observe that the square loss produces better results in the dominant majority of NLP and ASR experiments. Cross-entropy appears to have a slight edge on computer vision tasks.

We argue that there is little compelling empirical or theoretical evidence indicating a clear-cut advantage to the cross-entropy loss. Indeed, in our experiments, performance on nearly all non-vision tasks can be improved, sometimes significantly, by switching to the square loss. Furthermore, training with square loss appears to be less sensitive to the randomness in initialization. We posit that

training using the square loss for classification needs to be a part of best pra
ctices of modern deep learning on equal footing with cross-entropy.
****
A Communication Efficient Federated Kernel k-Means
Xiaochen Zhou,Xudong Wang
A federated kernel k-means algorithm is developed in this paper. This algorith
m resolves two challenging issues: 1) how to distributedly solve the optimizatio
n problem of kernel k-means under federated settings; 2) how to maintain commu
nication efficiency in the algorithm. To tackle the first challenge, a distribut
ed stochastic proximal gradient descent (DSPGD) algorithm is developed to determ
ine an approximated solution to the optimization problem of kernel k-means. To
 tackle the second challenge, a communication efficient mechanism (CEM) is desig
ned to reduce the communication cost. Besides, the federated kernel k-means pr
ovides two levels of privacy preservation: 1) users' local data are not exposed
to the cloud server; 2) the cloud server cannot recover users' local data from t
he local computational results via matrix operations. Theoretical analysis shows
: 1) DSPGD with CEM converges with an $O(1/T)$ rate, where T is the number of
iterations; 2) the communication cost of DSPGD with CEM is unrelated to the numb
er of data samples; 3) the clustering quality of the federated kernel k-means
approaches that of the standard kernel k-means, with a $(1+\epsilon)$ approxim
ate ratio. The experimental results show that the federated kernel k-means ach
ieves the highest clustering quality with the communication cost reduced by more
 than 60% in most cases.
****
Global inducing point variational posteriors for Bayesian neural networks and de
ep Gaussian processes
Sebastian W. Ober,Laurence Aitchison
We derive the optimal approximate posterior over the top-layer weights in a Baye
sian neural network for regression, and show that it exhibits strong dependencie
s on the lower-layer weights. We adapt this result to develop a correlated appro
ximate posterior over the weights at all layers in a Bayesian neural network. We
 extend this approach to deep Gaussian processes, unifying inference in the two
model classes. Our approximate posterior uses learned "global" inducing points,
which are defined only at the input layer and propagated through the network to
obtain inducing inputs at subsequent layers. By contrast, standard, "local", ind
ucing point methods from the deep Gaussian process literature optimise a separat
e set of inducing inputs at every layer, and thus do not model correlations acro
ss layers. Our method gives state-of-the-art performance for a variational Bayes
ian method, without data augmentation or tempering, on CIFAR-10 of 86.7%.
****
Asynchronous Modeling: A Dual-phase Perspective for Long-Tailed Recognition
Hu Zhang,Linchao Zhu,Yi Yang
This work explores deep learning based classification model on real-world datase
ts with a long-tailed distribution. Most of previous works deal with the long-ta
iled classification problem by re-balancing the overall distribution within the
whole dataset or directly transferring knowledge from data-rich classes to data-
poor ones. In this work, we consider the gradient distortion in long-tailed clas
sification when the gradient on data-rich classes and data-poor ones are incorpo
rated simultaneously, i.e., shifted gradient direction towards data-rich classes
 as well as the enlarged variance by the gradient fluctuation on data-poor class
es. Motivated by such phenomenon, we propose to disentangle the distinctive effe
cts of data-rich and data-poor gradient and asynchronously train a model via a d
ual-phase learning process. The first phase only concerns the data-rich classes.
 In the second phase, besides the standard classification upon data-poor classes
, we propose an exemplar memory bank to reserve representative examples and a me
mory-retentive loss via graph matching to retain the relation between two phases
. The extensive experimental results on four commonly used long-tailed benchmark
s including CIFAR100-LT, Places-LT, ImageNet-LT and iNaturalist 2018 highlight t
he excellent performance of our proposed method.
****

Maximum Entropy competes with Maximum Likelihood
Armen Allahverdyan
Maximum entropy (MAXENT) method has a large number of
applications in theoretical and applied machine learning, since it
provides a convenient non-parametric tool for estimating unknown
probabilities. The method is a major contribution of statistical
physics to probabilistic inference. However, a systematic approach
towards its validity limits is currently missing. Here we study MAXENT
in a Bayesian decision theory set-up, i.e. assuming that there exists a
well-defined prior Dirichlet density for unknown probabilities, and that
the average Kullback-Leibler (KL) distance can be employed for deciding
on the quality and applicability of various estimators. These allow to
evaluate the relevance of various MAXENT constraints, check its general
applicability, and compare MAXENT with estimators having various degrees
of dependence on the prior, {\it viz.} the regularized maximum
likelihood (ML) and the Bayesian estimators. We show that MAXENT applies
in sparse data regimes, but needs specific types of prior information.
In particular, MAXENT can outperform the optimally regularized
ML provided that there are prior rank correlations between the estimated
random quantity and its probabilities.
****
Autoencoder Image Interpolation by Shaping the Latent Space
Alon Oring,Zohar Yakhini,Yacov Hel-Or
One of the fascinating properties of deep learning is the ability of the network
 to reveal the underlying factors characterizing elements in datasets of differe
nt types. Autoencoders represent an effective approach for computing these facto
rs. Autoencoders have been studied in the context of enabling interpolation betw
een data points by decoding convex combinations of latent vectors. However, this
 interpolation often leads to artifacts or produces unrealistic results during r
econstruction. We argue that these incongruities are due to the structure of the
 latent space and to the fact that such naively interpolated latent vectors devi
ate from the data manifold. In this paper, we propose a regularization technique
 that shapes the latent representation to follow a manifold that is consistent w
ith the training images and that forces the manifold to be smooth and locally co
nvex. This regularization not only enables faithful interpolation between data p
oints, as we show herein, but can also be used as a general regularization techn
ique to avoid overfitting or to produce new samples for data augmentation
****
LINGUINE: LearnIng to pruNe on subGraph convolUtIon NEtworks
Yihan He,Wei Cao,Shun Zheng,Zhifeng Gao,Jiang Bian
Graph Convolutional Network (GCN) has become one of the most successful methods
for graph representation learning. Training and evaluating GCNs on large graphs
is challenging since full-batch GCNs have high overhead in memory and computatio
n. In recent years, research communities have been developing stochastic samplin
g methods to handle large graphs when it is unreal to put the whole graph into a
 single batch. The performance of the model depends largely on the quality and s
ize of subgraphs in the batch-training. Existing sampling approaches mostly focu
s on approximating the full-graph structure but care less about redundancy and r
andomness in sampling subgraphs. To address these issues and explore a better me
chanism of producing high-quality subgraphs to train GCNs, we proposed the \text
tt{Linguine} framework where we designed a meta-model to prune the subgraph smar
tly. To efficiently obtain the meta-model, we designed a joint training scenario
 with the idea of hardness based learning. The empirical study shows that our me
thod could augment the accuracy of the current state-of-art and reduce the error
 incurred by the redundancies in the subgraph structure. We also explored the re
asoning behind smart pruning via its visualization.
****
Exploring Transferability of Perturbations in Deep Reinforcement Learning
Ezgi Korkmaz
The use of Deep Neural Networks (DNNs) as function approximators has led to stri

king progress for reinforcement learning algorithms and applications. At the sam
e time, deep reinforcement learning agents have inherited the vulnerability of D
NNs to imperceptible adversarial perturbations to their inputs. Prior work on ad
versarial perturbations for deep reinforcement learning has generally relied on
calculating an adversarial perturbation customized to each state visited by the
agent. In this paper we propose a more realistic threat model in which the adver
sary computes the perturbation only once based on a single state. Furthermore, w
e show that to cause a deep reinforcement learning agent to fail it is enough to
 have only one adversarial offset vector in a black-box setting. We conduct expe
riments in various games from the Atari environment, and use our single-state ad
versaries to demonstrate the transferability of perturbations both between state
s of one MDP, and between entirely different MDPs. We believe our adversary fram
ework reveals fundamental properties of the environments used in deep reinforcem
ent learning training, and is a tangible step towards building robust and reliab
le deep reinforcement learning agents.
****
Local SGD Meets Asynchrony
Bapi Chatterjee,Vyacheslav Kungurtsev,Dan Alistarh
Distributed variants of stochastic gradient descent (SGD) are central to trainin
g deep neural networks on massive datasets.
Several scalable versions of data-parallel SGD have been developed, leveraging
asynchrony, communication-compression, and local gradient steps. Current researc
h seeks a balance between distributed scalability--seeking to minimize the amoun
t of synchronization needed--and generalization performance--seeking to achieve
the same or better accuracy relative to the sequential baseline. However, a key
issue in this regime is largely unaddressed: if ``local" data-parallelism is agg
ressively applied to better utilize the computing resources available with worke
rs, generalization performance of the trained model degrades.

In this paper, we present a method to improve the "local scalability" of decentr
alized SGD. In particular, we propose two key techniques: (a) shared-memory base
d asynchronous gradient updates at decentralized workers keeping the local minib
atch size small, and (b) an asynchronous non-blocking in-place averaging overlap
ping the local updates, thus essentially utilizing all compute resources at all
times without the need for large minibatches. Empirically, the additional noise
introduced in the procedure proves to be a boon for better generalization. On th
e theoretical side, we show that this method guarantees ergodic convergence for
non-convex objectives, and achieves the classic sublinear rate under standard as
sumptions.
On the practical side, we show that it improves upon the performance of local SG
D and related schemes, without compromising accuracy.
****
Differentiable Combinatorial Losses through Generalized Gradients of Linear Prog
rams
Xi Gao,Han Zhang,Aliakbar Panahi,Tom Arodz
Combinatorial problems with linear objective function play a central role in man
y computer science applications, and efficient algorithms for solving them are w
ell known. However, the solutions to these problems are not differentiable with
respect to the parameters specifying the problem instance – for example, shortes
t distance between two nodes in a graph is not a differentiable function of grap
h edge weights. Recently, attempts to integrate combinatorial and, more broadly,
 convex optimization solvers into gradient-trained models resulted in several ap
proaches for differentiating over the solution vector to the optimization proble
m. However, in many cases, the interest is in differentiating over only the obje
ctive value, not the solution vector, and using existing approaches introduces u
nnecessary overhead. Here, we show how to perform gradient descent directly over
 the objective value of the solution to combinatorial problems. We demonstrate a
dvantage of the approach in examples involving sequence-to-sequence modeling usi
ng differentiable encoder-decoder architecture with softmax or Gumbel-softmax, a
nd in weakly supervised learning involving a convolutional, residual feed-forwar

d network for image classification.

****
Model-Based Reinforcement Learning via Latent-Space Collocation
Oleh Rybkin,Chuning Zhu,Anusha Nagabandi,Kostas Daniilidis,Igor Mordatch,Sergey Levine
The ability to construct and execute long-term plans enables intelligent agents to solve complex multi-step tasks and prevents myopic behavior only seeking the short-term reward. Recent work has achieved significant progress on building agents that can predict and plan from raw visual observations. However, existing visual planning methods still require a densely shaped reward that provides the algorithm with a short-term signal that is always easy to optimize. These algorithms fail when the shaped reward is not available as they use simplistic planning methods such as sampling-based random shooting and are unable to plan for a distant goal. Instead, to achieve long-horizon visual control, we propose to use collocation-based planning, a powerful optimal control technique that plans forward a sequence of states while constraining the transitions to be physical. We propose a planning algorithm that adapts collocation to visual planning by leveraging probabilistic latent variable models. A model-based reinforcement learning agent equipped with our planning algorithm significantly outperforms prior model-based agents on challenging visual control tasks with sparse rewards and long-term goals.
****
Simple and Effective VAE Training with Calibrated Decoders
Oleh Rybkin,Kostas Daniilidis,Sergey Levine
Variational autoencoders (VAEs) provide an effective and simple method for modeling complex distributions. However, training VAEs often requires considerable hyperparameter tuning to determine the optimal amount of information retained by the latent variable. We study the impact of calibrated decoders, which learn the uncertainty of the decoding distribution and can determine this amount of information automatically, on the VAE performance. While many methods for learning calibrated decoders have been proposed, many of the recent papers that employ VAEs rely on heuristic hyperparameters and ad-hoc modifications instead. We perform the first comprehensive comparative analysis of calibrated decoder and provide recommendations for simple and effective VAE training. Our analysis covers a range of datasets and several single-image and sequential VAE models. We further propose a simple but novel modification to the commonly used Gaussian decoder, which computes the prediction variance analytically. We observe empirically that using heuristic modifications is not necessary with our method.
****
Disentangling Adversarial Robustness in Directions of the Data Manifold
Jiancong Xiao,Liusha Yang,Zhi-Quan Luo
Using generative models (GAN or VAE) to craft adversarial examples, i.e. generative adversarial examples, has received increasing attention in recent years. Previous studies showed that the generative adversarial examples work differently compared to that of the regular adversarial examples in many aspects, such as attack rates, perceptibility, and generalization. But the reasons causing the differences between regular and generative adversarial examples are unclear. In this work, we study the theoretical properties of the attacking mechanisms of the two kinds of adversarial examples in the Gaussian mixture data model case. We prove that adversarial robustness can be disentangled in directions of the data manifold. Specifically, we find that: 1. Regular adversarial examples attack in directions of small variance of the data manifold, while generative adversarial examples attack in directions of large variance. 2. Standard adversarial training increases model robustness by extending the data manifold boundary in directions of small variance, while on the contrary, adversarial training with generative adversarial examples increases model robustness by extending the data manifold boundary directions of large variance. In experiments, we demonstrate that these phenomena also exist on real datasets. Finally, we study the robustness trade-off between generative and regular adversarial examples. We show that the conflict be

tween regular and generative adversarial examples is much smaller than the confl
ict between regular adversarial examples of different norms.
****

Self-supervised Learning from a Multi-view Perspective
Yao-Hung Hubert Tsai,Yue Wu,Ruslan Salakhutdinov,Louis-Philippe Morency
As a subset of unsupervised representation learning, self-supervised representat
ion learning adopts self-defined signals as supervision and uses the learned rep
resentation for downstream tasks, such as object detection and image captioning.
 Many proposed approaches for self-supervised learning follow naturally a multi-
view perspective, where the input (e.g., original images) and the self-supervise
d signals (e.g., augmented images) can be seen as two redundant views of the dat
a. Building from this multi-view perspective, this paper provides an information
-theoretical framework to better understand the properties that encourage succes
sful self-supervised learning. Specifically, we demonstrate that self-supervised
 learned representations can extract task-relevant information and discard task-
irrelevant information. Our theoretical framework paves the way to a larger spac
e of self-supervised learning objective design. In particular, we propose a comp
osite objective that bridges the gap between prior contrastive and predictive le
arning objectives, and introduce an additional objective term to discard task-ir
relevant information. To verify our analysis, we conduct controlled experiments
to evaluate the impact of the composite objectives. We also explore our framewor
k's empirical generalization beyond the multi-view perspective, where the cross-
view redundancy may not be clearly observed.
****

Interpretable Neural Architecture Search via Bayesian Optimisation with Weisfeil
er-Lehman Kernels
Binxin Ru,Xingchen Wan,Xiaowen Dong,Michael Osborne
Current neural architecture search (NAS) strategies focus only on finding a sing
le, good, architecture. They offer little insight into why a specific network is
 performing well, or how we should modify the architecture if we want further im
provements. We propose a Bayesian optimisation (BO) approach for NAS that combin
es the Weisfeiler-Lehman graph kernel with a Gaussian process surrogate. Our met
hod not only optimises the architecture in a highly data-efficient manner, but a
lso affords interpretability by discovering useful network features and their co
rresponding impact on the network performance. Moreover, our method is capable o
f capturing the topological structures of the architectures and is scalable to l
arge graphs, thus making the high-dimensional and graph-like search spaces amena
ble to BO. We demonstrate empirically that our surrogate model is capable of ide
ntifying useful motifs which can guide the generation of new architectures. We f
inally show that our method outperforms existing NAS approaches to achieve the s
tate of the art on both closed- and open-domain search spaces.
****

Learning Causal Semantic Representation for Out-of-Distribution Prediction
Chang Liu,Xinwei Sun,Jindong Wang,Tao Li,Tao Qin,Wei Chen,Tie-Yan Liu
Conventional supervised learning methods, especially deep ones, are found to be
 sensitive to out-of-distribution (OOD) examples, largely because the learned rep
resentation mixes the semantic factor with the variation factor due to their dom
ain-specific correlation, while only the semantic factor causes the output. To a
ddress the problem, we propose a Causal Semantic Generative model (CSG) based on
 causality to model the two factors separately, and learn it on a single trainin
g domain for prediction without (OOD generalization) or with unsupervised data (
domain adaptation) in a test domain. We prove that CSG identifies the semantic f
actor on the training domain, and the invariance principle of causality subseque
ntly guarantees the boundedness of OOD generalization error and the success of a
daptation. We also design novel and delicate learning methods for both effective
 learning and easy prediction, following the first principle of variational Baye
s and the graphical structure of CSG. Empirical study demonstrates the effect of
 our methods to improve test accuracy for OOD generalization and domain adaptati
on.
****

Practical Locally Private Federated Learning with Communication Efficiency
Yan Feng,Tao Xiong,Ruofan Wu,Yuan Qi
Federated learning (FL) is a technique that trains machine learning models from decentralized data sources. We study FL under local differential privacy constraints, which provides strong protection against sensitive data disclosures via obfuscating the data before leaving the client. We identify two major concerns in designing practical privacy-preserving FL algorithms: communication efficiency and high-dimensional compatibility. We then develop a gradient-based learning algorithm called \emph{sqSGD} (selective quantized stochastic gradient descent) that addresses both concerns. The proposed algorithm is based on a novel privacy-preserving quantization scheme that uses a constant number of bits per dimension per client. Then we improve the base algorithm in two ways: first, we apply a gradient subsampling strategy that offers simultaneously better training performance and smaller communication costs under a fixed privacy budget. Secondly, we utilize randomized rotation as a preprocessing step to reduce quantization error. We also initialize a discussion about the role of quantization and perturbation in FL algorithm design with privacy and communication constraints. Finally, the practicality of the proposed framework is demonstrated on benchmark datasets. Experiment results show that sqSGD successfully learns large models like LeNet and ResNet with local privacy constraints. In addition, with fixed privacy and communication level, the performance of sqSGD significantly dominates that of baseline algorithms.
****
Activation Relaxation: A Local Dynamical Approximation to Backpropagation in the Brain
Beren Millidge,Alexander Tschantz,Anil K Seth,Christopher Buckley
The backpropagation of error algorithm (backprop) has been instrumental in the recent success of deep learning. However, a key question remains as to whether backprop can be formulated in a manner suitable for implementation in neural circuitry. The primary challenge is to ensure that any candidate formulation uses only local information, rather than relying on global signals as in standard backprop. Recently several algorithms for approximating backprop using only local signals have been proposed. However, these algorithms typically impose other requirements which challenge biological plausibility: for example, requiring complex and precise connectivity schemes, or multiple sequential backwards phases with information being stored across phases. Here, we propose a novel algorithm, Activation Relaxation (AR), which is motivated by constructing the backpropagation gradient as the equilibrium point of a dynamical system. Our algorithm converges rapidly and robustly to the correct backpropagation gradients, requires only a single type of computational unit, utilises only a single parallel backwards relaxation phase, and can operate on arbitrary computation graphs. We illustrate these properties by training deep neural networks on visual classification tasks, and describe simplifications to the algorithm which remove further obstacles to neurobiological implementation (for example, the weight-transport problem, and the use of nonlinear derivatives), while preserving performance.
****
CT-Net: Channel Tensorization Network for Video Classification
Kunchang Li,Xianhang Li,Yali Wang,Jun Wang,Yu Qiao
3D convolution is powerful for video classification but often computationally expensive, recent studies mainly focus on decomposing it on spatial-temporal and/or channel dimensions. Unfortunately, most approaches fail to achieve a preferable balance between convolutional efficiency and feature-interaction sufficiency. For this reason, we propose a concise and novel Channel Tensorization Network (CT-Net), by treating the channel dimension of input feature as a multiplication of K sub-dimensions. On one hand, it naturally factorizes convolution in a multiple dimension way, leading to a light computation burden. On the other hand, it can effectively enhance feature interaction from different channels, and progressively enlarge the 3D receptive field of such interaction to boost classification accuracy. Furthermore, we equip our CT-Module with a Tensor Excitation (TE) mechanism. It can learn to exploit spatial, temporal and channel attentio

n in a high-dimensional manner, to improve the cooperative power of all the feat
ure dimensions in our CT-Module. Finally, we flexibly adapt ResNet as our CT-Net
. Extensive experiments are conducted on several challenging video benchmarks, e
.g., Kinetics-400, Something-Something V1 and V2. Our CT-Net outperforms a numbe
r of recent SOTA approaches, in terms of accuracy and/or efficiency.
****

Learning Invariant Representations for Reinforcement Learning without Reconstruc
tion

Amy Zhang,Rowan Thomas McAllister,Roberto Calandra,Yarin Gal,Sergey Levine

We study how representation learning can accelerate reinforcement learning from
rich observations, such as images, without relying either on domain knowledge or
 pixel-reconstruction. Our goal is to learn representations that provide for eff
ective downstream control and invariance to task-irrelevant details. Bisimulatio
n metrics quantify behavioral similarity between states in continuous MDPs, whic
h we propose using to learn robust latent representations which encode only the
task-relevant information from observations. Our method trains encoders such tha
t distances in latent space equal bisimulation distances in state space. We demo
nstrate the effectiveness of our method at disregarding task-irrelevant informat
ion using modified visual MuJoCo tasks, where the background is replaced with mo
ving distractors and natural videos, while achieving SOTA performance. We also t
est a first-person highway driving task where our method learns invariance to cl
ouds, weather, and time of day. Finally, we provide generalization results drawn
 from properties of bisimulation metrics, and links to causal inference.
****

Intraclass clustering: an implicit learning ability that regularizes DNNs

Simon Carbonnelle,Christophe De Vleeschouwer

Several works have shown that the regularization mechanisms underlying deep neur
al networks' generalization performances are still poorly understood. In this pa
per, we hypothesize that deep neural networks are regularized through their abil
ity to extract meaningful clusters among the samples of a class. This constitute
s an implicit form of regularization, as no explicit training mechanisms or supe
rvision target such behaviour. To support our hypothesis, we design four differe
nt measures of intraclass clustering, based on the neuron- and layer-level repre
sentations of the training data. We then show that these measures constitute acc
urate predictors of generalization performance across variations of a large set
of hyperparameters (learning rate, batch size, optimizer, weight decay, dropout
rate, data augmentation, network depth and width).
****

Daylight: Assessing Generalization Skills of Deep Reinforcement Learning Agents

Ezgi Korkmaz

Deep reinforcement learning algorithms have recently achieved significant succes
s in learning high-performing policies from purely visual observations. The abil
ity to perform end-to-end learning from raw high dimensional input alone has led
 to deep reinforcement learning algorithms being deployed in a variety of fields
. Thus, understanding and improving the ability of deep reinforcement learning a
gents to generalize to unseen data distributions is of critical importance. Much
 recent work has focused on assessing the generalization of deep reinforcement l
earning agents by introducing specifically crafted adversarial perturbations to
their inputs. In this paper, we propose another approach that we call daylight:
a framework to assess the generalization skills of trained deep reinforcement le
arning agents. Rather than focusing on worst-case analysis of distribution shift
, our approach is based on black-box perturbations that correspond to semantical
ly meaningful changes to the environment or the agent's visual observation syste
m ranging from brightness to compression artifacts. We demonstrate that even the
 smallest changes in the environment cause the performance of the agents to degr
ade significantly in various games from the Atari environment despite having ord
ers of magnitude lower perceptual similarity distance compared to state-of-the-a
rt adversarial attacks. We show that our framework captures a diverse set of ban
ds in the Fourier spectrum, giving a better overall understanding of the agent's
 generalization capabilities. We believe our work can be crucial towards buildin

g resilient and generalizable deep reinforcement learning agents.
****
Learning Robust State Abstractions for Hidden-Parameter Block MDPs
Amy Zhang,Shagun Sodhani,Khimya Khetarpal,Joelle Pineau
Many control tasks exhibit similar dynamics that can be modeled as having common latent structure. Hidden-Parameter Markov Decision Processes (HiP-MDPs) explicitly model this structure to improve sample efficiency in multi-task settings. However, this setting makes strong assumptions on the observability of the state that limit its application in real-world scenarios with rich observation spaces. In this work, we leverage ideas of common structure from the HiP-MDP setting, and extend it to enable robust state abstractions inspired by Block MDPs. We derive instantiations of this new framework for both multi-task reinforcement learning (MTRL) and meta-reinforcement learning (Meta-RL) settings. Further, we provide transfer and generalization bounds based on task and state similarity, along with sample complexity bounds that depend on the aggregate number of samples across tasks, rather than the number of tasks, a significant improvement over prior work. To further demonstrate efficacy of the proposed method, we empirically compare and show improvement over multi-task and meta-reinforcement learning baselines.
****
Few-Round Learning for Federated Learning
Younghyun Park,Dong-Jun Han,Do-Yeon Kim,Jun Seo,Jaekyun Moon
Federated learning (FL) presents an appealing opportunity for individuals who are willing to make their private data available for building a communal model without revealing their data contents to anyone else. Of central issues that may limit a widespread adoption of FL is the significant communication resources required in the exchange of updated model parameters between the server and individual clients over many communication rounds. In this work, we focus on limiting the number of model exchange rounds in FL to some small fixed number R, to control the communication burden. Following the spirit of meta-learning for few-shot learning, we take a meta-learning strategy to train the model so that once the meta-training phase is over, only R rounds of FL would produce a model that will satisfy the needs of all participating clients. A key advantage of employing meta-training is that the main labeled dataset used in training could differ significantly (e.g., different classes of images) from the actual data sample presented at inference time. Compared to the meta-training approaches to optimize personalized local models at distributed devices, our method better handles the potential lack of data variability at individual nodes. Extensive experimental results indicate that meta-training geared to few-round learning provide large performance improvements compared to various baselines.
****
Predictive Coding Approximates Backprop along Arbitrary Computation Graphs
Beren Millidge,Alexander Tschantz,Christopher Buckley
The backpropagation of error (backprop) is a powerful algorithm for training machine learning architectures through end-to-end differentiation. Recently it has been shown that backprop in multilayer-perceptrons (MLPs) can be approximated using predictive coding, a biologically-plausible process theory of cortical computation which relies solely on local and Hebbian updates. The power of backprop, however, lies not in its instantiation in MLPs, but rather in the concept of automatic differentiation which allows for the optimisation of any differentiable program expressed as a computation graph. Here, we demonstrate that predictive coding converges asymptotically (and in practice rapidly) to exact backprop gradients on arbitrary computation graphs using only local learning rules. We apply this result to develop a straightforward strategy to translate core machine learning architectures into their predictive coding equivalents. We construct predictive coding CNNs, RNNs, and the more complex LSTMs, which include a non-layer-like branching internal graph structure and multiplicative interactions. Our models perform equivalently to backprop on challenging machine learning benchmarks, while utilising only local and (mostly) Hebbian plasticity. Our method raises the potential that standard machine learning algorithms could in principle be directl

y implemented in neural circuitry, and may also contribute to the development of completely distributed neuromorphic architectures.
****

CAFENet: Class-Agnostic Few-Shot Edge Detection Network
Younghyun Park,Jun Seo,Jaekyun Moon
We tackle a novel few-shot learning challenge, few-shot semantic edge detection, aiming to localize boundaries of novel categories using only a few labeled samples. Reliable boundary information has been shown to boost the performance of semantic segmentation and localization, while also playing a key role in its own right in object reconstruction, image generation and medical imaging. Few-shot semantic edge detection allows recovery of accurate boundaries with just a few examples. In this work, we present a Class-Agnostic Few-shot Edge detection Network (CAFENet) based on meta-learning strategy. CAFENet employs a semantic segmentation module in small-scale to compensate for lack of semantic information in edge labels. The predicted segmentation mask is used to generate an attention map to highlight the target object region, and make the decoder module concentrate on that region. We also propose a new regularization method based on multi-split matching. In meta-training, the metric-learning problem with high-dimensional vectors are divided into smaller subproblems with low-dimensional sub-vectors. Since there are no existing datasets for few-shot semantic edge detection, we construct two new datasets, FSE-1000 and SBD-5i, and evaluate the performance of the proposed CAFENet on them. Extensive simulation results confirm that the proposed CAFENet achieves better performance compared to the baseline methods using fine-tuning or few-shot segmentation.
****

Isometric Transformation Invariant and Equivariant Graph Convolutional Networks
Masanobu Horie,Naoki Morita,Toshiaki Hishinuma,Yu Ihara,Naoto Mitsume
Graphs are one of the most important data structures for representing pairwise relations between objects. Specifically, a graph embedded in a Euclidean space is essential to solving real problems, such as physical simulations. A crucial requirement for applying graphs in Euclidean spaces to physical simulations is learning and inferring the isometric transformation invariant and equivariant features in a computationally efficient manner. In this paper, we propose a set of transformation invariant and equivariant models based on graph convolutional networks, called IsoGCNs. We demonstrate that the proposed model has a competitive performance compared to state-of-the-art methods on tasks related to geometrical and physical simulation data. Moreover, the proposed model can scale up to graphs with 1M vertices and conduct an inference faster than a conventional finite element analysis, which the existing equivariant models cannot achieve.
****

Learning Safe Multi-agent Control with Decentralized Neural Barrier Certificates
Zengyi Qin,Kaiqing Zhang,Yuxiao Chen,Jingkai Chen,Chuchu Fan
We study the multi-agent safe control problem where agents should avoid collisions to static obstacles and collisions with each other while reaching their goals. Our core idea is to learn the multi-agent control policy jointly with learning the control barrier functions as safety certificates. We propose a new joint-learning framework that can be implemented in a decentralized fashion, which can adapt to an arbitrarily large number of agents. Building upon this framework, we further improve the scalability by incorporating neural network architectures that are invariant to the quantity and permutation of neighboring agents. In addition, we propose a new spontaneous policy refinement method to further enforce the certificate condition during testing. We provide extensive experiments to demonstrate that our method significantly outperforms other leading multi-agent control approaches in terms of maintaining safety and completing original tasks. Our approach also shows substantial generalization capability in that the control policy can be trained with 8 agents in one scenario, while being used on other scenarios with up to 1024 agents in complex multi-agent environments and dynamics. Videos and source code can be found at https://realm.mit.edu/blog/learning-safe-multi-agent-control-decentralized-neural-barrier-certificates.
****

Improving Mutual Information based Feature Selection by Boosting Unique Relevanc
e
Shiyu Liu,Mehul Motani
Mutual Information (MI) based feature selection makes use of MI to evaluate each
 feature and eventually shortlist a relevant feature subset, in order to address
 issues associated with high-dimensional datasets.
Despite the effectiveness of MI in feature selection, we have noticed that many
state-of-the-art algorithms disregard the so-called unique relevance (UR) of fea
tures, which is a necessary condition for the optimal feature subset. In fact, i
n our study of seven state-of-the-art and classical MIBFS algorithms, we find th
at all of them underperform as they ignore UR of features and arrive at a subopt
imal selected feature subset which contains a non-negligible number of redundant
 features. We point out that the heart of the problem is that all these MIBFS al
gorithms follow the criterion of Maximize Relevance with Minimum Redundancy (MRw
MR), which does not explicitly target UR. This motivates us to augment the exist
ing criterion with the objective of boosting unique relevance (BUR), leading to
a new criterion called MRwMR-BUR. We conduct extensive experiments with several
MIBFS algorithms with and without incorporating UR. The results indicate that th
e algorithms that boost UR consistently outperform their unboosted counterparts
in terms of peak accuracy and number of features required. Furthermore, we propo
se a classifier based approach to estimate UR that further improves the performa
nce of MRwMR-BUR based algorithms.
****
Learning Incompressible Fluid Dynamics from Scratch - Towards Fast, Differentiab
le Fluid Models that Generalize
Nils Wandel,Michael Weinmann,Reinhard Klein
Fast and stable fluid simulations are an essential prerequisite for applications
 ranging from computer-generated imagery to computer-aided design in research an
d development. However, solving the partial differential equations of incompress
ible fluids is a challenging task and traditional numerical approximation scheme
s come at high computational costs. Recent deep learning based approaches promis
e vast speed-ups but do not generalize to new fluid domains, require fluid simul
ation data for training, or rely on complex pipelines that outsource major parts
 of the fluid simulation to traditional methods.

In this work, we propose a novel physics-constrained training approach that gene
ralizes to new fluid domains, requires no fluid simulation data, and allows conv
olutional neural networks to map a fluid state from time-point t to a subsequent
 state at time t+dt in a single forward pass. This simplifies the pipeline to tr
ain and evaluate neural fluid models. After training, the framework yields model
s that are capable of fast fluid simulations and can handle various fluid phenom
ena including the Magnus effect and Kármán vortex streets. We present an interac
tive real-time demo to show the speed and generalization capabilities of our tra
ined models. Moreover, the trained neural networks are efficient differentiable
fluid solvers as they offer a differentiable update step to advance the fluid si
mulation in time. We exploit this fact in a proof-of-concept optimal control exp
eriment. Our models significantly outperform a recent differentiable fluid solve
r in terms of computational speed and accuracy.
****
ChipNet: Budget-Aware Pruning with Heaviside Continuous Approximations
Rishabh Tiwari,Udbhav Bamba,Arnav Chavan,Deepak Gupta
Structured pruning methods are among the effective strategies for extracting sma
ll resource-efficient convolutional neural networks from their dense counterpart
s with minimal loss in accuracy. However, most existing methods still suffer fro
m one or more limitations, that include 1) the need for training the dense model
 from scratch with pruning-related parameters embedded in the architecture, 2) r
equiring model-specific hyperparameter settings, 3) inability to include budget-
related constraint in the training process, and 4) instability under scenarios o
f extreme pruning. In this paper, we present ChipNet, a deterministic pruning st
rategy that employs continuous Heaviside function and a novel crispness loss to

identify a highly sparse network out of an existing dense network. Our choice of continuous Heaviside function is inspired by the field of design optimization, where the material distribution task is posed as a continuous optimization problem, but only discrete values (0 or 1) are practically feasible and expected as final outcomes. Our approach's flexible design facilitates its use with different choices of budget constraints while maintaining stability for very low target budgets. Experimental results show that ChipNet outperforms state-of-the-art structured pruning methods by remarkable margins of up to 16.1% in terms of accuracy. Further, we show that the masks obtained with ChipNet are transferable across datasets. For certain cases, it was observed that masks transferred from a model trained on feature-rich teacher dataset provide better performance on the student dataset than those obtained by directly pruning on the student data itself.
****

AdaSpeech: Adaptive Text to Speech for Custom Voice
Mingjian Chen,Xu Tan,Bohan Li,Yanqing Liu,Tao Qin,sheng zhao,Tie-Yan Liu
Custom voice, a specific text to speech (TTS) service in commercial speech platforms, aims to adapt a source TTS model to synthesize personal voice for a target speaker using few speech from her/him. Custom voice presents two unique challenges for TTS adaptation: 1) to support diverse customers, the adaptation model needs to handle diverse acoustic conditions which could be very different from source speech data, and 2) to support a large number of customers, the adaptation parameters need to be small enough for each target speaker to reduce memory usage while maintaining high voice quality. In this work, we propose AdaSpeech, an adaptive TTS system for high-quality and efficient customization of new voices. We design several techniques in AdaSpeech to address the two challenges in custom voice: 1) To handle different acoustic conditions, we model the acoustic information in both utterance and phoneme level. Specifically, we use one acoustic encoder to extract an utterance-level vector and another one to extract a sequence of phoneme-level vectors from the target speech during pre-training and fine-tuning; in inference, we extract the utterance-level vector from a reference speech and use an acoustic predictor to predict the phoneme-level vectors. 2) To better trade off the adaptation parameters and voice quality, we introduce conditional layer normalization in the mel-spectrogram decoder of AdaSpeech, and fine-tune this part in addition to speaker embedding for adaptation. We pre-train the source TTS model on LibriTTS datasets and fine-tune it on VCTK and LJSpeech datasets (with different acoustic conditions from LibriTTS) with few adaptation data, e.g., 20 sentences, about 1 minute speech. Experiment results show that AdaSpeech achieves much better adaptation quality than baseline methods, with only about 5K specific parameters for each speaker, which demonstrates its effectiveness for custom voice. The audio samples are available at https://speechresearch.github.io/adaspeech/.
****

Density Constrained Reinforcement Learning
Zengyi Qin,Yuxiao Chen,Chuchu Fan
Constrained reinforcement learning (CRL) plays an important role in solving safety-critical and resource-limited tasks. However, existing methods typically rely on tuning reward or cost parameters to encode the constraints, which can be tedious and tend to not generalize well. Instead of building sophisticated cost functions for constraints, we present a pioneering study of imposing constraints directly on the state density function of the system. Density functions have clear physical meanings and can express a variety of constraints in a straightforward fashion. We prove the duality between the density function and Q function in CRL and use it to develop an effective primal-dual algorithm to solve density constrained reinforcement learning problems. We provide theoretical guarantees of the optimality of our approach and use a comprehensive set of case studies including standard benchmarks to show that our method outperforms other leading CRL methods in terms of achieving higher reward while respecting the constraints.
****

Convergent Adaptive Gradient Methods in Decentralized Optimization
Xiangyi Chen,Belhal Karimi,Weijie Zhao,Ping Li

Adaptive gradient methods including Adam, AdaGrad, and their variants have been very successful for training deep learning models, such as neural networks, in the past few years. Meanwhile, given the need for distributed training procedures, distributed optimization algorithms are at the center of attention. With the growth of computing power and the need for using machine learning models on mobile devices, the communication cost of distributed training algorithms needs careful consideration. In that regard, more and more attention is shifted from the traditional parameter server training paradigm to the decentralized one, which usually requires lower communication costs. In this paper, we rigorously incorporate adaptive gradient methods into decentralized training procedures and introduce novel convergent decentralized adaptive gradient methods. Specifically, we propose a general algorithmic framework that can convert existing adaptive gradient methods to their decentralized counterparts. In addition, we thoroughly analyze the convergence behavior of the proposed algorithmic framework and show that if a given adaptive gradient method converges, under some specific conditions, then its decentralized counterpart is also convergent.
****
Cross-model Back-translated Distillation for Unsupervised Machine Translation
Phi Xuan Nguyen,Shafiq Joty,Kui Wu,AiTi Aw
Recent unsupervised machine translation (UMT) systems usually employ three main principles: initialization, language modeling and iterative back-translation, though they may apply them differently. Crucially, iterative back-translation and denoising auto-encoding for language modeling provide data diversity to train the UMT systems. However, these diversification processes may have reached their limit. We introduce a novel component to the standard UMT framework called Cross-model Back-translated Distillation (CBD), that is aimed to induce another level of data diversification that existing principles lack. CBD is applicable to all previous UMT approaches. In our experiments, it boosts the performance of the standard UMT methods by 1.5-2.0 BLEU. In particular, in WMT'14 English-French, WMT'16 German-English and English-Romanian, CBD outperforms cross-lingual masked language model (XLM) by 2.3, 2.2 and 1.6 BLEU, respectively. It also yields 1.5-3.3 BLEU improvements in IWSLT English-French and English-German tasks. Through extensive experimental analyses, we show that CBD is effective because it embraces data diversity while other similar variants do not.
****
Efficient randomized smoothing by denoising with learned score function
Kyungmin Lee,Seyoon Oh
The randomized smoothing with various noise distributions is a promising approach to protect classifiers from ℓ_p adversarial attacks. However, it requires an ensemble of classifiers trained with different noise types and magnitudes, which is computationally expensive. In this work, we present an efficient method for randomized smoothing that does not require any re-training of classifiers. We built upon denoised smoothing, which prepends denoiser to the pre-trained classifier. We investigate two approaches to the image denoising problem for randomized smoothing and show that using the score function suits for both. Moreover, we present an efficient algorithm that can scale to randomized smoothing and can be applied regardless of noise types or levels. To validate, we demonstrate the effectiveness of our methods through extensive experiments on CIFAR-10 and ImageNet, under various ℓ_p adversaries.
****
Using Synthetic Data to Improve the Long-range Forecasting of Time Series Data
Shiyu Liu,Mehul Motani
Effective long-range forecasting of time series data remains an unsolved and open problem. One possible approach is to use generative models to improve long-range forecasting, but the challenge then is how to generate high-quality synthetic data. In this paper, we propose a conditional Wasserstein GAN with Gradient and Error Penalty (cWGAN-GEP), aiming to generate accurate synthetic data that preserves the temporal dynamics between the conditioning input and generated data. By using such synthetic data, we develop a long-range forecasting method called Generative Forecasting (GenF). GenF consists of three key components: (i) a cWGAN

-GEP based generator, to generate synthetic data for next few time steps. (ii) a predictor which makes long-range predictions based on generated and observed data. (iii) an information theoretic clustering (ITC) algorithm to better train the cWGAN-GEP based generator and the predictor. Our experimental results on three public datasets demonstrate that GenF significantly outperforms a diverse range of state-of-the-art benchmarks and classical approaches. In most cases, we find an improvement of at least 10% over all studied methods. Lastly, we conduct an ablation study to demonstrate the effectiveness of the cWGAN-GEP and the ITC algorithm.
****

Adversarial Feature Desensitization
Pouya Bashivan,Mojtaba Faramarzi,Touraj Laleh,Blake Aaron Richards,Irina Rish
Deep neural networks can now perform many tasks that were once thought to be only feasible for humans. While reaching impressive performance under standard settings, such networks are known to be susceptible to adversarial attacks -- slight but carefully constructed perturbations of the inputs which drastically decrease the network performance. Here we propose a new way to improve the network robustness against adversarial attacks by focusing on robust representation learning based on adversarial training procedure, called here Adversarial Feature Desensitization (AFD). AFD desensitizes the representation via an adversarial game between the embedding network and an adversarial discriminator introduced on top of the standard predictive model, which is trained to distinguish between the clean and perturbed inputs from their high-level representations. Our method substantially improves the state-of-the-art in robust classification on MNIST, CIFAR10, and CIFAR100 datasets. More importantly, we demonstrate that AFD has better generalization ability than previous methods, as the learned features maintain their robustness across a wide range of perturbations, including perturbations not seen during training. These results indicate that reducing feature sensitivity is a promising approach for ameliorating the problem of adversarial attacks in deep neural networks.
****

Task-Agnostic and Adaptive-Size BERT Compression
Jin Xu,Xu Tan,Renqian Luo,Kaitao Song,Li Jian,Tao Qin,Tie-Yan Liu
While pre-trained language models such as BERT and RoBERTa have achieved impressive results on various natural language processing tasks, they have huge numbers of parameters and suffer from huge computational and memory costs, which make them difficult for real-world deployment. Hence, model compression should be performed in order to reduce the computation and memory cost of pre-trained models. In this work, we aim to compress BERT and address the following two challenging practical issues: (1) The compression algorithm should be able to output multiple compressed models with different sizes and latencies, so as to support devices with different kinds of memory and latency limitations; (2) the algorithm should be downstream task agnostic, so that the compressed models are generally applicable for different downstream tasks. We leverage techniques in neural architecture search (NAS) and propose NAS-BERT, an efficient method for BERT compression. NAS-BERT trains a big supernet on a carefully designed search space containing various architectures and outputs multiple compressed models with adaptive sizes and latency. Furthermore, the training of NAS-BERT is conducted on standard self-supervised pre-training tasks (e.g., masked language model) and does not depend on specific downstream tasks. Thus, the models it produces can be used across various downstream tasks. The technical challenge of NAS-BERT is that training a big supernet on the pre-training task is extremely costly. We employ several techniques including block-wise search, search space pruning, and performance approximation to improve search efficiency and accuracy. Extensive experiments on GLUE benchmark datasets demonstrate that NAS-BERT can find lightweight models with better accuracy than previous approaches, and can be directly applied to different downstream tasks with adaptive model sizes for different requirements of memory or latency.
****

Estimating and Evaluating Regression Predictive Uncertainty in Deep Object Detec

tors
Ali Harakeh,Steven L. Waslander
Predictive uncertainty estimation is an essential next step for the reliable deployment of deep object detectors in safety-critical tasks. In this work, we focus on estimating predictive distributions for bounding box regression output with variance networks. We show that in the context of object detection, training variance networks with negative log likelihood (NLL) can lead to high entropy predictive distributions regardless of the correctness of the output mean. We propose to use the energy score as a non-local proper scoring rule and find that when used for training, the energy score leads to better calibrated and lower entropy predictive distributions than NLL. We also address the widespread use of non-proper scoring metrics for evaluating predictive distributions from deep object detectors by proposing an alternate evaluation approach founded on proper scoring rules. Using the proposed evaluation tools, we show that although variance networks can be used to produce high quality predictive distributions, ad-hoc approaches used by seminal object detectors for choosing regression targets during training do not provide wide enough data support for reliable variance learning. We hope that our work helps shift evaluation in probabilistic object detection to better align with predictive uncertainty evaluation in other machine learning domains. Code for all models, evaluation, and datasets is available at: https://github.com/asharakeh/probdet.git.
****
Geometry matters: Exploring language examples at the decision boundary
Debajyoti Datta,Shashwat Kumar,Laura Barnes,Tom Fletcher
A growing body of recent evidence has highlighted the limitations of natural language processing (NLP) datasets and classifiers. These include the presence of annotation artifacts in datasets, classifiers relying on shallow features like a single word (e.g., if a movie review has the word "romantic", the review tends to be positive), or unnecessary words (e.g., learning a proper noun to classify a movie as positive or negative). The presence of such artifacts has subsequently led to the development of challenging datasets to force the model to generalize better. While a variety of heuristic strategies, such as counterfactual examples and contrast sets, have been proposed, the theoretical justification about what makes these examples difficult for the classifier is often lacking or unclear. In this paper, using tools from information geometry, we propose a theoretical way to quantify the difficulty of an example in NLP. Using our approach, we explore difficult examples for several deep learning architectures. We discover that BERT, CNN and fasttext are susceptible to word substitutions in high difficulty examples. These classifiers tend to perform poorly on the FIM test set. (generated by sampling and perturbing difficult examples, with accuracy dropping below 50%). We replicate our experiments on 5 NLP datasets (YelpReviewPolarity, AGNEWS, SogouNews, YelpReviewFull and Yahoo Answers). On YelpReviewPolarity we observe a correlation coefficient of -0.4 between resilience to perturbations and the difficulty score. Similarly we observe a correlation of 0.35 between the difficulty score and the empirical success probability of random substitutions. Our approach is simple, architecture agnostic and can be used to study the fragilities of text classification models. All the code used will be made publicly available, including a tool to explore the difficult examples for other datasets.

****
Self-Labeling of Fully Mediating Representations by Graph Alignment
Martijn Oldenhof,Adam Arany,Yves Moreau,Jaak Simm
To be able to predict a molecular graph structure (W) given a 2D image of a chemical compound (U) is a challenging problem in machine learning. We are interested to learn $f: U \rightarrow W$ where we have a fully mediating representation V such that f factors into $U \rightarrow V \rightarrow W$. However, observing V requires detailed and expensive labels. We propose \textbf{graph aligning} approach that generates rich or detailed labels given normal labels W. In this paper we investigate the scenario of domain adaptation from the source domain where we have access to the expensive labels V to the target domain where o

nly normal labels W are available. Focusing on the problem of predicting chemica l compound graphs from 2D images the fully mediating layer is represented using the planar embedding of the chemical graph structure we are predicting. The use of a fully mediating layer implies some assumptions on the mechanism of the unde rlying process. However if the assumptions are correct it should allow the machi ne learning model to be more interpretable, generalize better and be more data efficient at training time.
The empirical results show that, using only 4000 data points, we obtain up to 4x improvement of performance after domain adaptation to target domain compared to pretrained model only on the source domain. After domain adaptation, the model is even able to detect atom types that were never seen in the original source do main. Finally, on the Maybridge data set the proposed self-labeling approach rea ched higher performance than the current state of the art.
****

Defining Benchmarks for Continual Few-Shot Learning
Antreas Antoniou,Massimiliano Patacchiola,Mateusz Ochal,Amos Storkey
In recent years there has been substantial progress in few-shot learning, where a model is trained on a small labeled dataset related to a specific task, and in continual learning, where a model has to retain knowledge acquired on a sequenc e of datasets. Both of these fields are different abstractions of the same real world scenario, where a learner has to adapt to limited information from differe nt changing sources and be able to generalize in and from each of them. Combinin g these two paradigms, where a model is trained on several sequential few-shot t asks, and then tested on a validation set stemming from all those tasks, helps b y explicitly defining the competing requirements for both efficient integration and continuity. In this paper we propose such a setting, naming it Continual Few -Shot Learning (CFSL). We first define a theoretical framework for CFSL, then we propose a range of flexible benchmarks to unify the evaluation criteria. As par t of the benchmark, we introduce a compact variant of ImageNet, called SlimageNe t64, which retains all original 1000 classes but only contains 200 instances of each one (a total of 200K data-points) downscaled to 64 by 64 pixels. We provide baselines for the proposed benchmarks using a number of popular few-shot and co ntinual learning methods, exposing previously unknown strengths and weaknesses o f those algorithms. The dataloader and dataset will be released with an open-sou rce license.
****

Learning Efficient Planning-based Rewards for Imitation Learning
Xingrui Yu,Yueming Lyu,Ivor Tsang
Imitation learning from limited demonstrations is challenging. Most inverse rein forcement learning (IRL) methods are unable to perform as good as the demonstrat or, especially in a high-dimensional environment, e.g, the Atari domain. To addr ess this challenge, we propose a novel reward learning method, which streamlines a differential planning module with dynamics modeling. Our method learns useful planning computations with a meaningful reward function that focuses on the res ulting region of an agent executing an action. Such a planning-based reward func tion leads to policies with better generalization ability. Empirical results wit h multiple network architectures and reward instances show that our method can o utperform state-of-the-art IRL methods on multiple Atari games and continuous co ntrol tasks. Our method achieves performance that is averagely 1,139.1% of the d emonstration.
****

Domain-Robust Visual Imitation Learning with Mutual Information Constraints
Edoardo Cetin,Oya Celiktutan
Human beings are able to understand objectives and learn by simply observing oth ers perform a task. Imitation learning methods aim to replicate such capabilitie s, however, they generally depend on access to a full set of optimal states and actions taken with the agent's actuators and from the agent's point of view. In this paper, we introduce a new algorithm - called Disentangling Generative Adver sarial Imitation Learning (DisentanGAIL) - with the purpose of bypassing such co nstraints. Our algorithm enables autonomous agents to learn directly from high d

imensional observations of an expert performing a task, by making use of adversa rial learning with a latent representation inside the discriminator network. Suc h latent representation is regularized through mutual information constraints to incentivize learning only features that encode information about the completion levels of the task being demonstrated. This allows to obtain a shared feature s pace to successfully perform imitation while disregarding the differences betwee n the expert's and the agent's domains. Empirically, our algorithm is able to ef ficiently imitate in a diverse range of control problems including balancing, ma nipulation and locomotive tasks, while being robust to various domain difference s in terms of both environment appearance and agent embodiment.
****

Clustering-friendly Representation Learning via Instance Discrimination and Feat ure Decorrelation
Yaling Tao,Kentaro Takagi,Kouta Nakata
Clustering is one of the most fundamental tasks in machine learning. Recently, d eep clustering has become a major trend in clustering techniques. Representation learning often plays an important role in the effectiveness of deep clustering, and thus can be a principal cause of performance degradation. In this paper, we propose a clustering-friendly representation learning method using instance dis crimination and feature decorrelation. Our deep-learning-based representation le arning method is motivated by the properties of classical spectral clustering. I nstance discrimination learns similarities among data and feature decorrelation removes redundant correlation among features. We utilize an instance discriminat ion method in which learning individual instance classes leads to learning simil arity among instances. Through detailed experiments and examination, we show tha t the approach can be adapted to learning a latent space for clustering. We desi gn novel softmax-formulated decorrelation constraints for learning. In evaluatio ns of image clustering using CIFAR-10 and ImageNet-10, our method achieves accur acy of 81.5% and 95.4%, respectively. We also show that the softmax-formulated c onstraints are compatible with various neural networks.
****

CDT: Cascading Decision Trees for Explainable Reinforcement Learning
Zihan Ding,Pablo Hernandez-Leal,Gavin Weiguang Ding,Changjian Li,Ruitong Huang
Deep Reinforcement Learning (DRL) has recently achieved significant advances in various domains. However, explaining the policy of RL agents still remains an op en problem due to several factors, one being the complexity of explaining neural networks decisions. Recently, a group of works have used decision-tree-based mo dels to learn explainable policies. Soft decision trees (SDTs) and discretized d ifferentiable decision trees (DDTs) have been demonstrated to achieve both good performance and share the benefit of having explainable policies. In this work, we further improve the results for tree-based explainable RL in both performance and explainability. Our proposal, Cascading Decision Trees (CDTs) apply represe ntation learning on the decision path to allow richer expressivity. Empirical re sults show that in both situations, where CDTs are used as policy function appro ximators or as imitation learners to explain black-box policies, CDTs can achiev e better performances with more succinct and explainable models than SDTs. As a second contribution our study reveals limitations of explaining black-box polici es via imitation learning with tree-based explainable models, due to its inheren t instability.
****

A Gradient Flow Framework For Analyzing Network Pruning
Ekdeep Singh Lubana,Robert Dick
Recent network pruning methods focus on pruning models early-on in training. To estimate the impact of removing a parameter, these methods use importance measur es that were originally designed to prune trained models. Despite lacking justif ication for their use early-on in training, such measures result in surprisingly low accuracy loss. To better explain this behavior, we develop a general framew ork that uses gradient flow to unify state-of-the-art importance measures throug h the norm of model parameters. We use this framework to determine the relations hip between pruning measures and evolution of model parameters, establishing sev

eral results related to pruning models early-on in training: (i) magnitude-based pruning removes parameters that contribute least to reduction in loss, resulting in models that converge faster than magnitude-agnostic methods; (ii) loss-preservation based pruning preserves first-order model evolution dynamics and its use is therefore justified for pruning minimally trained models; and (iii) gradient-norm based pruning affects second-order model evolution dynamics, such that increasing gradient norm via pruning can produce poorly performing models. We validate our claims on several VGG-13, MobileNet-V1, and ResNet-56 models trained on CIFAR-10/CIFAR-100.
****

Progressive Skeletonization: Trimming more fat from a network at initialization
Pau de Jorge,Amartya Sanyal,Harkirat Behl,Philip Torr,Grégory Rogez,Puneet K. Dokania
Recent studies have shown that skeletonization (pruning parameters) of networks at initialization provides all the practical benefits of sparsity both at inference and training time, while only marginally degrading their performance. However, we observe that beyond a certain level of sparsity (approx 95%), these approaches fail to preserve the network performance, and to our surprise, in many cases perform even worse than trivial random pruning. To this end, we propose an objective to find a skeletonized network with maximum foresight connection sensitivity (FORCE) whereby the trainability, in terms of connection sensitivity, of a pruned network is taken into consideration. We then propose two approximate procedures to maximize our objective (1) Iterative SNIP: allows parameters that were unimportant at earlier stages of skeletonization to become important at later stages; and (2) FORCE: iterative process that allows exploration by allowing already pruned parameters to resurrect at later stages of skeletonization. Empirical analysis on a large suite of experiments show that our approach, while providing at least as good performance as other recent approaches on moderate pruning levels, provide remarkably improved performance on high pruning levels (could remove up to 99.5% parameters while keeping the networks trainable).
****

An Adversarial Attack via Feature Contributive Regions
Yaguan Qian,Jiamin Wang,Xiang Ling,Zhaoquan Gu,Bin Wang,Chunming Wu
Recently, to deal with the vulnerability to generate examples of CNNs, there are many advanced algorithms that have been proposed. These algorithms focus on modifying global pixels directly with small perturbations, and some work involves modifying local pixels. However, the global attacks have the problem of perturbations' redundancy and the local attacks are not effective. To overcome this challenge, we achieve a trade-off between the perturbation power and the number of perturbed pixels in this paper. The key idea is to find the feature contributive regions (FCRs) of the images. Furthermore, in order to create an adversarial example similar to the corresponding clean image as much as possible, we redefine a loss function as the objective function of the optimization in this paper and then using gradient descent optimization algorithm to find the efficient perturbations. Various experiments have been carried out on CIFAR-10 and ILSVRC2012 datasets, which show the excellence of this method, and in addition, the FCRs attack shows strong attack ability in both white-box and black-box settings.
****

On the Curse of Memory in Recurrent Neural Networks: Approximation and Optimization Analysis
Zhong Li,Jiequn Han,Weinan E,Qianxiao Li
We study the approximation properties and optimization dynamics of recurrent neural networks (RNNs) when applied to learn input-output relationships in temporal data. We consider the simple but representative setting of using continuous-time linear RNNs to learn from data generated by linear relationships. Mathematically, the latter can be understood as a sequence of linear functionals. We prove a universal approximation theorem of such linear functionals and characterize the approximation rate. Moreover, we perform a fine-grained dynamical analysis of training linear RNNs by gradient methods. A unifying theme uncovered is the non-trivial effect of memory, a notion that can be made precise in our framework,

on both approximation and optimization: when there is long-term memory in the target, it takes a large number of neurons to approximate it. Moreover, the training process will suffer from slow downs. In particular, both of these effects become exponentially more pronounced with increasing memory - a phenomenon we call the “curse of memory”. These analyses represent a basic step towards a concrete mathematical understanding of new phenomenons that may arise in learning temporal relationships using recurrent architectures.
****

On the Stability of Multi-branch Network
Huishuai Zhang,Da Yu,Wei Chen,Tie-Yan Liu
Multi-branch architectures are widely used in state-of-the-art neural networks. Their empirical success relies on some design wisdom, like adding normalization layers or/and scaling down the initialization. In this paper, we investigate the multi-branch architecture from the stability perspective. Specifically, we establish the forward/backward stability of multi-branch network, which leads to several new findings. Our analysis shows that only scaling down the initialization may not be enough for training multi-branch network successfully because of the uncontrollable backward process. We also unveil a new role of the normalization layer in terms of stabilizing the multi-branch architectures. More importantly, we propose a new design ``STAM aggregation" that can guarantee to STAbilize the forward/backward process of Multi-branch networks irrespective of the number of branches. We demonstrate that with STAM aggregation, the same training strategy is applicable to models with different numbers of branches, which can reduce the hyper-parameter tuning burden. Our experiments verify our theoretical findings and also demonstrate that the STAM aggregation can improve the performance of multi-branch networks considerably.
****

Deepening Hidden Representations from Pre-trained Language Models
Junjie Yang,hai zhao
Transformer-based pre-trained language models have proven to be effective for learning contextualized language representation. However, current approaches only take advantage of the output of the encoder's final layer when fine-tuning the downstream tasks. We argue that only taking single layer's output restricts the power of pre-trained representation. Thus we deepen the representation learned by the model by fusing the hidden representation in terms of an explicit HIdden Representation Extractor (HIRE), which automatically absorbs the complementary representation with respect to the output from the final layer. Utilizing RoBERTa as the backbone encoder, our proposed improvement over the pre-trained models is shown effective on multiple natural language understanding tasks and help our model rival with the state-of-the-art models on the GLUE benchmark.
****

Do not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning
Da Yu,Huishuai Zhang,Wei Chen,Tie-Yan Liu
The privacy leakage of the model about the training data can be bounded in the differential privacy mechanism. However, for meaningful privacy parameters, a differentially private model degrades the utility drastically when the model comprises a large number of trainable parameters. In this paper, we propose an algorithm \emph{Gradient Embedding Perturbation (GEP)} towards training differentially private deep models with decent accuracy. Specifically, in each gradient descent step, GEP first projects individual private gradient into a non-sensitive anchor subspace, producing a low-dimensional gradient embedding and a small-norm residual gradient. Then, GEP perturbs the low-dimensional embedding and the residual gradient separately according to the privacy budget. Such a decomposition permits a small perturbation variance, which greatly helps to break the dimensional barrier of private learning. With GEP, we achieve decent accuracy with low computational cost and modest privacy guarantee for deep models. Especially, with privacy bound $\epsilon=8$, we achieve 74.9% test accuracy on CIFAR10 and 95.1% test accuracy on SVHN, significantly improving over existing results.
****

What Do Deep Nets Learn? Class-wise Patterns Revealed in the Input Space
Shihao Zhao,Xingjun Ma,Yisen Wang,James Bailey,Bo Li,Yu-Gang Jiang
Deep neural networks (DNNs) have been widely adopted in different applications to achieve state-of-the-art performance. However, they are often applied as a black box with limited understanding of what the model has learned from the data. In this paper, we focus on image classification and propose a method to visualize and understand the class-wise patterns learned by DNNs trained under three different settings including natural, backdoored and adversarial. Different from existing class-wise deep representation visualizations, our method searches for a single predictive pattern in the input (i.e. pixel) space for each class. Based on the proposed method, we show that DNNs trained on natural (clean) data learn abstract shapes along with some texture, and backdoored models learn a small but highly predictive pattern for the backdoor target class. Interestingly, the existence of class-wise predictive patterns in the input space indicates that even DNNs trained on clean data can have backdoors, and the class-wise patterns identified by our method can be readily applied to “backdoor” attack the model. In the adversarial setting, we show that adversarially trained models learn more simplified shape patterns. Our method can serve as a useful tool to better understand DNNs trained on different datasets under different settings.
****
More or Less: When and How to Build Convolutional Neural Network Ensembles
Abdul Wasay,Stratos Idreos
Convolutional neural networks are utilized to solve increasingly more complex problems and with more data. As a result, researchers and practitioners seek to scale the representational power of such models by adding more parameters. However, increasing parameters requires additional critical resources in terms of memory and compute, leading to increased training and inference cost. Thus a consistent challenge is to obtain as high as possible accuracy within a parameter budget. As neural network designers navigate this complex landscape, they are guided by conventional wisdom that is informed from past empirical studies. We identify a critical part of this design space that is not well-understood: How to decide between the alternatives of expanding a single convolutional network model or increasing the number of networks in the form of an ensemble. We study this question in detail across various network architectures and data sets. We build an extensive experimental framework that captures numerous angles of the possible design space in terms of how a new set of parameters can be used in a model. We consider a holistic set of metrics such as training time, inference time, and memory usage. The framework provides a robust assessment by making sure it controls for the number of parameters. Contrary to conventional wisdom, we show that when we perform a holistic and robust assessment, we uncover a wide design space, where ensembles provide better accuracy, train faster, and deploy at speed comparable to single convolutional networks with the same total number of parameters.
****
A Simple Unified Information Regularization Framework for Multi-Source Domain Adaptation
Geon Yeong Park,Sang wan Lee
Adversarial learning strategy has demonstrated remarkable performance in dealing with single-source unsupervised Domain Adaptation (DA) problems, and it has recently been applied to multi-source DA problems. While most of the existing DA methods use multiple domain discriminators, the effect of using multiple discriminators on the quality of latent space representations has been poorly understood. Here we provide theoretical insights into potential pitfalls of using multiple domain discriminators: First, domain-discriminative information is inevitably distributed across multiple discriminators. Second, it is not scalable in terms of computational resources. Third, the variance of stochastic gradients from multiple discriminators may increase, which significantly undermines training stability. To fully address these issues, we situate adversarial DA in the context of information regularization. First, we present a unified information regularization framework for multi-source DA. It provides a theoretical justification for using a single and unified domain discriminator to encourage the synergistic integr

ation of the information gleaned from each domain. Second, this motivates us to implement a novel neural architecture called a Multi-source Information-regularized Adaptation Networks (MIAN). The proposed model significantly reduces the variance of stochastic gradients and increases computational-efficiency. Large-scale simulations on various multi-source DA scenarios demonstrate that MIAN, despite its structural simplicity, reliably outperforms other state-of-the-art methods by a large margin especially for difficult target domains.
****
Towards Robustness Against Natural Language Word Substitutions
Xinshuai Dong,Anh Tuan Luu,Rongrong Ji,Hong Liu
Robustness against word substitutions has a well-defined and widely acceptable form, i.e., using semantically similar words as substitutions, and thus it is considered as a fundamental stepping-stone towards broader robustness in natural language processing. Previous defense methods capture word substitutions in vector space by using either l_2-ball or hyper-rectangle, which results in perturbation sets that are not inclusive enough or unnecessarily large, and thus impedes mimicry of worst cases for robust training. In this paper, we introduce a novel Adversarial Sparse Convex Combination (ASCC) method. We model the word substitution attack space as a convex hull and leverages a regularization term to enforce perturbation towards an actual substitution, thus aligning our modeling better with the discrete textual space. Based on ASCC method, we further propose ASCC-defense, which leverages ASCC to generate worst-case perturbations and incorporates adversarial training towards robustness. Experiments show that ASCC-defense outperforms the current state-of-the-arts in terms of robustness on two prevailing NLP tasks, i.e., sentiment analysis and natural language inference, concerning several attacks across multiple model architectures. Besides, we also envision a new class of defense towards robustness in NLP, where our robustly trained word vectors can be plugged into a normally trained model and enforce its robustness without applying any other defense techniques.
****
On Low Rank Directed Acyclic Graphs and Causal Structure Learning
Zhuangyan Fang,Shengyu Zhu,Jiji Zhang,Yue Liu,Zhitang Chen,Yangbo He
Despite several important advances in recent years, learning causal structures represented by directed acyclic graphs (DAGs) remains a challenging task in high dimensional settings when the graphs to be learned are not sparse. In this paper, we propose to exploit a low rank assumption regarding the (weighted) adjacency matrix of a DAG causal model to mitigate this problem. We demonstrate how to adapt existing methods for causal structure learning to take advantage of this assumption and establish several useful results relating interpretable graphical conditions to the low rank assumption. In particular, we show that the maximum rank is highly related to hubs, suggesting that scale-free networks which are frequently encountered in real applications tend to be low rank. We also provide empirical evidence for the utility of our low rank adaptations, especially on relatively large and dense graphs. Not only do they outperform existing algorithms when the low rank condition is satisfied, the performance is also competitive even though the rank of the underlying DAG may not be as low as is assumed.
****
Unsupervised Hierarchical Concept Learning
Sumegh Roychowdhury,Sumedh Anand Sontakke,Mausoom Sarkar,Nikaash Puri,Milan Aggarwal,Pinkesh Badjatiya,Balaji Krishnamurthy,Laurent Itti
Discovering concepts (or temporal abstractions) in an unsupervised manner from demonstration data in the absence of an environment is an important problem. Organizing these discovered concepts hierarchically at different levels of abstraction is useful in discovering patterns, building ontologies, and generating tutorials from demonstration data. However, recent work to discover such concepts without access to any environment does not discover relationships (or a hierarchy) between these discovered concepts. In this paper, we present a Transformer-based concept abstraction architecture UNHCLE (pronounced uncle) that extracts a hierarchy of concepts in an unsupervised way from demonstration data. We empirically demonstrate how UNHCLE discovers meaningful hierarchies using datasets from Che

ss and Cooking domains. Finally, we show how UNHCLE learns meaningful language l
abels for concepts by using demonstration data augmented with natural language f
or cooking and chess. All of our code is available at https://github.com/UNHCLE
/UNHCLE

****
On Dropout, Overfitting, and Interaction Effects in Deep Neural Networks
Ben Lengerich,Eric Xing,Rich Caruana
We examine Dropout through the perspective of interactions. Given N variables,
 there are $\mathcal{O}(N^2)$ possible pairwise interactions, $\mathcal{O}(N^3)$
 possible 3-way interactions, i.e. $\mathcal{O}(N^k)$ possible interactions of $
k$ variables. Conversely, the probability of an interaction of k variables sur
viving Dropout at rate p is $\mathcal{O}((1-p)^k)$. In this paper, we show tha
t these rates cancel, and as a result, Dropout selectively regularizes against l
earning higher-order interactions. We prove this new perspective analytically fo
r Input Dropout and empirically for Activation Dropout. This perspective on Drop
out has several practical implications: (1) higher Dropout rates should be used
when we need stronger regularization against spurious high-order interactions, (
2) caution must be used when interpreting Dropout-based feature saliency measure
s, and (3) networks trained with Input Dropout are biased estimators, even with
infinite data. We also compare Dropout to regularization via weight decay and ea
rly stopping and find that it is difficult to obtain the same regularization aga
inst high-order interactions with these methods.
****
Revisiting Hierarchical Approach for Persistent Long-Term Video Prediction
Wonkwang Lee,Whie Jung,Han Zhang,Ting Chen,Jing Yu Koh,Thomas Huang,Hyungsuk Yoo
n,Honglak Lee,Seunghoon Hong
Learning to predict the long-term future of video frames is notoriously challeng
ing due to the inherent ambiguities in a distant future and dramatic amplificati
on of prediction error over time. Despite the recent advances in the literature,
 existing approaches are limited to moderately short-term prediction (less than
a few seconds), while extrapolating it to a longer future quickly leads to destr
uction in structure and content. In this work, we revisit the hierarchical model
s in video prediction. Our method generates future frames by first estimating a
sequence of dense semantic structures and subsequently translating the estimated
 structures to pixels by video-to-video translation model. Despite the simplicit
y, we show that modeling structures and their dynamics in categorical structure
space with stochastic sequential estimator leads to surprisingly successful long
-term prediction. We evaluate our method on two challenging video prediction sce
narios, \emph{car driving} and \emph{human dancing}, and demonstrate that it can
 generate complicated scene structures and motions over a very long time horizon
 (\ie~thousands frames), setting a new standard of video prediction with orders
of magnitude longer prediction time than existing approaches. Video results are
available at https://1konny.github.io/HVP/.
****
Self-Reflective Variational Autoencoder
Ifigeneia Apostolopoulou,Elan Rosenfeld,Artur Dubrawski
The Variational Autoencoder (VAE) is a powerful framework for learning probabili
stic latent variable generative models. However, typical assumptions on the appr
oximate posterior distributions can substantially restrict its capacity for infe
rence and generative modeling. Variational inference based on neural autoregress
ive models respects the conditional dependencies of the exact posterior, but thi
s flexibility comes at a cost: the resulting models are expensive to train in hi
gh-dimensional regimes and can be slow to produce samples. In this work, we intr
oduce an orthogonal solution, which we call self-reflective inference. By redesi
gning the hierarchical structure of existing VAE architectures, self-reflection
ensures that the stochastic flow preserves the factorization of the exact poster
ior, sequentially updating the latent codes in a manner consistent with the gene
rative model. We empirically demonstrate the advantages of matching the variatio
nal posterior to the exact posterior---on binarized MNIST self-reflective infere

nce achieves state-of-the-art performance without resorting to complex, computat
ionally expensive components such as autoregressive layers. Moreover, we design
a variational normalizing flow that employs the proposed architecture, yielding
predictive benefits compared to its purely generative counterpart. Our proposed
modification is quite general and it complements the existing literature; self-r
eflective inference can naturally leverage advances in distribution estimation a
nd generative modeling to improve the capacity of each layer in the hierarchy.
****

Symmetry-Aware Actor-Critic for 3D Molecular Design
Gregor N. C. Simm,Robert Pinsler,Gábor Csányi,José Miguel Hernández-Lobato
Automating molecular design using deep reinforcement learning (RL) has the poten
tial to greatly accelerate the search for novel materials. Despite recent progre
ss on leveraging graph representations to design molecules, such methods are fun
damentally limited by the lack of three-dimensional (3D) information. In light o
f this, we propose a novel actor-critic architecture for 3D molecular design tha
t can generate molecular structures unattainable with previous approaches. This
is achieved by exploiting the symmetries of the design process through a rotatio
nally covariant state-action representation based on a spherical harmonics serie
s expansion. We demonstrate the benefits of our approach on several 3D molecular
 design tasks, where we find that building in such symmetries significantly impr
oves generalization and the quality of generated molecules.
****

Generalizing Tree Models for Improving Prediction Accuracy
Jaemin Yoo,Lee Sael
Can we generalize and improve the representation power of tree models? Tree mode
ls are often favored over deep neural networks due to their interpretable struct
ures in problems where the interpretability is required, such as in the classifi
cation of feature-based data where each feature is meaningful. However, most tre
e models have low accuracies and easily overfit to training data. In this work,
we propose Decision Transformer Network (DTN), our highly accurate and interpret
able tree model based on our generalized framework of tree models, decision tran
sformers. Decision transformers allow us to describe tree models in the context
of deep learning. Our DTN is proposed based on improving the generalizable compo
nents of the decision transformer, which increases the representation power of t
ree models while preserving the inherent interpretability of the tree structure.
 Our extensive experiments on 121 feature-based datasets show that DTN outperfor
ms the state-of-the-art tree models and even deep neural networks.
****

Learnable Embedding sizes for Recommender Systems
Siyi Liu,Chen Gao,Yihong Chen,Depeng Jin,Yong Li
The embedding-based representation learning is commonly used in deep learning re
commendation models to map the raw sparse features to dense vectors. The traditi
onal embedding manner that assigns a uniform size to all features has two issues
. First, the numerous features inevitably lead to a gigantic embedding table tha
t causes a high memory usage cost. Second, it is likely to cause the over-fittin
g problem for those features that do not require too large representation capaci
ty. Existing works that try to address the problem always cause a significant dr
op in recommendation performance or suffers from the limitation of unaffordable
training time cost. In this paper, we proposed a novel approach, named PEP (shor
t for Plug-in Embedding Pruning), to reduce the size of the embedding table whil
e avoiding the drop of recommendation accuracy. PEP prunes embedding parameter w
here the pruning threshold(s) can be adaptively learned from data. Therefore we
can automatically obtain a mixed-dimension embedding-scheme by pruning redundant
 parameters for each feature. PEP is a general framework that can plug in variou
s base recommendation models. Extensive experiments demonstrate it can efficient
ly cut down embedding parameters and boost the base model's performance. Specifi
cally, it achieves strong recommendation performance while reducing 97-99% param
eters. As for the computation cost, PEP only brings an additional 20-30% time co
st compare with base models.
****

Goal-Driven Imitation Learning from Observation by Inferring Goal Proximity
Andrew Szot,Youngwoon Lee,Shao-Hua Sun,Joseph J Lim
Humans can effectively learn to estimate how close they are to completing a desi red task simply by watching others fulfill the task. To solve the task, they can then take actions towards states with higher estimated proximity to the goal. F rom this intuition, we propose a simple yet effective method for imitation learn ing that learns a goal proximity function from expert demonstrations and online agent experience, and then uses the learned proximity to provide a dense reward signal for training a policy to solve the task. By predicting task progress as t he temporal distance to the goal, the goal proximity function improves generaliz ation to unseen states over methods that aim to directly imitate expert behavior s. We demonstrate that our proposed method efficiently learns a set of goal-driv en tasks from state-only demonstrations in navigation, robotic arm manipulation, and locomotion tasks.
****
Bypassing the Random Input Mixing in Mixup
Hongyu Guo
Mixup and its variants have promoted a surge of interest due to their capability of boosting the accuracy of deep models. For a random sample pair, such approa ches generate a set of synthetic samples through interpolating both the inputs and their corresponding one-hot labels. Current methods either interpolate rando m features from an input pair or learn to mix salient features from the pair. Nevertheless, the former methods can create misleading synthetic samples or r emove important features from the given inputs, and the latter strategies incur significant computation cost for selecting descriptive input regions. In this pa per, we show that the effort needed for the input mixing can be bypassed. For a given sample pair, averaging the features from the two inputs and then assigning it with a set of soft labels can effectively regularize the training. We empi rically show that the proposed approach performs on par with state-of-the-art s trategies in terms of predictive accuracy.
****
Data-aware Low-Rank Compression for Large NLP Models
Patrick CHen,Hsiang-Fu Yu,Inderjit S Dhillon,Cho-Jui Hsieh
The representations learned by large-scale NLP models such as BERT have been wid ely used in various tasks. However, the increasing model size of the pre-trained models also brings the efficiency challenges, including the inference speed and the model size when deploying the model on devices. Specifically, most operatio ns in BERT consist of matrix multiplications. These matrices are not low-rank an d thus canonical matrix decomposition could not find an efficient approximation. In this paper, we observe that the learned representation of each layer lies in a low-dimensional space. Based on this observation, we propose DRONE (data-awar e low-rank compression), a provably optimal low-rank decomposition of weight mat rices, which has a simple closed form solution that can be efficiently computed. DRONE is generic, could be applied to both fully-connected and self-attention l ayers, and does not require any fine-tuning or distillation steps. Experimental results show that DRONE could improve both model size and inference speed with limited loss of accuracy. Specifically, DRONE alone achieves 1.92x faster on MRP C task with only 1.5%loss of accuracy, and when combined with distillation, DRON E achieves over 12.3x faster on various natural language inference tasks.
****
Generative Learning With Euler Particle Transport
Yuan Gao,Jian Huang,Yuling Jiao,Jin Liu
We propose an Euler particle transport (EPT) approach for generative learning. The proposed approach is motivated by the problem of finding the optimal transp ort map from a reference distribution to a target distribution characterized by the Monge-Ampere equation. Interpreting the infinitesimal linearization of the M onge-Ampere equation from the perspective of gradient flows in measure spaces le ads to a stochastic McKean-Vlasov equation. We use the forward Euler method to s olve this equation. The resulting forward Euler map pushes forward a reference d istribution to the target. This map is the composition of a sequence of simple r

esidual maps, which are computationally stable and easy to train. The key task i
n training is the estimation of the density ratios or differences that determine
 the residual maps. We estimate the density ratios (differences) based on the Br
egman divergence with a gradient penalty using deep density-ratio (difference) f
itting. We show that the proposed density-ratio (difference) estimators do not s
uffer from the ``curse of dimensionality" if data is supported on a lower-dimens
ional manifold. Numerical experiments with multi-mode synthetic datasets and com
parisons with the existing methods on real benchmark datasets support our theore
tical results and demonstrate the effectiveness of the proposed method.
****

Learning Flexible Classifiers with Shot-CONditional Episodic (SCONE) Training
Eleni Triantafillou,Vincent Dumoulin,Hugo Larochelle,Richard Zemel
Early few-shot classification work advocates for episodic training, i.e. trainin
g over learning episodes each posing a few-shot classification task. However, th
e role of this training regime remains poorly understood, and its usefulness is
still debated. Standard classification training methods (``pre-training'') follo
wed by episodic fine-tuning have recently achieved strong results. This work aim
s to understand the role of this episodic fine-tuning phase through an explorati
on of the effect of the ``shot'' setting (number of examples per class) that is
used during fine-tuning. We discover that fine-tuning on episodes of a particula
r shot can specialize the pre-trained model to solving episodes of that shot at
the expense of performance on other shots, in agreement with a trade-off recentl
y observed in the context of end-to-end episodic training. To amend this, we pro
pose a shot-conditional form of episodic fine-tuning, inspired from recent work
that trains a single model on a distribution of losses. Our investigation shows
that this improves overall performance, without suffering disproportionately on
any shot. We also examine the usefulness of this approach on the large-scale Met
a-Dataset benchmark where test episodes exhibit varying shots and imbalanced cla
sses. We find that our flexible model improves performance in that challenging e
nvironment.
****

Breaking the Expressive Bottlenecks of Graph Neural Networks
Mingqi Yang,Yanming Shen,Heng Qi,Baocai Yin
Recently, the Weisfeiler-Lehman (WL) graph isomorphism test was used to measure
the expressiveness of graph neural networks (GNNs), showing that the neighborhoo
d aggregation GNNs were at most as powerful as 1-WL test in distinguishing graph
 structures. There were also improvements proposed in analogy to k-WL test ($k
>1$). However, the aggregators in these GNNs are far from injective as required
by the WL test, and suffer from weak distinguishing strength, making it become e
xpressive bottlenecks. In this paper, we improve the expressiveness by exploring
 powerful aggregators. We reformulate aggregation with the corresponding aggrega
tion coefficient matrix, and then systematically analyze the requirements of the
 aggregation coefficient matrix for building more powerful aggregators and even
injective aggregators. It can also be viewed as the strategy for preserving the
rank of hidden features, and implies that basic aggregators correspond to a spec
ial case of low-rank transformations. We also show the necessity of applying non
linear units ahead of aggregation, which is different from most aggregation-base
d GNNs. Based on our theoretical analysis, we develop two GNN layers, ExpandingC
onv and CombConv. Experimental results show that our models significantly boost
performance, especially for large and densely connected graphs.
****

Iterative Empirical Game Solving via Single Policy Best Response
Max Smith,Thomas Anthony,Michael Wellman
Policy-Space Response Oracles (PSRO) is a general algorithmic framework for lear
ning policies in multiagent systems by interleaving empirical game analysis with
 deep reinforcement learning (DRL).
At each iteration, DRL is invoked to train a best response to a mixture of oppon
ent policies.
The repeated application of DRL poses an expensive computational burden as we lo
ok to apply this algorithm to more complex domains.

We introduce two variations of PSRO designed to reduce the amount of simulation required during DRL training.
Both algorithms modify how PSRO adds new policies to the empirical game, based on learned responses to a single opponent policy.
The first, Mixed-Oracles, transfers knowledge from previous iterations of DRL, requiring training only against the opponent's newest policy.
The second, Mixed-Opponents, constructs a pure-strategy opponent by mixing existing strategy's action-value estimates, instead of their policies.
Learning against a single policy mitigates conflicting experiences on behalf of a learner facing an unobserved distribution of opponents.
We empirically demonstrate that these algorithms substantially reduce the amount of simulation during training required by PSRO, while producing equivalent or better solutions to the game.
****

Differential-Critic GAN: Generating What You Want by a Cue of Preferences
Yinghua Yao,Yuangang Pan,Ivor Tsang,Xin Yao
This paper proposes Differential-Critic Generative Adversarial Network (DiCGAN) to learn the distribution of user-desired data when only partial instead of the entire dataset possesses the desired properties. Existing approaches select the desired samples first and train regular GANs on the selected samples to derive the user-desired data distribution. DiCGAN introduces a differential critic that can learn the preference direction from the pairwise preferences over the entire dataset. The resultant critic would guide the generation of the desired data instead of the whole data. Specifically, apart from the Wasserstein GAN loss, a ranking loss of the pairwise preferences is defined over the critic. It endows the difference of critic values between each pair of samples with the pairwise preference relation. The higher critic value indicates that the sample is preferred by the user. Thus training the generative model for higher critic values would encourage generating the user-preferred samples. Extensive experiments show that our DiCGAN can learn the user-desired data distributions.
****

Learning Neural Generative Dynamics for Molecular Conformation Generation
Minkai Xu,Shitong Luo,Yoshua Bengio,Jian Peng,Jian Tang
We study how to generate molecule conformations (i.e., 3D structures) from a molecular graph. Traditional methods, such as molecular dynamics, sample conformations via computationally expensive simulations. Recently, machine learning methods have shown great potential by training on a large collection of conformation data. Challenges arise from the limited model capacity for capturing complex distributions of conformations and the difficulty in modeling long-range dependencies between atoms. Inspired by the recent progress in deep generative models, in this paper, we propose a novel probabilistic framework to generate valid and diverse conformations given a molecular graph. We propose a method combining the advantages of both flow-based and energy-based models, enjoying: (1) a high model capacity to estimate the multimodal conformation distribution; (2) explicitly capturing the complex long-range dependencies between atoms in the observation space. Extensive experiments demonstrate the superior performance of the proposed method on several benchmarks, including conformation generation and distance modeling tasks, with a significant improvement over existing generative models for molecular conformation sampling.
****

Learning the Pareto Front with Hypernetworks
Aviv Navon,Aviv Shamsian,Ethan Fetaya,Gal Chechik
Multi-objective optimization (MOO) problems are prevalent in machine learning. These problems have a set of optimal solutions, called the Pareto front, where each point on the front represents a different trade-off between possibly conflicting objectives. Recent MOO methods can target a specific desired ray in loss space however, most approaches still face two grave limitations: (i) A separate model has to be trained for each point on the front; and (ii) The exact trade-off must be known before the optimization process. Here, we tackle the problem of learning the entire Pareto front, with the capability of selecting a desired operat

ing point on the front after training. We call this new setup Pareto-Front Learning (PFL).

We describe an approach to PFL implemented using HyperNetworks, which we term Pareto HyperNetworks (PHNs). PHN learns the entire Pareto front simultaneously using a single hypernetwork, which receives as input a desired preference vector and returns a Pareto-optimal model whose loss vector is in the desired ray. The unified model is runtime efficient compared to training multiple models and generalizes to new operating points not used during training. We evaluate our method on a wide set of problems, from multi-task regression and classification to fairness. PHNs learn the entire Pareto front at roughly the same time as learning a single point on the front and at the same time reach a better solution set. PFL opens the door to new applications where models are selected based on preferences that are only available at run time.
****
Characterizing Lookahead Dynamics of Smooth Games
Junsoo Ha,Gunhee Kim
As multi-agent systems proliferate in machine learning research, games have attracted much attention as a framework to understand optimization of multiple interacting objectives. However, a key challenge in game optimization is that, in general, there is no guarantee for usual gradient-based methods to converge to a local solution of the game. The latest work by Chavdarova et al. (2020) report that Lookahead optimizer (Zhang et al. (2019)) significantly improves the performance of Generative Adversarial Networks (GANs) and reduces the rotational force of bilinear games. While promising, their observations were purely empirical, and Lookahead optimization of smooth games still lacks theoretical understanding. In this paper, we fill this gap by theoretically characterizing Lookahead dynamics of smooth games. We provide an intuitive geometric explanation on how and when Lookahead can improve game dynamics in terms of stability and convergence. Furthermore, we present sufficient conditions under which Lookahead optimization of bilinear games provably stabilizes or accelerates convergence to a Nash equilibrium of the game. Finally, we show that Lookahead optimizer preserves locally asymptotically stable equilibria of base dynamics and can either stabilize or accelerate the local convergence to a given equilibrium with proper assumptions. We verify each of our theoretical predictions by conducting numerical experiments on two-player zero-sum (non-linear) games.
****
Recurrent Neural Network Architecture based on Dynamic Systems Theory for Data Driven Modelling of Complex Physical Systems
Deniz Neufeld
While dynamic systems can be modelled as sequence-to-sequence tasks by deep learning using different network architectures like DNN, CNN, RNNs or neural ODEs, the resulting models often provide poor understanding of the underlying system properties. We propose a new recurrent network architecture, the Dynamic Recurrent Network, where the computation function is based on the discrete difference equations of basic linear system transfer functions known from dynamic system identification. This results in a more explainable model, since the learnt weights can provide insight on a system's time dependent behaviour. It also introduces the sequences' sampling rate as an additional model parameter, which can be leveraged, for example, for time series data augmentation and model robustness checks. The network is trained using traditional gradient descent optimization and can be used in combination with other state of the art neural network layers. We show that our new layer type yields results comparable to or better than other recurrent layer types on several system identification tasks.
****
A new accelerated gradient method inspired by continuous-time perspective
Yasong Feng,Weiguo Gao
Nesterov's accelerated method are widely used in problems with machine learning background including deep learning. To give more insight about the acceleration phenomenon, an ordinary differential equation was obtained from Nesterov's accel

erated method by taking step sizes approaching zero, and the relationship betwee n Nesterov's method and the differential equation is still of research interest. In this work, we give the precise order of the iterations of Nesterov's acceler ated method converging to the solution of derived differential equation as step sizes go to zero. We then present a new accelerated method with higher order. Th e new method is more stable than ordinary method for large step size and converg es faster. We further apply the new method to matrix completion problem and show its better performance through numerical experiments.
****

Intervention Generative Adversarial Nets
Jiadong Liang,Liangyu Zhang,Cheng Zhang,Zhihua Zhang
In this paper we propose a novel approach for stabilizing the training process o f Generative Adversarial Networks as well as alleviating the mode collapse probl em. The main idea is to incorporate a regularization term that we call intervent ion into the objective. We refer to the resulting generative model as Interventi on Generative Adversarial Networks (IVGAN). By perturbing the latent representat ions of real images obtained from an auxiliary encoder network with Gaussian inv ariant interventions and penalizing the dissimilarity of the distributions of th e resulting generated images, the intervention term provides more informative gr adient for the generator, significantly improving training stability and encoura ging modecovering behaviour. We demonstrate the performance of our approach via solid theoretical analysis and thorough evaluation on standard real-world datase ts as well as the stacked MNIST dataset.

****

Ask Your Humans: Using Human Instructions to Improve Generalization in Reinforce ment Learning
Valerie Chen,Abhinav Gupta,Kenneth Marino
Complex, multi-task problems have proven to be difficult to solve efficiently in a sparse-reward reinforcement learning setting. In order to be sample efficient , multi-task learning requires reuse and sharing of low-level policies. To facil itate the automatic decomposition of hierarchical tasks, we propose the use of s tep-by-step human demonstrations in the form of natural language instructions an d action trajectories. We introduce a dataset of such demonstrations in a crafti ng-based grid world. Our model consists of a high-level language generator and l ow-level policy, conditioned on language. We find that human demonstrations help solve the most complex tasks. We also find that incorporating natural language allows the model to generalize to unseen tasks in a zero-shot setting and to lea rn quickly from a few demonstrations. Generalization is not only reflected in th e actions of the agent, but also in the generated natural language instructions in unseen tasks. Our approach also gives our trained agent interpretable behavio rs because it is able to generate a sequence of high-level descriptions of its a ctions.
****

Density estimation on low-dimensional manifolds: an inflation-deflation approach
Christian Horvat
Normalizing Flows (NFs) are universal density estimators based on Neuronal Netwo rks. However, this universality is limited: the density's support needs to be di ffeomorphic to a Euclidean space. In this paper, we propose a novel method to ov ercome this limitation without sacrificing the universality. The proposed method inflates the data manifold by adding noise in the normal space, trains an NF on this inflated manifold and, finally, deflates the learned density. Our main res ult provides sufficient conditions on the manifold and the specific choice of no ise under which the corresponding estimator is exact. Our method has the same co mputational complexity as NFs, and does not require to compute an inverse flow. We also show that, if the embedding dimension is much larger than the manifold d imension, noise in the normal space can be well approximated by some Gaussian no ise. This allows using our method for approximating arbitrary densities on non-f lat manifolds provided that the manifold dimension is known.
****

Correcting Momentum in Temporal Difference Learning
Emmanuel Bengio,Joelle Pineau,Doina Precup
A common optimization tool used in deep reinforcement learning is momentum, which consists in accumulating and discounting past gradients, reapplying them at each iteration. We argue that, unlike in supervised learning, momentum in Temporal Difference (TD) learning accumulates gradients that become doubly stale: not only does the gradient of the loss change due to parameter updates, the loss itself changes due to bootstrapping. We first show that this phenomenon exists, and then propose a first-order correction term to momentum. We show that this correction term improves sample efficiency in policy evaluation by correcting target value drift. An important insight of this work is that deep RL methods are not always best served by directly importing techniques from the supervised setting.
****
TRIP: Refining Image-to-Image Translation via Rival Preferences
Yinghua Yao,Yuangang Pan,Ivor Tsang,Xin Yao
We propose a new model to refine image-to-image translation via an adversarial ranking process. In particular, we simultaneously train two modules: a generator that translates an input image to the desired image with smooth subtle changes with respect to some specific attributes; and a ranker that ranks rival preferences consisting of the input image and the desired image. Rival preferences refer to the adversarial ranking process: (1) the ranker thinks no difference between the desired image and the input image in terms of the desired attributes; (2) the generator fools the ranker to believe that the desired image changes the attributes over the input image as desired. Real image preferences are introduced to guide the ranker to rank image pairs regarding the interested attributes only. With an effective ranker, the generator would “win” the adversarial game by producing high-quality images that present desired changes over the attributes compared to the input image. The experiments demonstrate that our TRIP can generate high-fidelity images which exhibit smooth changes with the strength of the attributes.
****
Efficient Neural Machine Translation with Prior Word Alignment
Jeonghyeok Park,hai zhao
Prior word alignment has been shown indeed helpful for a better translation if such prior is good enough and can be acquired in a convenient way at the same time. Traditionally, word alignment can be learned through statistical machine translation (SMT) models. In this paper, we propose a novel method that infuses prior word alignment information into neural machine translation (NMT) to provide hints or guidelines for the target sentence at running time. To this end, previous works of similar approaches should build dictionaries for specific domains, or constraint the decoding process, or both. While being effective to some extent, these methods may greatly affect decoding speed and hurt translation flexibility and efficiency. Instead, this paper introduces an enhancement learning model, which can learn how to directly replace specific source words with their target counterparts according to prior alignment information. The proposed model is then inserted into a neural MT model and augments MT input with the additional target information from the learning model in an effective and more efficient way. Our novel method achieves BLEU improvements (up to 1.1) over a strong baseline model on English-Korean and English-Romanian translation tasks.
****
Globally Injective ReLU networks
Michael Puthawala,Konik Kothari,Matti Lassas,Ivan Dokmani■,Maarten V. de Hoop
Injectivity plays an important role in generative models where it enables inference; in inverse problems and compressed sensing with generative priors it is a precursor to well posedness. We establish sharp characterizations of injectivity of fully-connected and convolutional ReLU layers and networks. First, through a layerwise analysis, we show that an expansivity factor of two is necessary and sufficient for injectivity by constructing appropriate weight matrices. We show that global injectivity with iid Gaussian matrices, a commonly used tractable model, requires larger expansivity between 3.4 and 10.5. We also characterize the s

tability of inverting an injective network via worst-case Lipschitz constants of the inverse. We then use arguments from differential topology to study injectivity of deep networks and prove that any Lipschitz map can be approximated by an injective ReLU network. Finally, using an argument based on random projections, we show that an end-to-end---rather than layerwise---doubling of the dimension suffices for injectivity. Our results establish a theoretical basis for the study of nonlinear inverse and inference problems using neural networks.
****

Taming GANs with Lookahead-Minmax
Tatjana Chavdarova,Matteo Pagliardini,Sebastian U Stich,François Fleuret,Martin Jaggi
Generative Adversarial Networks are notoriously challenging to train. The underlying minmax optimization is highly susceptible to the variance of the stochastic gradient and the rotational component of the associated game vector field. To tackle these challenges, we propose the Lookahead algorithm for minmax optimization, originally developed for single objective minimization only. The backtracking step of our Lookahead–minmax naturally handles the rotational game dynamics, a property which was identified to be key for enabling gradient ascent descent methods to converge on challenging examples often analyzed in the literature. Moreover, it implicitly handles high variance without using large mini-batches, known to be essential for reaching state of the art performance. Experimental results on MNIST, SVHN, CIFAR-10, and ImageNet demonstrate a clear advantage of combining Lookahead–minmax with Adam or extragradient, in terms of performance and improved stability, for negligible memory and computational cost. Using 30-fold fewer parameters and 16-fold smaller minibatches we outperform the reported performance of the class-dependent BigGAN on CIFAR-10 by obtaining FID of 12.19 without using the class labels, bringing state-of-the-art GAN training within reach of common computational resources.
****

Uniform Priors for Data-Efficient Transfer
Samarth Sinha,Karsten Roth,Anirudh Goyal,Marzyeh Ghassemi,Hugo Larochelle,Animesh Garg
Deep Neural Networks have shown great promise on a variety of downstream applications; but their ability to adapt and generalize to new data and tasks remains a challenging problem. However, the ability to perform few or zero-shot adaptation to novel tasks is important for the scalability and deployment of machine learning models. It is therefore crucial to understand what makes for good, transferable features in deep networks that best allow for such adaptation. In this paper, we shed light on this by showing that features that are most transferable have high uniformity in the embedding space and propose a uniformity regularization scheme that encourages better transfer and feature reuse. We evaluate the regularization on its ability to facilitate adaptation to unseen tasks and data, for which we conduct a thorough experimental study covering four relevant, and distinct domains: few-shot Meta-Learning, Deep Metric Learning, Zero-Shot Domain Adaptation, as well as Out-of-Distribution classification. Across all experiments, we show that uniformity regularization consistently offers benefits over baseline methods and is able to achieve state-of-the-art performance in Deep Metric Learning and Meta-Learning.
****

Why Does Decentralized Training Outperform Synchronous Training In The Large Batch Setting?
Wei Zhang,Mingrui Liu,Yu Feng,Brian Kingsbury,Yuhai Tu
Distributed Deep Learning (DDL) is essential for large-scale Deep Learning (DL) training. Using a sufficiently large batch size is critical to achieving DDL runtime speedup. In a large batch setting, the learning rate must be increased to compensate for the reduced number of parameter updates. However, a large batch size may converge to sharp minima with poor generalization, and a large learning rate may harm convergence. Synchronous Stochastic Gradient Descent (SSGD) is the de facto DDL optimization method. Recently, Decentralized Parallel SGD (DPSGD) has been proven to achieve a similar convergence rate as SGD and to guarantee lin

ear speedup for non-convex optimization problems. While there was anecdotal evid ence that DPSGD outperforms SSGD in the large-batch setting, no systematic study has been conducted to explain why this is the case. Based on a detailed analysi s of the DPSGD learning dynamics, we find that DPSGD introduces additional lands cape-dependent noise, which has two benefits in the large-batch setting: 1) it a utomatically adjusts the learning rate to improve convergence; 2) it enhances we ight space search by escaping local traps (e.g., saddle points) to find flat min ima with better generalization. We conduct extensive studies over 12 state-of-th e-art DL models/tasks and demonstrate that DPSGD consistently outperforms SSGD i n the large batch setting;
and DPSGD converges in cases where SSGD diverges for large learning rates. Our findings are consistent across different application domains, Computer Vision a nd Automatic Speech Recognition, and different neural network models, Convolutio nal Neural Networks and Long Short-Term Memory Recurrent Neural Networks.
****

Uncertainty in Gradient Boosting via Ensembles
Andrey Malinin,Liudmila Prokhorenkova,Aleksei Ustimenko
For many practical, high-risk applications, it is essential to quantify uncertai nty in a model's predictions to avoid costly mistakes. While predictive uncertai nty is widely studied for neural networks, the topic seems to be under-explored for models based on gradient boosting. However, gradient boosting often achieves state-of-the-art results on tabular data. This work examines a probabilistic en semble-based framework for deriving uncertainty estimates in the predictions of gradient boosting classification and regression models. We conducted experiments on a range of synthetic and real datasets and investigated the applicability of ensemble approaches to gradient boosting models that are themselves ensembles o f decision trees. Our analysis shows that ensembles of gradient boosting models successfully detect anomalous inputs while having limited ability to improve the predicted total uncertainty. Importantly, we also propose a concept of a virtua l ensemble to get the benefits of an ensemble via only one gradient boosting mod el, which significantly reduces complexity.
****

Experience Replay with Likelihood-free Importance Weights
Samarth Sinha,Jiaming Song,Animesh Garg,Stefano Ermon
The use of past experiences to accelerate temporal difference (TD) learning of v alue functions, or experience replay, is a key component in deep reinforcement l earning. In this work, we propose to reweight experiences based on their likelih ood under the stationary distribution of the current policy, and justify this wi th a contraction argument over the Bellman evaluation operator. The resulting TD objective encourages small approximation errors on the value function over freq uently encountered states. To balance bias and variance in practice, we use a li kelihood-free density ratio estimator between on-policy and off-policy experienc es, and use the ratios as the prioritization weights. We apply the proposed appr oach empirically on three competitive methods, Soft Actor Critic (SAC), Twin Del ayed Deep Deterministic policy gradient (TD3) and Data-regularized Q (DrQ), over 11 tasks from OpenAI gym and DeepMind control suite. We achieve superior sample complexity on 35 out of 45 method-task combinations compared to the best baseli ne and similar sample complexity on the remaining 10.
****

Cluster & Tune: Enhance BERT Performance in Low Resource Text Classification
Eyal Shnarch,Ariel Gera,Alon Halfon,Lena Dankin,Leshem Choshen,Ranit Aharonov,No am Slonim
In data-constrained cases, the common practice of fine-tuning BERT for a target text classification task is prone to producing poor performance. In such low res ources scenarios, we suggest performing an unsupervised classification task prio r to fine-tuning on the target task.
Specifically, as such an intermediate task, we perform unsupervised clustering, training BERT on predicting the cluster labels. We test this hypothesis on vario us data sets, and show that this additional classification step can reduce the d emand for labeled examples.

We further discuss under which conditions this task is helpful and why.
****
Adversarial score matching and improved sampling for image generation
Alexia Jolicoeur-Martineau,Rémi Piché-Taillefer,Ioannis Mitliagkas,Remi Tachet des Combes
Denoising Score Matching with Annealed Langevin Sampling (DSM-ALS) has recently found success in generative modeling. The approach works by first training a neural network to estimate the score of a distribution, and then using Langevin dynamics to sample from the data distribution assumed by the score network. Despite the convincing visual quality of samples, this method appears to perform worse than Generative Adversarial Networks (GANs) under the Fréchet Inception Distance, a standard metric for generative models. We show that this apparent gap vanishes when denoising the final Langevin samples using the score network.
In addition, we propose two improvements to DSM-ALS: 1) Consistent Annealed Sampling as a more stable alternative to Annealed Langevin Sampling, and 2) a hybrid training formulation, composed of both Denoising Score Matching and adversarial objectives. By combining these two techniques and exploring different network architectures, we elevate score matching methods and obtain results competitive with state-of-the-art image generation on CIFAR-10.
****
Straight to the Gradient: Learning to Use Novel Tokens for Neural Text Generation
Xiang Lin,SIMENG HAN,Shafiq Joty
Advanced large-scale neural language models have led to significant success in many natural language generation tasks. However, the most commonly used training objective, Maximum Likelihood Estimation (MLE), has been shown to be problematic, where the trained model prefers using dull and repetitive phrases. In this work, we introduce ScaleGrad, a modification straight to the gradient of the loss function, to remedy the degeneration issues of the standard MLE objective. By directly maneuvering the gradient information, ScaleGrad makes the model learn to use novel tokens during training. Empirical results show the effectiveness of our method not only in open-ended generation, but also in directed generation. With the simplicity in architecture, our method can serve as a general training objective that is applicable to most of the neural text generation tasks.
****
FastSpeech 2: Fast and High-Quality End-to-End Text to Speech
Yi Ren,Chenxu Hu,Xu Tan,Tao Qin,Sheng Zhao,Zhou Zhao,Tie-Yan Liu
Non-autoregressive text to speech (TTS) models such as FastSpeech can synthesize speech significantly faster than previous autoregressive models with comparable quality. The training of FastSpeech model relies on an autoregressive teacher model for duration prediction (to provide more information as input) and knowledge distillation (to simplify the data distribution in output), which can ease the one-to-many mapping problem (i.e., multiple speech variations correspond to the same text) in TTS. However, FastSpeech has several disadvantages: 1) the teacher-student distillation pipeline is complicated and time-consuming, 2) the duration extracted from the teacher model is not accurate enough, and the target mel-spectrograms distilled from teacher model suffer from information loss due to data simplification, both of which limit the voice quality. In this paper, we propose FastSpeech 2, which addresses the issues in FastSpeech and better solves the one-to-many mapping problem in TTS by 1) directly training the model with ground-truth target instead of the simplified output from teacher, and 2) introducing more variation information of speech (e.g., pitch, energy and more accurate duration) as conditional inputs. Specifically, we extract duration, pitch and energy from speech waveform and directly take them as conditional inputs in training and use predicted values in inference. We further design FastSpeech 2s, which is the first attempt to directly generate speech waveform from text in parallel, enjoying the benefit of fully end-to-end inference. Experimental results show that 1) FastSpeech 2 achieves a 3x training speed-up over FastSpeech, and FastSpeech 2s enjoys even faster inference speed; 2) FastSpeech 2 and 2s outperform FastSpeech in voice quality, and FastSpeech 2 can even surpass autoregressive models.

Audio samples are available at https://speechresearch.github.io/fastspeech2/.
****
Interpretable Models for Granger Causality Using Self-explaining Neural Networks
Ri■ards Marcinkevi■s,Julia E Vogt
Exploratory analysis of time series data can yield a better understanding of complex dynamical systems. Granger causality is a practical framework for analysing interactions in sequential data, applied in a wide range of domains. In this paper, we propose a novel framework for inferring multivariate Granger causality under nonlinear dynamics based on an extension of self-explaining neural networks. This framework is more interpretable than other neural-network-based techniques for inferring Granger causality, since in addition to relational inference, it also allows detecting signs of Granger-causal effects and inspecting their variability over time. In comprehensive experiments on simulated data, we show that our framework performs on par with several powerful baseline methods at inferring Granger causality and that it achieves better performance at inferring interaction signs. The results suggest that our framework is a viable and more interpretable alternative to sparse-input neural networks for inferring Granger causality.
****
Reducing Implicit Bias in Latent Domain Learning
Lucas Deecke,Timothy Hospedales,Hakan Bilen
A fundamental shortcoming of deep neural networks is their specialization to a single task and domain. While recent techniques in multi-domain learning enable the learning of more domain-agnostic features, their success relies firmly on the presence of domain labels, typically requiring manual annotation and careful curation of datasets. Here we focus on latent domain learning, a highly realistic, yet less explored scenario: learning from data from different domains, without access to domain annotations. This is a particularly challenging problem, since standard models exhibit an implicit bias toward learning only the large domains in data, while disregarding smaller ones. To address this issue, we propose dynamic residual adapters that adaptively account for latent domains, and weighted domain transfer – a novel augmentation strategy designed specifically for this setting. Our techniques are evaluated on image classification tasks containing multiple unannotated domains, and we demonstrate they enhance performance, in particular, on the smallest of these.
****
Learning Deep Features in Instrumental Variable Regression
Liyuan Xu,Yutian Chen,Siddarth Srinivasan,Nando de Freitas,Arnaud Doucet,Arthur Gretton
Instrumental variable (IV) regression is a standard strategy for learning causal relationships between confounded treatment and outcome variables from observational data by using an instrumental variable, which affects the outcome only through the treatment. In classical IV regression, learning proceeds in two stages: stage 1 performs linear regression from the instrument to the treatment; and stage 2 performs linear regression from the treatment to the outcome, conditioned on the instrument. We propose a novel method, deep feature instrumental variable regression (DFIV), to address the case where relations between instruments, treatments, and outcomes may be nonlinear. In this case, deep neural nets are trained to define informative nonlinear features on the instruments and treatments. We propose an alternating training regime for these features to ensure good end-to-end performance when composing stages 1 and 2, thus obtaining highly flexible feature maps in a computationally efficient manner.
DFIV outperforms recent state-of-the-art methods on challenging IV benchmarks, including settings involving high dimensional image data. DFIV also exhibits competitive performance in off-policy policy evaluation for reinforcement learning, which can be understood as an IV regression task.
****
Causal Screening to Interpret Graph Neural Networks
Xiang Wang,Yingxin Wu,An Zhang,Xiangnan He,Tat-seng Chua
With the growing success of graph neural networks (GNNs), the explainability of

GNN is attracting considerable attention. However, current works on feature attribution, which frame explanation generation as attributing a prediction to the graph features, mostly focus on the statistical interpretability. They may struggle to distinguish causal and noncausal effects of features, and quantify redundancy among features, thus resulting in unsatisfactory explanations. In this work, we focus on the causal interpretability in GNNs and propose a method, Causal Screening, from the perspective of cause-effect. It incrementally selects a graph feature (i.e., edge) with large causal attribution, which is formulated as the individual causal effect on the model outcome. As a model-agnostic tool, Causal Screening can be used to generate faithful and concise explanations for any GNN model. Further, by conducting extensive experiments on three graph classi■cation datasets, we observe that Causal Screening achieves signi■cant improvements over state-of-the-art approaches w.r.t. two quantitative metrics: predictive accuracy, contrastivity, and safely passes sanity checks.

****

Statistical inference for individual fairness

Subha Maity,Songkai Xue,Mikhail Yurochkin,Yuekai Sun

As we rely on machine learning (ML) models to make more consequential decisions, the issue of ML models perpetuating unwanted social biases has come to the fore of the public's and the research community's attention. In this paper, we focus on the problem of detecting violations of individual fairness in ML models. We formalize the problem as measuring the susceptibility of ML models against a form of adversarial attack and develop a suite of inference tools for the adversarial loss. The tools allow practitioners to assess the individual fairness of ML models in a statistically-principled way: form confidence intervals for the adversarial loss and test hypotheses of model fairness with (asymptotic) non-coverage/Type I error rate control. We demonstrate the utility of our tools in a real-world case study.

****

Self-Supervised Policy Adaptation during Deployment

Nicklas Hansen,Rishabh Jangir,Yu Sun,Guillem Alenyà,Pieter Abbeel,Alexei A Efros,Lerrel Pinto,Xiaolong Wang

In most real world scenarios, a policy trained by reinforcement learning in one environment needs to be deployed in another, potentially quite different environment. However, generalization across different environments is known to be hard. A natural solution would be to keep training after deployment in the new environment, but this cannot be done if the new environment offers no reward signal. Our work explores the use of self-supervision to allow the policy to continue training after deployment without using any rewards. While previous methods explicitly anticipate changes in the new environment, we assume no prior knowledge of those changes yet still obtain significant improvements. Empirical evaluations are performed on diverse simulation environments from DeepMind Control suite and ViZDoom, as well as real robotic manipulation tasks in continuously changing environments, taking observations from an uncalibrated camera. Our method improves generalization in 31 out of 36 environments across various tasks and outperforms domain randomization on a majority of environments. Webpage and implementation: https://nicklashansen.github.io/PAD/.

****

Disentangled Generative Causal Representation Learning

Xinwei Shen,Furui Liu,Hanze Dong,Qing LIAN,Zhitang Chen,Tong Zhang

This paper proposes a Disentangled gEnerative cAusal Representation (DEAR) learning method. Unlike existing disentanglement methods that enforce independence of the latent variables, we consider the general case where the underlying factors of interests can be causally correlated. We show that previous methods with independent priors fail to disentangle causally related factors. Motivated by this finding, we propose a new disentangled learning method called DEAR that enables causal controllable generation and causal representation learning. The key ingredient of this new formulation is to use a structural causal model (SCM) as the prior for a bidirectional generative model. A generator is then trained jointly with an encoder using a suitable GAN loss. Theoretical justification on the propo

sed formulation is provided, which guarantees disentangled causal representation learning under appropriate conditions. We conduct extensive experiments on both synthesized and real datasets to demonstrate the effectiveness of DEAR in causal controllable generation, and the benefits of the learned representations for downstream tasks in terms of sample efficiency and distributional robustness.
****
Randomized Ensembled Double Q-Learning: Learning Fast Without a Model
Xinyue Chen,Che Wang,Zijian Zhou,Keith W. Ross
Using a high Update-To-Data (UTD) ratio, model-based methods have recently achieved much higher sample efficiency than previous model-free methods for continuous-action DRL benchmarks. In this paper, we introduce a simple model-free algorithm, Randomized Ensembled Double Q-Learning (REDQ), and show that its performance is just as good as, if not better than, a state-of-the-art model-based algorithm for the MuJoCo benchmark. Moreover, REDQ can achieve this performance using fewer parameters than the model-based method, and with less wall-clock run time. REDQ has three carefully integrated ingredients which allow it to achieve its high performance: (i) a UTD ratio $\gg 1$; (ii) an ensemble of Q functions; (iii) in-target minimization across a random subset of Q functions from the ensemble. Through carefully designed experiments, we provide a detailed analysis of REDQ and related model-free algorithms. To our knowledge, REDQ is the first successful model-free DRL algorithm for continuous-action spaces using a UTD ratio $\gg 1$.

****
Graph Deformer Network
Wenting Zhao,Yuan Fang,Zhen Cui,Tong Zhang,Jian Yang,Wei Liu
Convolution learning on graphs draws increasing attention recently due to its potential applications to a large amount of irregular data. Most graph convolution methods leverage the plain summation/average aggregation to avoid the discrepancy of responses from isomorphic graphs. However, such an extreme collapsing way would result in a structural loss and signal entanglement of nodes, which further cause the degradation of the learning ability. In this paper, we propose a simple yet effective graph deformer network (GDN) to fulfill anisotropic convolution filtering on graphs, analogous to the standard convolution operation on images. Local neighborhood subgraphs (acting like receptive fields) with different structures are deformed into a unified virtual space, coordinated by several anchor nodes. In space deformation, we transfer components of nodes therein into anchors by learning their correlation, and build a pseudo multi-granularity plane calibrated with anchors. Anisotropic convolutional kernels can be further performed over the anchor-coordinated space to well encode local variations of receptive fields. By parameterizing anchors and stacking coarsening layers, we build a graph deformer network in an end-to-end fashion. Theoretical analysis indicates its connection to previous work and shows the promising property of isomorphism testing. Extensive experiments on widely-used datasets validate the effectiveness of the proposed GDN in node and graph classifications.
****
There is no trade-off: enforcing fairness can improve accuracy
Subha Maity,Debarghya Mukherjee,Mikhail Yurochkin,Yuekai Sun
One of the main barriers to the broader adoption of algorithmic fairness in machine learning is the trade-off between fairness and performance of ML models: many practitioners are unwilling to sacrifice the performance of their ML model for fairness. In this paper, we show that this trade-off may not be necessary. If the algorithmic biases in an ML model are due to sampling biases in the training data, then enforcing algorithmic fairness may improve the performance of the ML model on unbiased test data. We study conditions under which enforcing algorithmic fairness helps practitioners learn the Bayes decision rule for (unbiased) test data from biased training data. We also demonstrate the practical implications of our theoretical results in real-world ML tasks.
****
DROPS: Deep Retrieval of Physiological Signals via Attribute-specific Clinical Prototypes

Dani Kiyasseh,Tingting Zhu,David A. Clifton
The ongoing digitization of health records within the healthcare industry results in large-scale datasets. Manually extracting clinically-useful insight from such datasets is non-trivial. However, doing so at scale while simultaneously leveraging patient-specific attributes such as sex and age can assist with clinical-trial enrollment, medical school educational endeavours, and the evaluation of the fairness of neural networks. To facilitate the reliable extraction of clinical information, we propose to learn embeddings, known as clinical prototypes (CPs), via supervised contrastive learning. We show that CPs can be efficiently used for large-scale retrieval and clustering of physiological signals based on multiple patient attributes. We also show that CPs capture attribute-specific semantic relationships.

****

Can one hear the shape of a neural network?: Snooping the GPU via Magnetic Side Channel
Henrique Teles Maia,Chang Xiao,Dingzeyu Li,Eitan Grinspun,Changxi Zheng
We examine the magnetic flux emanating from a graphics processing unit's (GPU's) power cable, as acquired by a cheap $3 induction sensor, and find that this signal betrays the detailed topology and hyperparameters of a black-box neural network model. The attack acquires the magnetic signal for one query with unknown input values, but known input dimension and batch size. The reconstruction is possible due to the modular layer sequence in which deep neural networks are evaluated. We find that each layer component's evaluation produces an identifiable magnetic signal signature, from which layer topology, width, function type, and sequence order can be inferred using a suitably trained classifier and an optimization based on integer programming. We study the extent to which network specifications can be recovered, and consider metrics for comparing network similarity. We demonstrate the potential accuracy of this side channel attack in recovering the details for a broad range of network architectures including also random designs. We consider applications that may exploit this novel side channel exposure, such as adversarial transfer attacks. In response, we discuss countermeasures to protect against our method and other similar snooping techniques.

****

PGPS : Coupling Policy Gradient with Population-based Search
Namyong Kim,Hyunsuk Baek,Hayong Shin
Gradient-based policy search algorithms (such as PPO, SAC or TD3) in deep reinforcement learning (DRL) have shown successful results on a range of challenging control tasks. However, they often suffer from flat or deceptive gradient problems. As an alternative to policy gradient methods, population-based evolutionary approaches have been applied to DRL. While population-based search algorithms show more robust learning in a broader range of tasks, they are usually inefficient in the use of samples. Recently, reported are a few attempts (such as CEMRL) to combine gradient with a population in searching optimal policy. This kind of hybrid algorithm takes advantage of both camps. In this paper, we propose yet another hybrid algorithm, which more tightly couples policy gradient with the population-based search. More specifically, we use the Cross-Entropy Method (CEM) for population-based search and Twin Delayed Deep Deterministic Policy Gradient (TD3) for policy gradient. In the proposed algorithm called Coupling Policy Gradient with Population-based Search (PGPS), a single TD3 agent, which learns by a gradient from all experiences generated by population, leads a population by providing its critic function Q as a surrogate to select better performing next-generation population from candidates. On the other hand, if the TD3 agent falls behind the CEM population, then the TD3 agent is updated toward the elite member of the CEM population using loss function augmented with the distance between the TD3 and the CEM elite. Experiments in a MuJoCo environment show that PGPS is robust to deceptive gradient and also outperforms the state-of-the-art algorithms.

****

Lifelong Learning of Compositional Structures
Jorge A Mendez,ERIC EATON

A hallmark of human intelligence is the ability to construct self-contained chun
ks of knowledge and adequately reuse them in novel combinations for solving diff
erent yet structurally related problems. Learning such compositional structures
has been a significant challenge for artificial systems, due to the combinatoria
l nature of the underlying search problem. To date, research into compositional
learning has largely proceeded separately from work on lifelong or continual lea
rning. We integrate these two lines of work to present a general-purpose framewo
rk for lifelong learning of compositional structures that can be used for solvin
g a stream of related tasks. Our framework separates the learning process into t
wo broad stages: learning how to best combine existing components in order to as
similate a novel problem, and learning how to adapt the set of existing componen
ts to accommodate the new problem. This separation explicitly handles the trade-
off between the stability required to remember how to solve earlier tasks and th
e flexibility required to solve new tasks, as we show empirically in an extensiv
e evaluation.
****

Dynamically Stable Infinite-Width Limits of Neural Classifiers
Eugene Golikov
Recent research has been focused on two different approaches to studying neural
networks training in the limit of infinite width (1) a mean-field (MF) and (2) a
constant neural tangent kernel (NTK) approximations. These two approaches have
different scaling of hyperparameters with the width of a network layer and as a
result, different infinite-width limit models. Restricting ourselves to single h
idden layer nets with zero-mean initialization trained for binary classification
with SGD, we propose a general framework to study how the limit behavior of neu
ral models depends on the scaling of hyperparameters with network width. Our fra
mework allows us to derive scaling for existing MF and NTK limits, as well as an
uncountable number of other scalings that lead to a dynamically stable limit be
havior of corresponding models. However, only a finite number of distinct limit
models are induced by these scalings. Each distinct limit model corresponds to a
unique combination of such properties as boundedness of logits and tangent kern
els at initialization or stationarity of tangent kernels. Existing MF and NTK li
mit models, as well as one novel limit model, satisfy most of the properties dem
onstrated by finite-width models. We also propose a novel initialization-correct
ed mean-field limit that satisfies all properties noted above, and its correspon
ding model is a simple modification for a finite-width model.
****

QPLEX: Duplex Dueling Multi-Agent Q-Learning
Jianhao Wang,Zhizhou Ren,Terry Liu,Yang Yu,Chongjie Zhang
We explore value-based multi-agent reinforcement learning (MARL) in the popular
paradigm of centralized training with decentralized execution (CTDE). CTDE has a
n important concept, Individual-Global-Max (IGM) principle, which requires the c
onsistency between joint and local action selections to support efficient local
decision-making. However, in order to achieve scalability, existing MARL methods
either limit representation expressiveness of their value function classes or r
elax the IGM consistency, which may suffer from instability risk or may not perf
orm well in complex domains. This paper presents a novel MARL approach, called d
uPLEX dueling multi-agent Q-learning (QPLEX), which takes a duplex dueling netwo
rk architecture to factorize the joint value function. This duplex dueling struc
ture encodes the IGM principle into the neural network architecture and thus ena
bles efficient value function learning. Theoretical analysis shows that QPLEX ac
hieves a complete IGM function class. Empirical experiments on StarCraft II micr
omanagement tasks demonstrate that QPLEX significantly outperforms state-of-the-
art baselines in both online and offline data collection settings, and also reve
al that QPLEX achieves high sample efficiency and can benefit from offline datas
ets without additional online exploration.
****

Meta-Learning of Structured Task Distributions in Humans and Machines
Sreejan Kumar,Ishita Dasgupta,Jonathan Cohen,Nathaniel Daw,Thomas Griffiths
In recent years, meta-learning, in which a model is trained on a family of tasks

(i.e. a task distribution), has emerged as an approach to training neural networks to perform tasks that were previously assumed to require structured representations, making strides toward closing the gap between humans and machines. However, we argue that evaluating meta-learning remains a challenge, and can miss whether meta-learning actually uses the structure embedded within the tasks. These meta-learners might therefore still be significantly different from humans learners. To demonstrate this difference, we first define a new meta-reinforcement learning task in which a structured task distribution is generated using a compositional grammar. We then introduce a novel approach to constructing a "null task distribution" with the same statistical complexity as this structured task distribution but without the explicit rule-based structure used to generate the structured task. We train a standard meta-learning agent, a recurrent network trained with model-free reinforcement learning, and compare it with human performance across the two task distributions. We find a double dissociation in which humans do better in the structured task distribution whereas agents do better in the null task distribution -- despite comparable statistical complexity. This work highlights that multiple strategies can achieve reasonable meta-test performance, and that careful construction of control task distributions is a valuable way to understand which strategies meta-learners acquire, and how they might differ from humans.

****

All-You-Can-Fit 8-Bit Flexible Floating-Point Format for Accurate and Memory-Efficient Inference of Deep Neural Networks

Juinn-Dar Huang,Cheng-Wei Huang,Tim-Wei Chen

Modern deep neural network (DNN) models generally require a huge amount of weight and activation values to achieve good inference outcomes. Those data inevitably demand a massive off-chip memory capacity/bandwidth, and the situation gets even worse if they are represented in high-precision floating-point formats. Effort has been made for representing those data in different 8-bit floating-point formats, nevertheless, a notable accuracy loss is still unavoidable. In this paper we introduce an extremely flexible 8-bit floating-point (FFP8) format whose defining factors – the bit width of exponent/fraction field, the exponent bias, and even the presence of the sign bit – are all configurable. We also present a methodology to properly determine those factors so that the accuracy of model inference can be maximized. The foundation of this methodology is based on a key observation – both the maximum magnitude and the value distribution are quite dissimilar between weights and activations in most DNN models. Experimental results demonstrate that the proposed FFP8 format achieves an extremely low accuracy loss of $0.1\%\sim 0.3\%$ for several representative image classification models even without the need of model retraining. Besides, it is easy to turn a classical floating-point processing unit into an FFP8-compliant one, and the extra hardware cost is minor.

****

PCPs: Patient Cardiac Prototypes

Dani Kiyasseh,Tingting Zhu,David A. Clifton

Existing deep learning methodologies within the medical domain are typically population-based and difficult to interpret. This limits their clinical utility as population-based findings may not generalize to the individual patient. To overcome these obstacles, we propose to learn patient-specific representations, entitled patient cardiac prototypes (PCPs), that efficiently summarize the cardiac state of a patient. We show that PCPs, learned in an end-to-end manner via contrastive learning, allow for the discovery of similar patients both within and across datasets, and can be exploited for dataset distillation as a compact substitute for the original dataset.

****

Revisiting the Train Loss: an Efficient Performance Estimator for Neural Architecture Search

Binxin Ru,Clare Lyle,Lisa Schut,Mark van der Wilk,Yarin Gal

Reliable yet efficient evaluation of generalisation performance of a proposed architecture is crucial to the success of neural architecture search (NAS). Tradit

ional approaches face a variety of limitations: training each architecture to co
mpletion is prohibitively expensive, early stopping estimates may correlate poor
ly with fully trained performance, and model-based estimators require large trai
ning sets. Instead, motivated by recent results linking training speed and gener
alisation with stochastic gradient descent, we propose to estimate the final tes
t performance based on the sum of training losses. Our estimator is inspired by
 the marginal likelihood, which is used for Bayesian model selection. Our model-f
ree estimator is simple, efficient, and cheap to implement, and does not require
 hyperparameter-tuning or surrogate training before deployment. We demonstrate e
mpirically that our estimator consistently outperforms other baselines under var
ious settings and can achieve a rank correlation of 0.95 with final test accurac
y on the NAS-Bench201 dataset within 50 epochs.
****

CLOPS: Continual Learning of Physiological Signals
Dani Kiyasseh,Tingting Zhu,David A. Clifton
Deep learning algorithms are known to experience destructive interference when i
nstances violate the assumption of being independent and identically distributed
 (i.i.d). This violation, however, is ubiquitous in clinical settings where data
 are streamed temporally and from a multitude of physiological sensors. To overc
ome this obstacle, we propose CLOPS, a replay-based continual learning strategy.
 In three continual learning scenarios based on three publically-available datas
ets, we show that CLOPS can outperform the state-of-the-art methods, GEM and MIR
. Moreover, we propose end-to-end trainable parameters, which we term task-insta
nce parameters, that can be used to quantify task difficulty and similarity. Thi
s quantification yields insights into both network interpretability and clinical
 applications, where task difficulty is poorly quantified.
****

Differentially Private Learning Needs Better Features (or Much More Data)
Florian Tramer,Dan Boneh
We demonstrate that differentially private machine learning has not yet reached
its ''AlexNet moment'' on many canonical vision tasks: linear models trained on
handcrafted features significantly outperform end-to-end deep neural networks fo
r moderate privacy budgets.
To exceed the performance of handcrafted features, we show that private learning
 requires either much more private data, or access to features learned on public
 data from a similar domain.
Our work introduces simple yet strong baselines for differentially private learn
ing that can inform the evaluation of future progress in this area.
****

What About Taking Policy as Input of Value Function: Policy-extended Value Funct
ion Approximator
Hongyao Tang,Zhaopeng Meng,Jianye HAO,Chen Chen,Daniel Graves,Dong Li,Wulong Liu
,Yaodong Yang
The value function lies in the heart of Reinforcement Learning (RL), which defin
es the long-term evaluation of a policy in a given state. In this paper, we prop
ose Policy-extended Value Function Approximator (PeVFA) which extends the conven
tional value to be not only a function of state but also an explicit policy repr
esentation. Such an extension enables PeVFA to preserve values of multiple polic
ies in contrast to a conventional one with limited capacity for only one policy,
 inducing the new characteristic of \emph{value generalization among policies}.
From both the theoretical and empirical lens, we study value generalization alon
g the policy improvement path (called local generalization), from which we deriv
e a new form of Generalized Policy Iteration with PeVFA to improve the conventio
nal learning process. Besides, we propose a framework to learn the representatio
n of an RL policy, studying several different approaches to learn an effective
policy representation from policy network parameters and state-action pairs thro
ugh contrastive learning and action prediction. In our experiments, Proximal Pol
icy Optimization (PPO) with PeVFA significantly outperforms its vanilla counterp
art in MuJoCo continuous control tasks, demonstrating the effectiveness of value
 generalization offered by PeVFA and policy representation learning.

****
CLOCS: Contrastive Learning of Cardiac Signals Across Space, Time, and Patients
Dani Kiyasseh,Tingting Zhu,David A. Clifton
The healthcare industry generates troves of unlabelled physiological data. This data can be exploited via contrastive learning, a self-supervised pre-training method that encourages representations of instances to be similar to one another. We propose a family of contrastive learning methods, CLOCS, that encourages representations across space, time, \textit{and} patients to be similar to one another. We show that CLOCS consistently outperforms the state-of-the-art methods, BYOL and SimCLR, when performing a linear evaluation of, and fine-tuning on, downstream tasks. We also show that CLOCS achieves strong generalization performance with only 25\% of labelled training data. Furthermore, our training procedure naturally generates patient-specific representations that can be used to quantify patient-similarity.
****
Improving robustness of softmax corss-entropy loss via inference information
Bingbing Song,Wei He,Renyang Liu,Shui Yu,Ruxin Wang,Mingming Gong,Tongliang Liu,Wei Zhou
Adversarial examples easily mislead the vision systems based on deep neural networks (DNNs) trained with the softmax cross entropy (SCE) loss. Such a vulnerability of DNN comes from the fact that SCE drives DNNs to fit on the training samples, whereas the resultant feature distributions between the training and adversarial examples are unfortunately misaligned. Several state-of-the-arts start from improving the inter-class separability of training samples by modifying loss functions, where we argue that the adversarial examples are ignored and thus limited robustness to adversarial attacks is resulted. In this paper, we exploit inference region which inspires us to involve a margin-like inference information to SCE, resulting in a novel inference-softmax cross entropy (I-SCE) loss, which is intuitively appealing and interpretable. The inference information is a guarantee to both the inter-class separability and the improved generalization to adversarial examples, which is furthermore demonstrated under the min-max framework. Extensive experiments show that under strong adaptive attacks, the DNN models trained with the proposed I-SCE loss achieve superior performance and robustness over the state-of-the-arts.
****
SoCal: Selective Oracle Questioning for Consistency-based Active Learning of Cardiac Signals
Dani Kiyasseh,Tingting Zhu,David A. Clifton
The ubiquity and rate of collection of cardiac signals produce large, unlabelled datasets. Active learning (AL) can exploit such datasets by incorporating human annotators (oracles) to improve generalization performance. However, the over-reliance of existing algorithms on oracles continues to burden physicians. To minimize this burden, we propose SoCal, a consistency-based AL framework that dynamically determines whether to request a label from an oracle or to generate a pseudo-label instead. We show that our framework decreases the labelling burden while maintaining strong performance, even in the presence of a noisy oracle.
****
PC2WF: 3D Wireframe Reconstruction from Raw Point Clouds
Yujia Liu,Stefano D'Aronco,Konrad Schindler,Jan Dirk Wegner
We introduce PC2WF, the first end-to-end trainable deep network architecture to convert a 3D point cloud into a wireframe model. The network takes as input an unordered set of 3D points sampled from the surface of some object, and outputs a wireframe of that object, i.e., a sparse set of corner points linked by line segments. Recovering the wireframe is a challenging task, where the numbers of both vertices and edges are different for every instance, and a-priori unknown. Our architecture gradually builds up the model: It starts by encoding the points into feature vectors. Based on those features, it identifies a pool of candidate vertices, then prunes those candidates to a final set of corner vertices and refines their locations. Next, the corners are linked with an exhaustive set of candidate edges, which is again pruned to obtain the final wireframe. All steps are

trainable, and errors can be backpropagated through the entire sequence. We validate the proposed model on a publicly available synthetic dataset, for which the ground truth wireframes are accessible, as well as on a new real-world dataset. Our model produces wireframe abstractions of good quality and outperforms several baselines.
****
Improving Abstractive Dialogue Summarization with Conversational Structure and Factual Knowledge
Lulu Zhao,Zeyuan Yang,Weiran Xu,Sheng Gao,Jun Guo
Recently, people have been paying more attention to the abstractive dialogue summarization task. Compared with news text, the information flows of the dialogue exchange between at least two interlocutors, which leads to the necessity of capturing long-distance cross-sentence relations. In addition, the generated summaries commonly suffer from fake facts because the key elements of dialogues often scatter in multiple utterances. However, the existing sequence-to-sequence models are difficult to address these issues. Therefore, it is necessary for researchers to explore the implicit conversational structure to ensure the richness and faithfulness of generated contents. In this paper, we present a Knowledge Graph Enhanced Dual-Copy network (KGEDC), a novel framework for abstractive dialogue summarization with conversational structure and factual knowledge. We use a sequence encoder to draw local features and a graph encoder to integrate global features via the sparse relational graph self-attention network, complementing each other. Besides, a dual-copy mechanism is also designed in decoding process to force the generation conditioned on both the source text and extracted factual knowledge. The experimental results show that our method produces significantly higher ROUGE scores than most of the baselines on both SAMSum corpus and Automobile Master corpus. Human judges further evaluate that outputs of our model contain more richer and faithful information.
****
Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability
Suraj Srinivas,Francois Fleuret
Current methods for the interpretability of discriminative deep neural networks commonly rely on the model's input-gradients, i.e., the gradients of the output logits w.r.t. the inputs. The common assumption is that these input-gradients contain information regarding $p_{\theta} (y\mid \mathbf{x})$, the model's discriminative capabilities, thus justifying their use for interpretability. However, in this work, we show that these input-gradients can be arbitrarily manipulated as a consequence of the shift-invariance of softmax without changing the discriminative function. This leaves an open question: given that input-gradients can be arbitrary, why are they highly structured and explanatory in standard models?

In this work, we re-interpret the logits of standard softmax-based classifiers as unnormalized log-densities of the data distribution and show that input-gradients can be viewed as gradients of a class-conditional generative model $p_{\theta}(\mathbf{x} \mid y)$ implicit in the discriminative model. This leads us to hypothesize that the highly structured and explanatory nature of input-gradients may be due to the alignment of this class-conditional model $p_{\theta}(\mathbf{x} \mid y)$ with that of the ground truth data distribution $p_{\text{data}} (\mathbf{x} \mid y)$. We test this hypothesis by studying the effect of density alignment on gradient explanations. To achieve this density alignment, we use an algorithm called score-matching, and propose novel approximations to this algorithm to enable training large-scale models.

Our experiments show that improving the alignment of the implicit density model with the data distribution enhances gradient structure and explanatory power while reducing this alignment has the opposite effect. This also leads us to conjecture that unintended density alignment in standard neural network training may explain the highly structured nature of input-gradients observed in practice. Overall, our finding that input-gradients capture information regarding an implicit

generative model implies that we need to re-think their use for interpreting discriminative models.
****
Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks
Yige Li,Xixiang Lyu,Nodens Koren,Lingjuan Lyu,Bo Li,Xingjun Ma
Deep neural networks (DNNs) are known vulnerable to backdoor attacks, a training time attack that injects a trigger pattern into a small proportion of training data so as to control the model's prediction at the test time. Backdoor attacks are notably dangerous since they do not affect the model's performance on clean examples, yet can fool the model to make the incorrect prediction whenever the trigger pattern appears during testing. In this paper, we propose a novel defense framework Neural Attention Distillation (NAD) to erase backdoor triggers from backdoored DNNs. NAD utilizes a teacher network to guide the finetuning of the backdoored student network on a small clean subset of data such that the intermediate-layer attention of the student network aligns with that of the teacher network. The teacher network can be obtained by an independent finetuning process on the same clean subset. We empirically show, against 6 state-of-the-art backdoor attacks, NAD can effectively erase the backdoor triggers using only 5\% clean training data without causing obvious performance degradation on clean examples. Our code is available at https://github.com/bboylyg/NAD.
****
Adaptive Optimizers with Sparse Group Lasso
Yun Yue,Suo Tong,Zhen Zhang,Yongchao Liu,Chunyang Wen,Huanjun Bao,Jinjie Gu,Yixiang Mu
We develop a novel framework that adds the regularizers to a family of adaptive optimizers in deep learning, such as MOMENTUM, ADAGRAD, ADAM, AMSGRAD, ADAHESSIAN, and create a new class of optimizers, which are named GROUP MOMENTUM, GROUP ADAGRAD, GROUP ADAM, GROUP AMSGRAD and GROUP ADAHESSIAN, etc., accordingly. We establish theoretically proven convergence guarantees in the stochastic convex settings, based on primal-dual methods. We evaluate the regularized effect of our new optimizers on three large-scale real-world ad click datasets with state-of-the-art deep learning models. The experimental results reveal that compared with the original optimizers with the post-processing procedure which use the magnitude pruning method, the performance of the models can be significantly improved on the same sparsity level. Furthermore, in comparison to the cases without magnitude pruning, our methods can achieve extremely high sparsity with significantly better or highly competitive performance.
****
Data-Efficient Reinforcement Learning with Self-Predictive Representations
Max Schwarzer,Ankesh Anand,Rishab Goel,R Devon Hjelm,Aaron Courville,Philip Bachman
While deep reinforcement learning excels at solving tasks where large amounts of data can be collected through virtually unlimited interaction with the environment, learning from limited interaction remains a key challenge. We posit that an agent can learn more efficiently if we augment reward maximization with self-supervised objectives based on structure in its visual input and sequential interaction with the environment. Our method, Self-Predictive Representations (SPR), trains an agent to predict its own latent state representations multiple steps into the future. We compute target representations for future states using an encoder which is an exponential moving average of the agent’s parameters and we make predictions using a learned transition model. On its own, this future prediction objective outperforms prior methods for sample-efficient deep RL from pixels. We further improve performance by adding data augmentation to the future prediction loss, which forces the agent’s representations to be consistent across multiple views of an observation. Our full self-supervised objective, which combines future prediction and data augmentation, achieves a median human-normalized score of 0.415 on Atari in a setting limited to 100k steps of environment interaction, which represents a 55% relative improvement over the previous state-of-the-art. Notably, even in this limited data regime, SPR exceeds expert human score

s on 7 out of 26 games. We’ve made the code associated with this work available at https://github.com/mila-iqia/spr.
****
Imagine That! Leveraging Emergent Affordances for 3D Tool Synthesis
Yizhe Wu,Sudhanshu Kasewa,Oliver Groth,Sasha Salter,Kevin Li Sun,Oiwi Parker Jones,Ingmar Posner
In this paper we explore the richness of information captured by the latent space of a vision-based generative model. The model combines unsupervised generative learning with a task-based performance predictor to learn and to exploit task-relevant object affordances given visual observations from a reaching task, involving a scenario and a stick-like tool. While the learned embedding of the generative model captures factors of variation in 3D tool geometry (e.g. length, width, and shape), the performance predictor identifies sub-manifolds of the embedding that correlate with task success. Within a variety of scenarios, we demonstrate that traversing the latent space via backpropagation from the performance predictor allows us to imagine tools appropriate for the task at hand. Our results indicate that affordances – like the utility for reaching – are encoded along smooth trajectories in latent space. Accessing these emergent affordances by considering only high-level performance criteria (such as task success) enables an agent to manipulate tool geometries in a targeted and deliberate way.
****
Experimental Design for Overparameterized Learning with Application to Single Shot Deep Active Learning
Neta Shoham,Haim Avron
Abstract The impressive performance exhibited by modern machine learning models hinges on the ability to train such models on a very large amounts of labeled data. However, since access to large volumes of labeled data is often limited or expensive, it is desirable to alleviate this bottleneck by carefully curating the training set. Optimal experimental design is a well-established paradigm for selecting data point to be labeled so to maximally inform the learning process. Unfortunately, classical theory on optimal experimental design focuses on selecting examples in order to learn underparameterized (and thus, non-interpolative) models, while modern machine learning models such as deep neural networks are overparameterized, and oftentimes are trained to be interpolative. As such, classical experimental design methods are not applicable in many modern learning setups. Indeed, the predictive performance of underparameterized models tends to be variance dominated, so classical experimental design focuses on variance reduction, while the predictive performance of overparameterized models can also be, as is shown in this paper, bias dominated or of mixed nature. In this paper we propose a design strategy that is well suited for overparameterized regression and interpolation, and we demonstrate the applicability of our method in the context of deep learning by proposing a new algorithm for single shot deep active learning.
****
CAT-SAC: Soft Actor-Critic with Curiosity-Aware Entropy Temperature
Junfan Lin,Changxin Huang,Xiaodan Liang,Liang Lin
The trade-off between exploration and exploitation has long been a crucial issue in reinforcement learning~(RL). Most of the existing RL methods handle this problem by adding action noise to the policies, such as the Soft Actor-Critic (SAC) that introduces an entropy temperature for maximizing both the external value and the entropy of the policy. However, this temperature is applied indiscriminately to all different environment states, undermining the potential of exploration. In this paper, we argue that the agent should explore more in an unfamiliar state, while less in a familiar state, so as to understand the environment more efficiently. To this purpose, we propose \textbf{C}uriosity-\textbf{A}ware entropy \textbf{T}emperature for SAC (CAT-SAC), which utilizes the curiosity mechanism in developing an instance-level entropy temperature. CAT-SAC uses the state prediction error to model curiosity because an unfamiliar state generally has a large prediction error. The curiosity is added to the target entropy to increase the entropy temperature for unfamiliar states and decrease the target entropy for

familiar states. By tuning the entropy specifically and adaptively, CAT-SAC is encouraged to explore when its curiosity is large, otherwise, it is encouraged to exploit. Experimental results on the difficult MuJoCo benchmark testify that the proposed CAT-SAC significantly improves the sample efficiency, outperforming the advanced model-based / model-free RL baselines.
****
Learning to Observe with Reinforcement Learning
Mehmet Koseoglu,Ece Kunduracioglu,Ayca Ozcelikkale
We consider a decision making problem where an autonomous agent decides on which actions to take based on the observations it collects from the environment. We are interested in revealing the information structure of the observation space illustrating which type of observations are the most important (such as position versus velocity) and the dependence of this on the state of agent (such as at the bottom versus top of a hill). We approach this problem by associating a cost with collecting observations which increases with the accuracy. We adopt a reinforcement learning (RL) framework where the RL agent learns to adjust the accuracy of the observations alongside learning to perform the original task. We consider both the scenario where the accuracy can be adjusted continuously and also the scenario where the agent has to choose between given preset levels, such as taking a sample perfectly or not taking a sample at all. In contrast to the existing work that mostly focuses on sample efficiency during training, our focus is on the behaviour during the actual task. Our results illustrate that the RL agent can learn to use the observation space efficiently and obtain satisfactory performance in the original task while collecting effectively smaller amount of data. By uncovering the relative usefulness of different types of observations and trade-offs within, these results also provide insights for further design of active data acquisition schemes.
****
Revisiting Prioritized Experience Replay: A Value Perspective
Ang A. Li,Zongqing Lu,Chenglin Miao
Reinforcement learning (RL) agents need to learn from past experiences. Prioritized experience replay that weighs experiences by their surprise (the magnitude of the temporal-difference error) significantly improves the learning efficiency for RL algorithms. Intuitively, surprise quantifies the unexpectedness of an experience to the learning agent. But how surprise is related to the importance of experience is not well understood. To address this problem, we derive three value metrics to quantify the importance of experience, which consider the extra reward would be earned by accessing the experience. We theoretically show these value metrics are upper-bounded by surprise for Q-learning. Furthermore, we successfully extend our theoretical framework to maximum-entropy RL by deriving the lower and upper bounds of these value metrics for soft Q-learning, which is also related to surprise. Our framework links two important quantities in RL, i.e., surprise and value of experience, and provides a theoretical basis to estimate the value of experience by surprise. We empirically show that the upper bounds hold in practice, and experience replay using the upper bound as priority improves maximum-entropy RL in Atari games.
****
HyperGrid Transformers: Towards A Single Model for Multiple Tasks
Yi Tay,Zhe Zhao,Dara Bahri,Donald Metzler,Da-Cheng Juan
Achieving state-of-the-art performance on natural language understanding tasks typically relies on fine-tuning a fresh model for every task. Consequently, this approach leads to a higher overall parameter cost, along with higher technical maintenance for serving multiple models. Learning a single multi-task model that is able to do well for all the tasks has been a challenging and yet attractive proposition. In this paper, we propose HyperGrid Transformers, a new Transformer architecture that leverages task-conditioned hyper networks for controlling its feed-forward layers. Specifically, we propose a decomposable hypernetwork that learns grid-wise projections that help to specialize regions in weight matrices for different tasks. In order to construct the proposed hypernetwork, our method learns the interactions and composition between a global (task-agnostic) state a

nd a local task-specific state. We conduct an extensive set of experiments on GL UE/SuperGLUE. On the SuperGLUE test set, we match the performance of the state-o f-the-art while being 16 times more parameter efficient. Our method helps brid ge the gap between fine-tuning and multi-task learning approaches.
****
Efficient Sampling for Generative Adversarial Networks with Reparameterized Mark ov Chains
Yifei Wang,Yisen Wang,Jiansheng Yang,Zhouchen Lin
Recently, sampling methods have been successfully applied to enhance the sample quality of Generative Adversarial Networks (GANs). However, in practice, they ty pically have poor sample efficiency because of the independent proposal sampling from the generator. In this work, we propose REP-GAN, a novel sampling method t hat allows general dependent proposals by REParameterizing the Markov chains int o the latent space of the generator. Theoretically, we show that our reparameter ized proposal admits a closed-form Metropolis-Hastings acceptance ratio. Empiric ally, extensive experiments on synthetic and real datasets demonstrate that our REP-GAN largely improves the sample efficiency and obtains better sample quality simultaneously.
****
One-class Classification Robust to Geometric Transformation
Hyunjun Ju,Dongha Lee,SeongKu Kang,Hwanjo Yu
Recent studies on one-class classification have achieved a remarkable performanc e, by employing the self-supervised classifier that predicts the geometric trans formation applied to in-class images. However, they cannot identify in-class ima ges at all when the input images are geometrically-transformed (e.g., rotated im ages), because their classification-based in-class scores assume that input imag es always have a fixed viewpoint, as similar to the images used for training. Po inting out that humans can easily recognize such transformed images as the same class, in this work, we aim to propose a one-class classifier robust to geometri cally-transformed inputs, named as GROC. To this end, we introduce a conformity score which indicates how strongly an input image agrees with one of the predefi ned in-class transformations, then utilize the conformity score with our propose d agreement measures for one-class classification. Our extensive experiments dem onstrate that GROC is able to accurately distinguish in-class images from out-of -class images regardless of whether the inputs are geometrically-transformed or not, whereas the existing methods fail.
****
Predicting the Outputs of Finite Networks Trained with Noisy Gradients
Gadi Naveh,Oded Ben-David,Haim Sompolinsky,Zohar Ringel
A recent line of works studied wide deep neural networks (DNNs) by approximating them as Gaussian Processes (GPs). A DNN trained with gradient flow was shown to map to a GP governed by the Neural Tangent Kernel (NTK), whereas earlier works showed that a DNN with an i.i.d. prior over its parameters maps to the so-called Neural Network Gaussian Process (NNGP). Here we consider a DNN training protoco l, involving noise, weight decay and finite width, whose outcome corresponds to a certain non-Gaussian stochastic process. An analytical framework is then intro duced to analyze this non-Gaussian process, whose deviation from a GP is control led by the finite width. Our contribution is three-fold:
(i) In the infinite width limit, we establish a correspondence between DNNs trai ned with noisy gradients and the NNGP, not the NTK.
(ii) We provide a general analytical form for the finite width correction (FWC) for DNNs with arbitrary activation functions and depth and use it to predict the outputs of empirical finite networks with high accuracy.
Analyzing the FWC behavior as a function of n, the training set size, we find that it is negligible for both the very small n regime, and, surprisingly, for the large n regime (where the GP error scales as $O(1/n)$).
(iii) We flesh-out algebraically how these FWCs can improve the performance of f inite convolutional neural networks (CNNs) relative to their GP counterparts on image classification tasks.
****

Distributed Training of Graph Convolutional Networks using Subgraph Approximation
Alexandra Angerd,Keshav Balasubramanian,Murali Annavaram
Modern machine learning techniques are successfully being adapted to data modeled as graphs. However, many real-world graphs are typically very large and do not fit in memory, often making the problem of training machine learning models on them intractable. Distributed training has been successfully employed to alleviate memory problems and speed up training in machine learning domains in which the input data is assumed to be independently identical distributed (i.i.d). However, distributing the training of non i.i.d data such as graphs that are used as training inputs in Graph Convolutional Networks (GCNs) causes accuracy problems since information is lost at the graph partitioning boundaries.

In this paper, we propose a training strategy that mitigates the lost information across multiple partitions of a graph through a subgraph approximation scheme. Our proposed approach augments each sub-graph with a small amount of edge and vertex information that is approximated from all other sub-graphs. The subgraph approximation approach helps the distributed training system converge at single-machine accuracy, while keeping the memory footprint low and minimizing synchronization overhead between the machines.
****
Long Range Arena : A Benchmark for Efficient Transformers
Yi Tay,Mostafa Dehghani,Samira Abnar,Yikang Shen,Dara Bahri,Philip Pham,Jinfeng Rao,Liu Yang,Sebastian Ruder,Donald Metzler
Transformers do not scale very well to long sequence lengths largely because of quadratic self-attention complexity. In the recent months, a wide spectrum of efficient, fast Transformers have been proposed to tackle this problem, more often than not claiming superior or comparable model quality to vanilla Transformer models. To this date, there is no well-established consensus on how to evaluate this class of models. Moreover, inconsistent benchmarking on a wide spectrum of tasks and datasets makes it difficult to assess relative model quality amongst many models. This paper proposes a systematic and unified benchmark, Long Range Arena, specifically focused on evaluating model quality under long-context scenarios. Our benchmark is a suite of tasks consisting of sequences ranging from $1K$ to $16K$ tokens, encompassing a wide range of data types and modalities such as text, natural, synthetic images, and mathematical expressions requiring similarity, structural, and visual-spatial reasoning. We systematically evaluate ten well-established long-range Transformer models (Reformers, Linformers, Linear Transformers, Sinkhorn Transformers, Performers, Synthesizers, Sparse Transformers, and Longformers) on our newly proposed benchmark suite. Long Range Arena paves the way towards better understanding this class of efficient Transformer models, facilitates more research in this direction, and presents new challenging tasks to tackle.
****
Acting in Delayed Environments with Non-Stationary Markov Policies
Esther Derman,Gal Dalal,Shie Mannor
The standard Markov Decision Process (MDP) formulation hinges on the assumption that an action is executed immediately after it was chosen. However, assuming it is often unrealistic and can lead to catastrophic failures in applications such as robotic manipulation, cloud computing, and finance. We introduce a framework for learning and planning in MDPs where the decision-maker commits actions that are executed with a delay of m steps. The brute-force state augmentation baseline where the state is concatenated to the last m committed actions suffers from an exponential complexity in m, as we show for policy iteration. We then prove that with execution delay, deterministic Markov policies in the original state-space are sufficient for attaining maximal reward, but need to be non-stationary. As for stationary Markov policies, we show they are sub-optimal in general. Consequently, we devise a non-stationary Q-learning style model-based algorithm that solves delayed execution tasks without resorting to state-augmentation. Experiments on tabular, physical, and Atari domains reveal that it converges quic

kly to high performance even for substantial delays, while standard approaches t
hat either ignore the delay or rely on state-augmentation struggle or fail due t
o divergence. The code is available at \url{https://github.com/galdl/rl_delay_ba
sic.git}.
****

Neurally Augmented ALISTA
Freya Behrens,Jonathan Sauder,Peter Jung
 It is well-established that many iterative sparse reconstruction algorithms can
 be unrolled to yield a learnable neural network for improved empirical performa
nce. A prime example is learned ISTA (LISTA) where weights, step sizes and thres
holds are learned from training data. Recently, Analytic LISTA (ALISTA) has been
 introduced, combining the strong empirical performance of a fully learned appro
ach like LISTA, while retaining theoretical guarantees of classical compressed s
ensing algorithms and significantly reducing the number of parameters to learn.
However, these parameters are trained to work in expectation, often leading to s
uboptimal reconstruction of individual targets. In this work we therefore intro
duce Neurally Augmented ALISTA, in which an LSTM network is used to compute step
 sizes and thresholds individually for each target vector during reconstruction.
 This adaptive approach is theoretically motivated by revisiting the recovery gu
arantees of ALISTA. We show that our approach further improves empirical perform
ance in sparse reconstruction, in particular outperforming existing algorithms b
y an increasing margin as the compression ratio becomes more challenging.
****

Selfish Sparse RNN Training
SHiwei Liu,Decebal Constantin Mocanu,Yulong Pei,Mykola Pechenizkiy
Sparse neural networks have been widely applied to reduce the necessary resource
 requirements to train and deploy over-parameterized deep neural networks. For i
nference acceleration, methods that induce sparsity from a pre-trained dense net
work (dense-to-sparse) work effectively. Recently, dynamic sparse training (DST)
 has been proposed to train sparse neural networks without pre-training a large
and dense network (sparse-to-sparse), so that the training process can also be a
ccelerated. However, previous sparse-to-sparse methods mainly focus on Multilaye
r Perceptron Networks (MLPs) and Convolutional Neural Networks (CNNs), failing t
o match the performance of dense-to-sparse methods in Recurrent Neural Networks
(RNNs) setting. In this paper, we propose an approach to train sparse RNNs with
a fixed parameter count in one single run, without compromising performance. Dur
ing training, we allow RNN layers to have a non-uniform redistribution across ce
ll weights for a better regularization. Further, we introduce SNT-ASGD, a varian
t of the averaged stochastic gradient optimizer, which significantly improves th
e performance of all sparse training methods for RNNs. Using these strategies, w
e achieve state-of-the-art sparse training results, even better than dense model
 results, with various types of RNNs on Penn TreeBank and Wikitext-2 datasets.
****

Translation Memory Guided Neural Machine Translation
Shaohui Kuang,Heng Yu,Weihua Luo,Qiang Wang
Many studies have proven that Translation Memory (TM) can help improve the trans
lation quality of neural machine translation (NMT). Existing ways either employ
extra encoder to encode information from TM or concatenate source sentence and T
M sentences as encoder's input. These previous methods don't model the semantic
relationship between the source sentence and TM sentences. Meanwhile, the traini
ng corpus related to TM is limited, and the sentence level retrieval approach fu
rther limits its scale.
In this paper, we propose a novel method to combine the strengths of both TM and
 NMT. We treat the matched sentence pair of TM as the additional signal and appl
y one encoder enhanced by the pre-trained language model (PLM) to encode the TM
information and source sentence together. Additionally, we extend the sentence l
evel retrieval method to the n-gram retrieval method that we don't need to calcu
late the similarity score. Further, we explore new methods to manipulate the inf
ormation flow from TM to the NMT decoder. We validate our proposed methods on a
mixed test set of multiple domains. Experiment results demonstrate that the prop

osed methods can significantly improve the translation quality and show strong a daptation for an unknown or new domain.
****
Domain-slot Relationship Modeling using a Pre-trained Language Encoder for Mult i-Domain Dialogue State Tracking
Jinwon An,Misuk Kim,Sungzoon Cho,Junseong Bang
Dialogue state tracking for multi-domain dialogues is challenging because the mo del should be able to track dialogue states across multiple domains and slots. P ast studies had its limitations in that they did not factor in the relationship among different domain-slot pairs. Although recent approaches did support relati onship modeling among the domain-slot pairs, they did not leverage a pre-trained language model, which has improved the performance of numerous natural language tasks, in the encoding process. Our approach fills the gap between these previo us studies. We propose a model for multi-domain dialogue state tracking that eff ectively models the relationship among domain-slot pairs using a pre-trained lan guage encoder. Inspired by the way the special $[CLS]$ token in BERT is used to aggregate the information of the whole sequence, we use multiple special tokens for each domain-slot pair that encodes information corresponding to its domain a nd slot. The special tokens are run together with the dialogue context through t he pre-trained language encoder, which effectively models the relationship among different domain-slot pairs. Our experimental results show that our model achie ves state-of-the-art performance on the MultiWOZ-2.1 and MultiWOZ-2.2 dataset.
****
Waste not, Want not: All-Alive Pruning for Extremely Sparse Networks
Daejin Kim,Hyunjung Shim,Jongwuk Lee
Network pruning has been widely adopted for reducing computational cost and memo ry consumption in low-resource devices. Recent studies show that saliency-based pruning can achieve high compression ratios (e.g., 80-90% of the parameters in o riginal networks are removed) without sacrificing much accuracy loss. Neverthele ss, finding the well-trainable networks with sparse parameters (e.g., < 10% of t he parameters remaining) is still challenging to network pruning, commonly belie ved to lack model capacity. In this work, we revisit the procedure of existing p runing methods and observe that dead connections, which do not contribute to mod el capacity, appear regardless of pruning methods. To this end, we propose a nov el pruning method, called all-alive pruning (AAP), producing the pruned networks with only trainable weights. Notably, AAP is broadly applicable to various sali ency-based pruning methods and model architectures. We demonstrate that AAP equi pped with existing pruning methods (i.e., iterative pruning, one-shot pruning, a nd dynamic pruning) consistently improves the accuracy of original methods at 12 8× - 4096× compression ratios on three benchmark datasets.
****
DO-GAN: A Double Oracle Framework for Generative Adversarial Networks
Aye Phyu Phyu Aung,Xinrun Wang,Runsheng Yu,Bo An,Senthilnath Jayavelu,Xiaoli Li
In this paper, we propose a new approach to train Generative Adversarial Network s (GAN) where we deploy a double-oracle framework using the generator and discri minator oracles. GAN is essentially a two-player zero-sum game between the gener ator and the discriminator. Training GANs is challenging as a pure Nash equilibr ium may not exist and even finding the mixed Nash equilibrium is difficult as GA Ns have a large-scale strategy space. In DO-GAN, we extend the double oracle fra mework to GANs. We first generalize the player strategies as the trained models of generator and discriminator from the best response oracles. We then compute t he meta-strategies using a linear program. Next, we prune the weakly-dominated p layer strategies to keep the oracles from becoming intractable. We apply our fra mework to established architectures such as vanilla GAN, Deep Convolutional GAN, Spectral Normalization GAN and Stacked GAN. Finally, we conduct evaluations on MNIST, CIFAR-10 and CelebA datasets and show that DO-GAN variants have significa nt improvements in both subjective qualitative evaluation and quantitative metri cs, compared with their respective GAN architectures.
****
Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Mo

dels
Yuge Shi,Brooks Paige,Philip Torr,Siddharth N
Multimodal learning for generative models often refers to the learning of abstra
ct concepts from the commonality of information in multiple modalities, such as
vision and language. While it has proven effective for learning generalisable re
presentations, the training of such models often requires a large amount of rela
ted multimodal data that shares commonality, which can be expensive to come by.
To mitigate this, we develop a novel contrastive framework for generative model
learning, allowing us to train the model not just by the commonality between mod
alities, but by the distinction between "related" and "unrelated" multimodal dat
a. We show in experiments that our method enables data-efficient multimodal lear
ning on challenging datasets for various multimodal VAE models. We also show tha
t under our proposed framework, the generative model can accurately identify rel
ated samples from unrelated ones, making it possible to make use of the plentifu
l unlabeled, unpaired multimodal data.
****

Learning Private Representations with Focal Entropy
Tassilo Klein,Moin Nabi
How can we learn a representation with good predictive power while preserving us
er privacy?
We present an adversarial representation learning method to sanitize sensitive c
ontent from the representation in an adversarial fashion.
Specifically, we propose focal entropy - a variant of entropy embedded in an adv
ersarial representation learning setting to leverage privacy sanitization. Focal
 entropy enforces maximum uncertainty in terms of confusion on the subset of pri
vacy-related similar classes, separated from the dissimilar ones. As such, our p
roposed sanitization method yields deep sanitization of private features yet is
conceptually simple and empirically powerful. We showcase feasibility in terms o
f classification of facial attributes and identity on the CelebA dataset as well
 as CIFAR-100. The results suggest that private components can be removed reliab
ly.
****

Deep Continuous Networks
Nergis Tomen,Silvia Laura Pintea,Jan van Gemert
CNNs and computational models of biological vision share some fundamental princi
ples, which, combined with recent developments in deep learning, have opened up
new avenues of research in neuroscience. However, in contrast to biological mode
ls, conventional CNN architectures are based on spatio-temporally discrete repre
sentations, and thus cannot accommodate certain aspects of biological complexity
 such as continuously varying receptive field sizes and temporal dynamics of neu
ronal responses. Here we propose deep continuous networks (DCNs), which combine
spatially continuous convolutional filter representations, with the continuous t
ime framework of neural ODEs. This allows us to learn the spatial support of the
 filters during training, as well as model the temporal evolution of feature map
s, linking DCNs closely to biological models. We show that DCNs are versatile. E
xperimentally, we demonstrate their applicability to a standard classification p
roblem, where they allow for parameter reductions and meta-parametrization. We i
llustrate the biological plausibility of the scale distributions learned by DCNs
 and explore their performance in a pattern completion task, which is inspired b
y models from computational neuroscience. Finally, we suggest that the continuou
s representations learned by DCNs may enable computationally efficient implement
ations.
****

The Lipschitz Constant of Self-Attention
Hyunjik Kim,George Papamakarios,Andriy Mnih
Lipschitz constants of neural networks have been explored in various contexts in
 deep learning, such as provable adversarial robustness, estimating Wasserstein
distance, stabilising training of GANs, and formulating invertible neural networ
ks. Such works have focused on bounding the Lipschitz constant of fully connecte
d or convolutional networks, composed of linear maps and pointwise non-lineariti

es. In this paper, we investigate the Lipschitz constant of self-attention, a no
n-linear neural network module widely used in sequence modelling. We prove that
the standard dot-product self-attention is *not* Lipschitz, and propose an alter
native L2 self-attention that *is* Lipschitz. We derive an upper bound on the Li
pschitz constant of L2 self-attention and provide empirical evidence for its asy
mptotic tightness. To demonstrate the practical relevance of our theoretical wor
k, we formulate invertible self-attention and use it in a Transformer-based arch
itecture for a character-level language modelling task.
****

Identifying Informative Latent Variables Learned by GIN via Mutual Information
Chen Zhang,Yitong Sun,Mingtian Zhang
How to learn a good representation of data is one of the most important topics o
f machine learning. Disentanglement of representations, though believed to be t
he core feature of good representations, has caused a lot of debates and discuss
ions in recent. Sorrenson et al. (2020), using the techniques developed in nonli
near independent analysis theory, show that general incompressible-flow networks
 (GIN) can recover the underlying latent variables that generate the data, and t
hus can provide a compact and disentangled representation. However, in this pape
r, we point out that the method taken by GIN for informative latent variables id
entification is not theoretically supported and can be disproved by experiments.
  We propose to use the mutual information between latent variables and the auxi
liary variable to correctly identify informative latent variables. We directly
verify the improvement brought by our method in experiments on synthetic data.
We further show the advantage of our method on various downstream tasks includin
g classification, outlier detection and adversarial attack defence.
****

A Reduction Approach to Constrained Reinforcement Learning
Tianchi Cai,Wenjie Shi,Lihong Gu,Xiaodong Zeng,Jinjie Gu
Many applications of reinforcement learning (RL) optimize a long-term reward sub
ject to risk, safety, budget, diversity or other constraints. Though constrained
 RL problem has been studied to incorporate various constraints, existing method
s either tie to specific families of RL algorithms or require storing infinitely
 many individual policies found by an RL oracle to approach a feasible solution.
 In this paper, we present a novel reduction approach for constrained RL problem
 that ensures convergence when using any off-the-shelf RL algorithm to construct
 an RL oracle yet requires storing at most constantly many policies. The key ide
a is to reduce the constrained RL problem to a distance minimization problem, an
d a novel variant of Frank-Wolfe algorithm is proposed for this task. Throughout
 the learning process, our method maintains at most constantly many individual p
olicies, where the constant is shown to be worst-case optimal to ensure converge
nce of any RL oracle. Our method comes with rigorous convergence and complexity
analysis, and does not introduce any extra hyper-parameter. Experiments on a gri
d-world navigation task demonstrate the efficiency of our method.

****

Mime: Mimicking Centralized Stochastic Algorithms in Federated Learning
Sai Praneeth Karimireddy,Martin Jaggi,Satyen Kale,Mehryar Mohri,Sashank J. Reddi
,Sebastian U Stich,Ananda Theertha Suresh
Federated learning is a challenging optimization problem due to the heterogeneit
y of the data across different clients. Such heterogeneity has been observed to
induce \emph{client drift} and significantly degrade the performance of algorith
ms designed for this setting.
In contrast, centralized learning with centrally collected data does not experie
nce such drift, and has seen great empirical and theoretical progress with innov
ations such as momentum, adaptivity, etc.
In this work, we propose a general framework {\sc Mime} which mitigates client-d
rift and adapts arbitrary centralized optimization algorithms (e.g. SGD, Adam, e
tc.) to federated learning.
{\sc Mime} uses a combination of \emph{control-variates} and \emph{server-level
statistics} (e.g. momentum) at every client-update step to ensure that each loca

l update mimics that of the centralized method. Our thorough theoretical and emp
irical analyses strongly establish \mime's superiority over other baselines.
****
Categorical Normalizing Flows via Continuous Transformations
Phillip Lippe,Efstratios Gavves
Despite their popularity, to date, the application of normalizing flows on categ
orical data stays limited. The current practice of using dequantization to map d
iscrete data to a continuous space is inapplicable as categorical data has no in
trinsic order. Instead, categorical data have complex and latent relations that
must be inferred, like the synonymy between words. In this paper, we investigate
 Categorical Normalizing Flows, that is normalizing flows for categorical data.
By casting the encoding of categorical data in continuous space as a variational
 inference problem, we jointly optimize the continuous representation and the mo
del likelihood. Using a factorized decoder, we introduce an inductive bias to mo
del any interactions in the normalizing flow. As a consequence, we do not only s
implify the optimization compared to having a joint decoder, but also make it po
ssible to scale up to a large number of categories that is currently impossible
with discrete normalizing flows. Based on Categorical Normalizing Flows, we prop
ose GraphCNF a permutation-invariant generative model on graphs. GraphCNF implem
ents a three step approach modeling the nodes, edges, and adjacency matrix stepw
ise to increase efficiency. On molecule generation, GraphCNF outperforms both on
e-shot and autoregressive flow-based state-of-the-art.

****
Offline Meta Learning of Exploration
Ron Dorfman,Aviv Tamar
Consider the following problem: given the complete training histories of N con
ventional RL agents, trained on N different tasks, design a meta-agent that ca
n quickly maximize reward in a new, unseen task from the same task distribution.
 In particular, while each conventional RL agent explored and exploited its own
different task, the meta-agent must identify regularities in the data that lead
to effective exploration/exploitation in the unseen task. This meta-learning pro
blem is an instance of a setting we term Offline Meta Reinforcement Learning (OM
RL). To solve our challenge, we take a Bayesian RL (BRL) view, and seek to learn
 a Bayes-optimal policy from the offline data. We extend the recently proposed V
ariBAD BRL algorithm to the off-policy setting, and demonstrate learning of appr
oximately Bayes-optimal exploration strategies from offline data using deep neur
al networks. For the particular problem described above, our method learns effec
tive exploration behavior that is qualitatively different from the exploration u
sed by any RL agent in the data. Furthermore, we find that when applied to the o
nline meta-RL setting (agent simultaneously collects data and improves its meta-
RL policy), our method is significantly more sample efficient than the state-of-
the-art VariBAD.
****
Loss Landscape Matters: Training Certifiably Robust Models with Favorable Loss L
andscape
Sungyoon Lee,Woojin Lee,Jinseong Park,Jaewook Lee
In this paper, we study the problem of training certifiably robust models. Certi
fiable training minimizes an upper bound on the worst-case loss over the allowed
 perturbation, and thus the tightness of the upper bound is an important factor
in building certifiably robust models. However, many studies have shown that Int
erval Bound Propagation (IBP) training uses much looser bounds but outperforms o
ther models that use tighter bounds. We identify another key factor that influen
ces the performance of certifiable training: \textit{smoothness of the loss land
scape}. We consider linear relaxation based methods and find significant differe
nces in the loss landscape across these methods. Based on this analysis, we prop
ose a certifiable training method that utilizes a tighter upper bound and has a
landscape with favorable properties. The proposed method achieves performance co
mparable to state-of-the-art methods under a wide range of perturbations.
****

Prototypical Representation Learning for Relation Extraction
Ning Ding,Xiaobin Wang,Yao Fu,Guangwei Xu,Rui Wang,Pengjun Xie,Ying Shen,Fei Hua
ng,Hai-Tao Zheng,Rui Zhang
Recognizing relations between entities is a pivotal task of relational learning.

Learning relation representations from distantly-labeled datasets is difficult b
ecause of the abundant label noise and complicated expressions in human language
.
This paper aims to learn predictive, interpretable, and robust relation represen
tations from distantly-labeled data that are effective in different settings, in
cluding supervised, distantly supervised, and few-shot learning.
Instead of solely relying on the supervision from noisy labels, we propose to le
arn prototypes for each relation from contextual information to best explore the
 intrinsic semantics of relations.
Prototypes are representations in the feature space abstracting the essential se
mantics of relations between entities in sentences.
We learn prototypes based on objectives with clear geometric interpretation, whe
re the prototypes are unit vectors uniformly dispersed in a unit ball, and state
ment embeddings are centered at the end of their corresponding prototype vectors
 on the surface of the ball.
This approach allows us to learn meaningful, interpretable prototypes for the fi
nal classification.
Results on several relation learning tasks show that our model significantly out
performs the previous state-of-the-art models.
We further demonstrate the robustness of the encoder and the interpretability of
 prototypes with extensive experiments.
****
On the Decision Boundaries of Neural Networks. A Tropical Geometry Perspective
Motasem Alfarra,Adel Bibi,Hasan Abed Al Kader Hammoud,Mohamed Gaafar,Bernard Gha
nem
This work tackles the problem of characterizing and understanding the decision b
oundaries of neural networks with piecewise linear non-linearity activations. We
 use tropical geometry, a new development in the area of algebraic geometry, to
characterize the decision boundaries of a simple network of the form (Affine, Re
LU, Affine). Our main finding is that the decision boundaries are a subset of a
tropical hypersurface, which is intimately related to a polytope formed by the c
onvex hull of two zonotopes. The generators of these zonotopes are functions of
the network parameters. This geometric characterization provides new perspective
s to three tasks. Specifically, we propose a new tropical perspective to the lot
tery ticket hypothesis, where we view the effect of different initializations on
 the tropical geometric representation of a network's decision boundaries. Moreo
ver, we propose new tropical based optimization problems that directly influence
 the decision boundaries of the network for the tasks of network pruning (removi
ng network parameters not contributing to the tropical geometric representation
of the decision boundaries) and the generation of adversarial attacks.
****
Generalized Energy Based Models
Michael Arbel,Liang Zhou,Arthur Gretton
We introduce the Generalized Energy Based Model (GEBM) for generative modelling.
 These models combine two trained components: a base distribution (generally an
 implicit model), which can learn the support of data with low intrinsic dimensi
on in a high dimensional space; and an energy function, to refine the probabilit
y mass on the learned support.
Both the energy function and base jointly constitute the final model, unlike GAN
s, which retain only the base distribution (the "generator").
GEBMs are trained by alternating between learning the energy and the base.
We show that both training stages are well-defined: the energy is learned by max
imising a generalized likelihood, and the resulting energy-based loss provides i
nformative gradients for learning the base.
Samples from the posterior on the latent space of the trained model can be obtai

ned via MCMC, thus finding regions in this space that produce better quality samples.
Empirically, the GEBM samples on image-generation tasks are of much better quality than those from the learned generator alone, indicating that all else being equal, the GEBM will outperform a GAN of the same complexity. When using normalizing flows as base measures, GEBMs succeed on density modelling tasks returning comparable performance to direct maximum likelihood of the same networks.
****
AriEL: Volume Coding for Sentence Generation Comparisons
Luca Celotti,Simon Brodeur,Jean Rouat
Mapping sequences of discrete data to a point in a continuous space makes it difficult to retrieve those sequences via random sampling. Mapping the input to a volume would make it easier to retrieve at test time, and that is the strategy followed by the family of approaches based on Variational Autoencoder. However the fact that they are at the same time optimizing for prediction and for smoothness of representation, forces them to trade-off between the two. We benchmark the performance of some of the standard methods in deep learning to generate sentences by uniformly sampling a continuous space. We do it by proposing AriEL, that constructs volumes in a continuous space, without the need of encouraging the creation of volumes through the loss function. We first benchmark on a toy grammar, that allows to automatically evaluate the language learned and generated by the models. Then, we benchmark on a real dataset of human dialogues. Our results indicate that the random access to the stored information can be significantly improved, since our method AriEL is able to generate a wider variety of correct language by randomly sampling the latent space. VAE follows in performance for the toy dataset while, AE and Transformer follow for the real dataset. This partially supports the hypothesis that encoding information into volumes instead of into points, leads to improved retrieval of learned information with random sampling. We hope this analysis can clarify directions to lead to better generators.
****
Predicting Classification Accuracy When Adding New Unobserved Classes
Yuli Slavutsky,Yuval Benjamini
Multiclass classifiers are often designed and evaluated only on a sample from the classes on which they will eventually be applied. Hence, their final accuracy remains unknown. In this work we study how a classifier’s performance over the initial class sample can be used to extrapolate its expected accuracy on a larger, unobserved set of classes. For this, we define a measure of separation between correct and incorrect classes that is independent of the number of classes: the "reversed ROC" (rROC), which is obtained by replacing the roles of classes and data-points in the common ROC. We show that the classification accuracy is a function of the rROC in multiclass classifiers, for which the learned representation of data from the initial class sample remains unchanged when new classes are added. Using these results we formulate a robust neural-network-based algorithm, "CleaneX", which learns to estimate the accuracy of such classifiers on arbitrarily large sets of classes. Unlike previous methods, our method uses both the observed accuracies of the classifier and densities of classification scores, and therefore achieves remarkably better predictions than current state-of-the-art methods on both simulations and real datasets of object detection, face recognition, and brain decoding.
****
In Search of Lost Domain Generalization
Ishaan Gulrajani,David Lopez-Paz
The goal of domain generalization algorithms is to predict well on distributions different from those seen during training.
While a myriad of domain generalization algorithms exist, inconsistencies in experimental conditions---datasets, network architectures, and model selection criteria---render fair comparisons difficult.
The goal of this paper is to understand how useful domain generalization algorithms are in realistic settings.
As a first step, we realize that model selection is non-trivial for domain gener

alization tasks, and we argue that algorithms without a model selection criterio
n remain incomplete.
Next we implement DomainBed, a testbed for domain generalization including seven
 benchmarks, fourteen algorithms, and three model selection criteria.
When conducting extensive experiments using DomainBed we find that when carefull
y implemented and tuned, ERM outperforms the state-of-the-art in terms of averag
e performance.
Furthermore, no algorithm included in DomainBed outperforms ERM by more than one
 point when evaluated under the same experimental conditions.
We hope that the release of DomainBed, alongside contributions from fellow resea
rchers, will streamline reproducible and rigorous advances in domain generalizat
ion.
****

Estimating informativeness of samples with Smooth Unique Information
Hrayr Harutyunyan,Alessandro Achille,Giovanni Paolini,Orchid Majumder,Avinash Ra
vichandran,Rahul Bhotika,Stefano Soatto
We define a notion of information that an individual sample provides to the trai
ning of a neural network, and we specialize it to measure both how much a sample
 informs the final weights and how much it informs the function computed by the
weights. Though related, we show that these quantities have a qualitatively dif
ferent behavior. We give efficient approximations of these quantities using a li
nearized network and demonstrate empirically that the approximation is accurate
for real-world architectures, such as pre-trained ResNets. We apply these measur
es to several problems, such as dataset summarization, analysis of under-sampled
 classes, comparison of informativeness of different data sources, and detection
 of adversarial and corrupted examples. Our work generalizes existing frameworks
, but enjoys better computational properties for heavily over-parametrized model
s, which makes it possible to apply it to real-world networks.
****

Causal Probabilistic Spatio-temporal Fusion Transformers in Two-sided Ride-Haili
ng Markets
Shixiang Wan,Shikai Luo,Hongtu Zhu
Achieving accurate spatio-temporal predictions in large-scale systems is extreme
ly valuable in many real-world applications, such as weather forecasts, retail f
orecasting, and urban traffic forecasting. So far, most existing methods for mul
ti-horizon, multi-task and multi-target predictions select important predicting
variables via their correlations with responses, and thus it is highly possible
that many forecasting models generated from those methods are not causal, leadin
g to poor interpretability. The aim of this paper is to develop a collaborative
causal spatio-temporal fusion transformer, named CausalTrans, to establish the c
ollaborative causal effects of predictors on multiple forecasting targets, such
as supply and demand in ride-sharing platforms. Specifically, we integrate the c
ausal attention with the Conditional Average Treatment Effect (CATE) estimation
method for causal inference. Moreover, we propose a novel and fast multi-head at
tention evolved from Taylor expansion instead of softmax, reducing time complexi
ty from $O(\mathcal{V}^2)$ to $O(\mathcal{V})$, where $\mathcal{V}$ is the numbe
r of nodes in a graph. We further design a spatial graph fusion mechanism to sig
nificantly reduce the parameters' scale. We conduct a wide range of experiments
to demonstrate the interpretability of causal attention, the effectiveness of va
rious model components, and the time efficiency of our CausalTrans. As shown in
these experiments, our CausalTrans framework can achieve up to 15$\%$ error redu
ction compared with various baseline methods.
****

Contextual Knowledge Distillation for Transformer Compression
Geondo Park,Gyeongman Kim,Eunho Yang
A computationally expensive and memory intensive neural network lies behind the
recent success of language representation learning. Knowledge distillation, a ma
jor technique for deploying such a vast language model in resource-scarce enviro
nments, transfers the knowledge on individual word representations learned witho
ut restrictions. In this paper, inspired by the recent observations that languag

e representations are relatively positioned and have more semantic knowledge as a whole, we present a new knowledge distillation strategy for language representation learning that transfers the contextual knowledge via two types of relationships across representations: Word Relation and Layer Transforming Relation. We validate the effectiveness of our method on challenging benchmarks of language understanding tasks. The code will be released.

****

3D Scene Compression through Entropy Penalized Neural Representation Functions
Thomas Bird,Johannes Ballé,Saurabh Singh,Philip Chou

Some forms of novel visual media enable the viewer to explore a 3D scene from essentially arbitrary viewpoints, by interpolating between a discrete set of original views. Compared to 2D imagery, these types of applications require much larger amounts of storage space, which we seek to reduce. Existing approaches for compressing 3D scenes are based on a separation of compression and rendering: each of the original views is compressed using traditional 2D image formats; the receiver decompresses the views and then performs the rendering. We unify these steps by directly compressing an implicit representation of the scene, a function that maps spatial coordinates to a radiance vector field, which can then be queried to render arbitrary viewpoints. The function is implemented as a neural network and jointly trained for reconstruction as well as compressibility, in an end-to-end manner, with the use of an entropy penalty on the parameters. Our method significantly outperforms a state-of-the-art conventional approach for scene compression, achieving simultaneously higher quality reconstructions and lower bitrates. Furthermore, we show that the performance at lower bitrates can be improved by jointly representing multiple scenes using a soft form of parameter sharing.

****

FILTRA: Rethinking Steerable CNN by Filter Transform
Bo Li,Qili Wang,Gim Hee Lee

Steerable CNN imposes the prior knowledge of transformation invariance or equivariance in the network architecture to enhance the the network robustness on geometry transformation of data and reduce overfitting. Filter transform has been an intuitive and widely used technique to construct steerable CNN in the past decades. Recently, group representation theory is used to analyze steerable CNN and reveals the function space structure of a steerable kernel function. However, it is not yet clear on how this theory is related to the filter transform technique. In this paper, we show that kernel constructed by filter transform can also be interpreted in the group representation theory. Meanwhile, we show that filter transformed kernels can be used to convolve input/output features in different group representation. This interpretation help complete the puzzle of steerable CNN theory and provides a novel and simple approach to implement steerable convolution operators. Experiments are executed on multiple datasets to verify the feasibilty of the proposed approach.

****

Transferring Inductive Biases through Knowledge Distillation
Samira Abnar,Mostafa Dehghani,Willem H. Zuidema

Having the right inductive biases can be crucial in many tasks or scenarios where data or computing resources are a limiting factor, or where training data is not perfectly representative of the conditions at test time. However, defining, designing, and efficiently adapting inductive biases is not necessarily straightforward. Inductive biases of a model affect its generalisation behaviour and influence the solution it converges to from different aspects. In this paper, we investigate the power of knowledge distillation in transferring the effects of inductive biases of a teacher model to a student model, when they have different architectures.

We consider different families of models: LSTMs vs. Transformers and CNNs vs. MLPs, in the context of tasks and scenarios with linguistics and vision applications, where having the right inductive biases is critical. We train our models in different setups: no knowledge distillation, self-distillation, and distillation

using a teacher with a better inductive bias for the task at hand. We show that in the later setup, compared to no distillation and self-distillation, we can not only improve the performance of the students, but also the solutions they converge become similar to their teachers with respect to a wide range of properties, including different task-specific performance metrics, per sample behaviour of the models, representational similarity and how the representational space of the models evolve during training, performance on out-of-distribution datasets, confidence calibration, and finally whether the converged solutions fall within the same basins of attractions.
****
Identifying Physical Law of Hamiltonian Systems via Meta-Learning
Seungjun Lee,Haesang Yang,Woojae Seong
Hamiltonian mechanics is an effective tool to represent many physical processes with concise yet well-generalized mathematical expressions. A well-modeled Hamiltonian makes it easy for researchers to analyze and forecast many related phenomena that are governed by the same physical law. However, in general, identifying a functional or shared expression of the Hamiltonian is very difficult. It requires carefully designed experiments and the researcher's insight that comes from years of experience. We propose that meta-learning algorithms can be potentially powerful data-driven tools for identifying the physical law governing Hamiltonian systems without any mathematical assumptions on the representation, but with observations from a set of systems governed by the same physical law. We show that a well meta-trained learner can identify the shared representation of the Hamiltonian by evaluating our method on several types of physical systems with various experimental settings.
****
Adapting to Reward Progressivity via Spectral Reinforcement Learning
Michael Dann,John Thangarajah
In this paper we consider reinforcement learning tasks with progressive rewards; that is, tasks where the rewards tend to increase in magnitude over time. We hypothesise that this property may be problematic for value-based deep reinforcement learning agents, particularly if the agent must first succeed in relatively unrewarding regions of the task in order to reach more rewarding regions. To address this issue, we propose Spectral DQN, which decomposes the reward into frequencies such that the high frequencies only activate when large rewards are found. This allows the training loss to be balanced so that it gives more even weighting across small and large reward regions. In two domains with extreme reward progressivity, where standard value-based methods struggle significantly, Spectral DQN is able to make much farther progress. Moreover, when evaluated on a set of six standard Atari games that do not overtly favour the approach, Spectral DQN remains more than competitive: While it underperforms one of the benchmarks in a single game, it comfortably surpasses the benchmarks in three games. These results demonstrate that the approach is not overfit to its target problem, and suggest that Spectral DQN may have advantages beyond addressing reward progressivity.
****
Understanding the effects of data parallelism and sparsity on neural network training
Namhoon Lee,Thalaiyasingam Ajanthan,Philip Torr,Martin Jaggi
We study two factors in neural network training: data parallelism and sparsity; here, data parallelism means processing training data in parallel using distributed systems (or equivalently increasing batch size), so that training can be accelerated; for sparsity, we refer to pruning parameters in a neural network model, so as to reduce computational and memory cost. Despite their promising benefits, however, understanding of their effects on neural network training remains elusive. In this work, we first measure these effects rigorously by conducting extensive experiments while tuning all metaparameters involved in the optimization. As a result, we find across various workloads of data set, network model, and optimization algorithm that there exists a general scaling trend between batch size and number of training steps to convergence for the effect of data parallelism, and further, difficulty of training under sparsity. Then, we develop a theore

tical analysis based on the convergence properties of stochastic gradient method s and smoothness of the optimization landscape, which illustrates the observed p henomena precisely and generally, establishing a better account of the effects o f data parallelism and sparsity on neural network training.
****

Rethinking Uncertainty in Deep Learning: Whether and How it Improves Robustness
Yilun Jin,Lixin Fan,Kam Woh Ng,Ce Ju,Qiang Yang
Deep neural networks (DNNs) are known to be prone to adversarial attacks, for wh ich many remedies are proposed. While adversarial training (AT) is regarded as t he most robust defense, it suffers from poor performance both on clean examples and under other types of attacks, e.g. attacks with larger perturbations. Meanwh ile, regularizers that encourage uncertain outputs, such as entropy maximization (EntM) and label smoothing (LS) can maintain accuracy on clean examples and imp rove performance under weak attacks, yet their ability to defend against strong attacks is still in doubt. In this paper, we revisit uncertainty promotion regul arizers, including EntM and LS, in the field of adversarial learning. We show th at EntM and LS alone provide robustness only under small perturbations. Contrari ly, we show that uncertainty promotion regularizers complement AT in a principle d manner, consistently improving performance on both clean examples and under va rious attacks, especially attacks with large perturbations. We further analyze h ow uncertainty promotion regularizers enhance the performance of AT from the per spective of Jacobian matrices $\nabla_X f(X;\theta)$, and find out that EntM eff ectively shrinks the norm of Jacobian matrices and hence promotes robustness.
****

Primal Wasserstein Imitation Learning
Robert Dadashi,Leonard Hussenot,Matthieu Geist,Olivier Pietquin
Imitation Learning (IL) methods seek to match the behavior of an agent with that of an expert. In the present work, we propose a new IL method based on a concep tually simple algorithm: Primal Wasserstein Imitation Learning (PWIL), which tie s to the primal form of the Wasserstein distance between the expert and the agen t state-action distributions. We present a reward function which is derived offl ine, as opposed to recent adversarial IL algorithms that learn a reward function through interactions with the environment, and which requires little fine-tunin g. We show that we can recover expert behavior on a variety of continuous contro l tasks of the MuJoCo domain in a sample efficient manner in terms of agent inte ractions and of expert interactions with the environment. Finally, we show that the behavior of the agent we train matches the behavior of the expert with the W asserstein distance, rather than the commonly used proxy of performance.
****

Matrix Data Deep Decoder - Geometric Learning for Structured Data Completion
Maria Schmidt,Alexander Bronstein
In this work, we present a fully convolutional end to end method to reconstruct corrupted sparse matrices of Non-Euclidean data. The classic example for such ma trices is recommender systems matrices where the rows/columns represent items/us ers and the entries are ratings. The method we present is inspired by the surpri sing and spectacular success of methods like$"$ deep image prior$"$ and $``$de ep decoder$"$ for corrupted image completion. In sharp contrast to previous Mat rix Completion methods wherein the latent matrix or its factors directly serve a s the optimization variable, in the method we present, the matrix is parameteriz ed as the weights of a graph neural network acting on a random noisy input. Then we are tuning the network parameters to get a result as close as possible to th e initial sparse matrix (using its factors) getting that way state of the art ma trix completion result. In addition to the conceptual simplicity of our method, which is just Non-Euclidean generalization of deep image priors, it holds fewer parameters than previously presented methods which makes the parameters more tra ckable and the method more computationally efficient and more applicable for the real-world tasks. The method also achieves state-of-the-art results for the mat rix completion task on the classical benchmarks in the field. The method also su rprisingly shows that untrained convolutional neural network can use a good prio r not only for image completion but also for Matrix Completion when redefined fo

r graphs.
****
Improving Post Training Neural Quantization: Layer-wise Calibration and Integer Programming
Itay Hubara,Yury Nahshan,Yair Hanani,Ron Banner,Daniel Soudry
Lately, post-training quantization methods have gained considerable attention, as they are simple to use, and require only a small unlabeled calibration set. This small dataset cannot be used to fine-tune the model without significant over-fitting. Instead, these methods only use the calibration set to set the activations' dynamic ranges. However, such methods always resulted in significant accuracy degradation, when used below 8-bits (except on small datasets). Here we aim to break the 8-bit barrier. To this end, we minimize the quantization errors of each layer separately by optimizing its parameters over the calibration set. We empirically demonstrate that this approach is: (1) much less susceptible to over-fitting than the standard fine-tuning approaches, and can be used even on a very small calibration set; and (2) more powerful than previous methods, which only set the activations' dynamic ranges. Furthermore, we demonstrate how to optimally allocate the bit-widths for each layer, while constraining accuracy degradation or model compression by proposing a novel integer programming formulation. Finally, we suggest model global statistics tuning, to correct biases introduced during quantization. Together, these methods yield state-of-the-art results for both vision and text models. For instance, on ResNet50, we obtain less than 1\% accuracy degradation --- with 4-bit weights and activations in all layers, but the smallest two. Our code is available at, https://github.com/papers-submission/CalibTIP
****
A frequency domain analysis of gradient-based adversarial examples
Bochen Lv,Pu Yang,Zehao Wang,Zhanxing Zhu
It is well known that deep neural networks are vulnerable to adversarial examples. We attempt to understand adversarial examples from the perspective of frequency analysis. Several works have empirically shown that the gradient-based adversarial attacks perform differently in the low-frequency and high-frequency part of the input data. But there is still a lack of theoretical justification of these phenomena. In this work, we both theoretically and empirically show that the adversarial perturbations gradually increase the concentration in the low-frequency domain of the spectrum during the training process of the model parameters. And the log-spectrum difference of the adversarial examples and clean image is more concentrated in the high-frequency part than the low-frequency part. We also find out that the ratio of the high-frequency and the low-frequency part in the adversarial perturbation is much larger than that in the corresponding natural image. Inspired by these important theoretical findings, we apply low-pass filter to potential adversarial examples before feeding them to the model. The results show that this preprocessing can significantly improve the robustness of the model.
****
On Noise Injection in Generative Adversarial Networks
Ruili Feng,Deli Zhao,Zheng-Jun Zha
Noise injection is an effective way of circumventing overfitting and enhancing generalization in machine learning, the rationale of which has been validated in deep learning as well. Recently, noise injection exhibits surprising performance when
  generating high-fidelity images in Generative Adversarial Networks (GANs). Despite its successful applications in GANs, the mechanism of its validity is still unclear. In this paper, we propose a geometric framework to theoretically analyze the role of noise injection in GANs. Based on Riemannian geometry, we successfully model the noise injection framework as fuzzy equivalence on geodesic normal coordinates. Guided by our theories, we find that existing methods are incomplete and a new strategy for noise injection is devised. Experiments on image generation and GAN inversion demonstrate the superiority of our method.
****

Optimizing Large-Scale Hyperparameters via Automated Learning Algorithm
Bin Gu,Guodong Liu,Yanfu Zhang,Xiang Geng,Heng Huang
Modern machine learning algorithms usually involve tuning multiple (from one to thousands) hyperparameters which play a pivotal role in terms of model generalizability. Globally choosing appropriate values of hyperparameters is extremely computationally challenging. Black-box optimization and gradient-based algorithms are two dominant approaches to hyperparameter optimization while they have totally distinct advantages. How to design a new hyperparameter optimization technique inheriting all benefits from both approaches is still an open problem. To address this challenging problem, in this paper, we propose a new hyperparameter optimization method with zeroth-order hyper-gradients (HOZOG). Specifically, we first exactly formulate hyperparameter optimization as an $\mathcal{A}$-based constrained optimization problem, where $\mathcal{A}$ is a black-box optimization algorithm (such as deep neural network). Then, we use the average zeroth-order hyper-gradients to update hyperparameters. We provide the feasibility analysis of using HOZOG to achieve hyperparameter optimization. The experimental results on three representative hyperparameter (the size is from 1 to 1250) optimization tasks demonstrate the benefits of HOZOG in terms of \textit{simplicity, scalability, flexibility, effectiveness and efficiency} compared with the state-of-the-art hyperparameter optimization methods.
****
Prediction and generalisation over directed actions by grid cells
Changmin Yu,Timothy Behrens,Neil Burgess
Knowing how the effects of directed actions generalise to new situations (e.g. moving North, South, East and West, or turning left, right, etc.) is key to rapid generalisation across new situations. Markovian tasks can be characterised by a state space and a transition matrix and recent work has proposed that neural grid codes provide an efficient representation of the state space, as eigenvectors of a transition matrix reflecting diffusion across states, that allows efficient prediction of future state distributions. Here we extend the eigenbasis prediction model, utilising tools from Fourier analysis, to prediction over arbitrary translation-invariant directed transition structures (i.e. displacement and diffusion), showing that a single set of eigenvectors can support predictions over arbitrary directed actions via action-specific eigenvalues. We show how to define a "sense of direction" to combine actions to reach a target state (ignoring task-specific deviations from translation-invariance), and demonstrate that adding the Fourier representations to a deep Q network aids policy learning in continuous control tasks. We show the equivalence between the generalised prediction framework and traditional models of grid cell firing driven by self-motion to perform path integration, either using oscillatory interference (via Fourier components as velocity-controlled oscillators) or continuous attractor networks (via analysis of the update dynamics). We thus provide a unifying framework for the role of the grid system in predictive planning, sense of direction and path integration: supporting generalisable inference over directed actions across different tasks.
****
Drop-Bottleneck: Learning Discrete Compressed Representation for Noise-Robust Exploration
Jaekyeom Kim,Minjung Kim,Dongyeon Woo,Gunhee Kim
We propose a novel information bottleneck (IB) method named Drop-Bottleneck, which discretely drops features that are irrelevant to the target variable. Drop-Bottleneck not only enjoys a simple and tractable compression objective but also additionally provides a deterministic compressed representation of the input variable, which is useful for inference tasks that require consistent representation. Moreover, it can jointly learn a feature extractor and select features considering each feature dimension's relevance to the target task, which is unattainable by most neural network-based IB methods. We propose an exploration method based on Drop-Bottleneck for reinforcement learning tasks. In a multitude of noisy and reward sparse maze navigation tasks in VizDoom (Kempka et al., 2016) and DMLab (Beattie et al., 2016), our exploration method achieves state-of-the-art perfo

rmance. As a new IB framework, we demonstrate that Drop-Bottleneck outperforms V ariational Information Bottleneck (VIB) (Alemi et al., 2017) in multiple aspects including adversarial robustness and dimensionality reduction.
****

PIVEN: A Deep Neural Network for Prediction Intervals with Specific Value Predic tion

Eli Simhayev,Gilad Katz,Lior Rokach

Improving the robustness of neural nets in regression tasks is key to their appl ication in multiple domains. Deep learning-based approaches aim to achieve this goal either by improving their prediction of specific values (i.e., point predic tion), or by producing prediction intervals (PIs) that quantify uncertainty. We present PIVEN, a deep neural network for producing both a PI and a prediction o f specific values. Unlike previous studies, PIVEN makes no assumptions regarding data distribution inside the PI, making its point prediction more effective for various real-world problems. Benchmark experiments show that our approach produ ces tighter uncertainty bounds than the current state-of-the-art approach for pr oducing PIs, while maintaining comparable performance to the state-of-the-art ap proach for specific value-prediction. Additional evaluation on large image datas ets further support our conclusions.
****

Variational Invariant Learning for Bayesian Domain Generalization

Zehao Xiao,Jiayi Shen,Xiantong Zhen,Ling Shao,Cees G. M. Snoek

Domain generalization addresses the out-of-distribution problem, which is challe nging due to the domain shift and the uncertainty caused by the inaccessibility to data from the target domains. In this paper, we propose variational invariant learning, a probabilistic inference framework that jointly models domain invari ance and uncertainty. We introduce variational Bayesian approximation into both the feature representation and classifier layers to facilitate invariant learnin g for better generalization across domains. In the probabilistic modeling framew ork, we introduce a domain-invariant principle to explore invariance across doma ins in a unified way. We incorporate the principle into the variational Bayesian layers in neural networks, achieving domain-invariant representations and class ifier. We empirically demonstrate the effectiveness of our proposal on four wide ly used cross-domain visual recognition benchmarks. Ablation studies demonstrate the benefits of our proposal and on all benchmarks our variational invariant le arning consistently delivers state-of-the-art performance.
****

BOIL: Towards Representation Change for Few-shot Learning

Jaehoon Oh,Hyungjun Yoo,ChangHwan Kim,Se-Young Yun

Model Agnostic Meta-Learning (MAML) is one of the most representative of gradien t-based meta-learning algorithms. MAML learns new tasks with a few data samples using inner updates from a meta-initialization point and learns the meta-initial ization parameters with outer updates. It has recently been hypothesized that re presentation reuse, which makes little change in efficient representations, is t he dominant factor in the performance of the meta-initialized model through MAML in contrast to representation change, which causes a significant change in repr esentations. In this study, we investigate the necessity of representation chang e for the ultimate goal of few-shot learning, which is solving domain-agnostic t asks. To this aim, we propose a novel meta-learning algorithm, called BOIL (Body Only update in Inner Loop), which updates only the body (extractor) of the mode l and freezes the head (classifier) during inner loop updates. BOIL leverages re presentation change rather than representation reuse. A frozen head cannot achie ve better results than even a random guessing classifier at the initial point of new tasks, and feature vectors (representations) have to move quickly to their corresponding frozen head vectors. We visualize this property using cosine simil arity, CKA, and empirical results without the head. Although the inner loop upda tes purely hinge on representation change, BOIL empirically shows significant pe rformance improvement over MAML, particularly on cross-domain tasks. The results imply that representation change in gradient-based meta-learning approaches is a critical component.

****
Exploring the Potential of Low-Bit Training of Convolutional Neural Networks
Kai Zhong,Xuefei Ning,Tianchen Zhao,Zhenhua Zhu,Shulin Zeng,Guohao Dai,Yu Wang,Huazhong Yang
In this paper, we propose a low-bit training framework for convolutional neural networks. Our framework focuses on reducing the energy and time consumption of convolution kernels, by quantizing all the convolutional operands (activation, weight, and error) to low bit-width. Specifically, we propose a multi-level scaling (MLS) tensor format, in which the element-wise bit-width can be largely reduced to simplify floating-point computations to nearly fixed-point. Then, we describe the dynamic quantization and the low-bit tensor convolution arithmetic to efficiently leverage the MLS tensor format. Experiments show that our framework achieves a superior trade-off between the accuracy and the bit-width than previous methods. When training ResNet-20 on CIFAR-10,
all convolution operands can be quantized to 1-bit mantissa and 2-bit exponent, while retaining the same accuracy as the full-precision training. When training ResNet-18 on ImageNet, with 4-bit mantissa and 2-bit exponent, our framework can achieve an accuracy loss of less than 1%. Energy consumption analysis shows that our design can achieve over $6.8\times$ higher energy efficiency than training with floating-point arithmetic.
****
MultiModalQA: complex question answering over text, tables and images
Alon Talmor,Ori Yoran,Amnon Catav,Dan Lahav,Yizhong Wang,Akari Asai,Gabriel Ilharco,Hannaneh Hajishirzi,Jonathan Berant
When answering complex questions, people can seamlessly combine information from visual, textual and tabular sources.
While interest in models that reason over multiple pieces of evidence has surged in recent years, there has been relatively little work on question answering models that reason across multiple modalities.
In this paper, we present MultiModalQA (MMQA): a challenging question answering dataset that requires joint reasoning over text, tables and images.
We create MMQA using a new framework for generating complex multi-modal questions at scale, harvesting tables from Wikipedia, and attaching images and text paragraphs using entities that appear in each table. We then define a formal language that allows us to take questions that can be answered from a single modality, and combine them to generate cross-modal questions. Last, crowdsourcing workers take these automatically generated questions and rephrase them into more fluent language.
We create 29,918 questions through this procedure, and empirically demonstrate the necessity of a multi-modal multi-hop approach to solve our task: our multi-hop model, ImplicitDecomp, achieves an average F1 of 51.7 over cross-modal questions, substantially outperforming a strong baseline that achieves 38.2 F1, but still lags significantly behind human performance, which is at 90.1 F1.
****
Interpretable Relational Representations for Food Ingredient Recommendation Systems
Kana Maruyama,Michael Spranger
Supporting chefs with ingredient recommender systems to create new recipes is challenging, as good ingredient combinations depend on many factors like taste, smell, cuisine style, texture among others. There have been few attempts to address these issues using machine   learning. Importantly, useful models do obviously need to be accurate but importantly -- especially for food professionals -- interpretable. In order to address these issues, we propose the Interpretable Relational Representation Model (IRRM). The main component of the model is a key-value memory network to represent relationships of ingredients. We propose and test two variants of the model.
One can learn latent relational representations over a trainable memory network (Implicit model), and the other can learn explainable relational representations over a pre-trained memory network that integrates an external knowledge base (Explicit model).

The relational representations resulting from the model are interpretable -- the
y allow to inspect why certain ingredient pairings have been suggested. The Expl
icit model additionally allows to integrate any number of manually specified con
straints.
We conduct experiments on two recipe datasets, including CulinaryDB with 45,772
recipes and Flavornet with 55,001 recipes, respectively. The experimental result
s show that our models are both predictive and informative.
****
Regularization Shortcomings for Continual Learning
Timothee LESORT,Andrei Stoian
In most machine learning algorithms, training data is assumed to be independent
and identically distributed (iid).
When it is not the case, the performances of the algorithms are challenged, lead
ing to the famous phenomenon of \textit{catastrophic forgetting}. Algorithms dea
ling with it are gathered in the \textit{Continual Learning} research field. In
this paper, we study the \textit{regularization} based approaches to continual l
earning and show that those approaches can not learn to discriminate classes fro
m different tasks in an elemental continual benchmark, the class-incremental set
ting.
We make theoretical reasoning to prove this shortcoming and illustrate it with e
xperiments.
Moreover, we show that it can have some important consequences on multi-tasks re
inforcement learning or in pre-trained models used for continual learning.
We believe this paper to be the first to propose a theoretical description of re
gularization shortcomings for continual learning.
****
Central Server Free Federated Learning over Single-sided Trust Social Networks
Chaoyang He,Conghui Tan,Hanlin Tang,Shuang Qiu,Ji Liu
Federated learning has become increasingly important for modern machine learning
, especially for data privacy-sensitive scenarios. Existing federated learning m
ostly adopts the central server-based architecture or centralized architecture.
However, in many social network scenarios, centralized federated learning is not
 applicable (e.g., a central agent or server connecting all users may not exist,
 or the communication cost to the central server is not affordable). In this pap
er, we consider a generic setting: 1) the central server may not exist, and 2) t
he social network is unidirectional or of single-sided trust (i.e., user A trust
s user B but user B may not trust user A). We propose a central server free fede
rated learning algorithm, named Online Push-Sum (OPS) method, to handle this cha
llenging but generic scenario. A rigorous regret analysis is also provided, whic
h shows interesting results on how users can benefit from communication with tru
sted users in the federated learning scenario. This work builds upon the fundame
ntal algorithm framework and theoretical guarantees for federated learning in th
e generic social network scenario.
****
On the Inductive Bias of a CNN for Distributions with Orthogonal Patterns
Alon Brutzkus,Amir Globerson
Training overparameterized convolutional neural networks with gradient based opt
imization is the most successful learning method for image classification. Howev
er, their generalization properties are far from understood. In this work, we co
nsider a simplified image classification task where images contain orthogonal pa
tches and are learned with a 3-layer overparameterized convolutional network and
 stochastic gradient descent (SGD). We empirically identify a novel phenomenon o
f SGD in our setting, where the dot-product between the learned pattern detector
s and their detected patterns are governed by the pattern statistics in the trai
ning set. We call this phenomenon Pattern Statistics Inductive Bias (PSI) and em
pirically verify it in a large number of instances. We prove that in our setting
, if a learning algorithm satisfies PSI then its sample complexity is $O(d^2\log
(d))$ where d is the filter dimension. In contrast, we show a VC dimension low
er bound which is exponential in d. We perform experiments with overparameteri
zed CNNs on a variant of MNIST with non-orthogonal patches, and show that the em

pirical observations are in line with our analysis.
****
Neural gradients are near-lognormal: improved quantized and sparse training
Brian Chmiel,Liad Ben-Uri,Moran Shkolnik,Elad Hoffer,Ron Banner,Daniel Soudry
While training can mostly be accelerated by reducing the time needed to propagate neural gradients (loss gradients with respect to the intermediate neural layer outputs) back throughout the model, most previous works focus on the quantization/pruning of weights and activations. These methods are often not applicable to neural gradients, which have very different statistical properties. Distinguished from weights and activations, we find that the distribution of neural gradients is approximately lognormal. Considering this, we suggest two closed-form analytical methods to reduce the computational and memory burdens of neural gradients. The first method optimizes the floating-point format and scale of the gradients. The second method accurately sets sparsity thresholds for gradient pruning. Each method achieves state-of-the-art results on ImageNet. To the best of our knowledge, this paper is the first to (1) quantize the gradients to 6-bit floating-point formats, or (2) achieve up to 85% gradient sparsity --- in each case without accuracy degradation.
Reference implementation accompanies the paper in the supplementary material.
****
Knowledge Distillation By Sparse Representation Matching
Dat Thanh Tran,Moncef Gabbouj,Alexandros Iosifidis
Knowledge Distillation refers to a class of methods that transfers the knowledge from a teacher network to a student network. In this paper, we propose Sparse Representation Matching (SRM), a method to transfer intermediate knowledge obtained from one Convolutional Neural Network (CNN) to another by utilizing sparse representation learning. SRM first extracts sparse representations of the hidden features of the teacher CNN, which are then used to generate both pixel-level and image-level labels for training intermediate feature maps of the student network. We formulate SRM as a neural processing block, which can be efficiently optimized using stochastic gradient descent and integrated into any CNN in a plug-and-play manner. Our experiments demonstrate that SRM is robust to architectural differences between the teacher and student networks, and outperforms other KD techniques across several datasets.
****
Group Equivariant Conditional Neural Processes
Makoto Kawano,Wataru Kumagai,Akiyoshi Sannai,Yusuke Iwasawa,Yutaka Matsuo
We present the group equivariant conditional neural process (EquivCNP), a meta-learning method with permutation invariance in a data set as in conventional conditional neural processes (CNPs), and it also has transformation equivariance in data space. Incorporating group equivariance, such as rotation and scaling equivariance, provides a way to consider the symmetry of real-world data. We give a decomposition theorem for permutation-invariant and group-equivariant maps, which leads us to construct EquivCNPs with an infinite-dimensional latent space to handle group symmetries. In this paper, we build architecture using Lie group convolutional layers for practical implementation. We show that EquivCNP with translation equivariance achieves comparable performance to conventional CNPs in a 1D regression task. Moreover, we demonstrate that incorporating an appropriate Lie group equivariance, EquivCNP is capable of zero-shot generalization for an image-completion task by selecting an appropriate Lie group equivariance.
****
K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters
Ruize Wang,Duyu Tang,Nan Duan,zhongyu wei,Xuanjing Huang,Jianshu Ji,Guihong Cao,Daxin Jiang,Ming Zhou
We study the problem of injecting knowledge into large pre-trained models like BERT and RoBERTa. Existing methods typically update the original parameters of pre-trained models when injecting knowledge. However, when multiple kinds of knowledge are injected, they may suffer from catastrophic forgetting. To address this, we propose K-Adapter, which remains the original parameters of the pre-trained model fixed and supports continual knowledge infusion. Taking RoBERTa as the p

re-trained model, K-Adapter has a neural adapter for each kind of infused knowle dge, like a plug-in connected to RoBERTa. There is no information flow between d ifferent adapters, thus different adapters are efficiently trained in a distribu ted way. We inject two kinds of knowledge, including factual knowledge obtained from automatically aligned text-triplets on Wikipedia and Wikidata, and linguist ic knowledge obtained from dependency parsing. Results on three knowledge-driven tasks (total six datasets) including relation classification, entity typing and question answering demonstrate that each adapter improves the performance, and the combination of both adapters brings further improvements. Probing experiment s further indicate that K-Adapter captures richer factual and commonsense knowle dge than RoBERTa.
****

Deployment-Efficient Reinforcement Learning via Model-Based Offline Optimization
Tatsuya Matsushima,Hiroki Furuta,Yutaka Matsuo,Ofir Nachum,Shixiang Gu
Most reinforcement learning (RL) algorithms assume online access to the environm ent, in which one may readily interleave updates to the policy with experience c ollection using that policy. However, in many real-world applications such as he alth, education, dialogue agents, and robotics, the cost or potential risk of de ploying a new data-collection policy is high, to the point that it can become pr ohibitive to update the data-collection policy more than a few times during lear ning. With this view, we propose a novel concept of deployment efficiency, measu ring the number of distinct data-collection policies that are used during policy learning. We observe that naïvely applying existing model-free offline RL algor ithms recursively does not lead to a practical deployment-efficient and sample-e fficient algorithm. We propose a novel model-based algorithm, Behavior-Regulariz ed Model-ENsemble (BREMEN), that not only performs better than or comparably as the state-of-the-art dynamic-programming-based and concurrently-proposed model-b ased offline approaches on existing benchmarks, but can also effectively optimiz e a policy offline using 10-20 times fewer data than prior works. Furthermore, t he recursive application of BREMEN achieves impressive deployment efficiency whi le maintaining the same or better sample efficiency, learning successful policie s from scratch on simulated robotic environments with only 5-10 deployments, com pared to typical values of hundreds to millions in standard RL baselines.
****

Fair Differential Privacy Can Mitigate the Disparate Impact on Model Accuracy
Wenyan Liu,Xiangfeng Wang,Xingjian Lu,Junhong Cheng,Bo Jin,Xiaoling Wang,Hongyua n Zha
The techniques based on the theory of differential privacy (DP) has become a sta ndard building block in the machine learning community. DP training mechanisms o ffer strong guarantees that an adversary cannot determine with high confidence a bout the training data based on analyzing the released model, let alone any deta ils of the instances. However, DP may disproportionately affect the underreprese nted and relatively complicated classes. That is, the reduction in utility is un equal for each class. This paper proposes a fair differential privacy algorithm (FairDP) to mitigate the disparate impact on model accuracy for each class. We c ast the learning procedure as a two-stage optimization problem, which integrates differential privacy with fairness. FairDP establishes a self-adaptive DP mecha nism and dynamically adjusts instance influence in each class depending on the t heoretical bias-variance bound. Our experimental evaluation shows the effectiven ess of FairDP in mitigating the disparate impact on model accuracy among the cla sses on several benchmark datasets and scenarios ranging from text to vision.
****

Efficient Conformal Prediction via Cascaded Inference with Expanded Admission
Adam Fisch,Tal Schuster,Tommi S. Jaakkola,Regina Barzilay
In this paper, we present a novel approach for conformal prediction (CP), in whi ch we aim to identify a set of promising prediction candidates---in place of a s ingle prediction. This set is guaranteed to contain a correct answer with high p robability, and is well-suited for many open-ended classification tasks. In the standard CP paradigm, the predicted set can often be unusably large and also cos tly to obtain. This is particularly pervasive in settings where the correct answ

er is not unique, and the number of total possible answers is high. We first exp and the CP correctness criterion to allow for additional, inferred "admissible" answers, which can substantially reduce the size of the predicted set while stil l providing valid performance guarantees. Second, we amortize costs by conformal izing prediction cascades, in which we aggressively prune implausible labels ear ly on by using progressively stronger classifiers---again, while still providing valid performance guarantees. We demonstrate the empirical effectiveness of our approach for multiple applications in natural language processing and computati onal chemistry for drug discovery.
****

Learned ISTA with Error-based Thresholding for Adaptive Sparse Coding
Li Ziang,Wu Kailun,Yiwen Guo,Changshui Zhang
The learned iterative shrinkage thresholding algorithm (LISTA) introduces deep u nfolding models with learnable thresholds in the shrinkage function for sparse c oding. Drawing on some theoretical insights, we advocate an error-based threshol ding (EBT) mechanism for LISTA, which leverages a function of the layer-wise rec onstruction error to suggest an appropriate threshold value for each observation on each layer. We show that the EBT mechanism well-disentangles the learnable p arameters in the shrinkage functions from the reconstruction errors, making them more adaptive to the various observations. With rigorous theoretical analyses, we show that the proposed EBT can lead to faster convergence on the basis of LIS TA and its variants, in addition to its higher adaptivity. Extensive experimenta l results confirm our theoretical analyses and verify the effectiveness of our m ethods.
****

Closing the Generalization Gap in One-Shot Object Detection
Claudio Michaelis,Matthias Bethge,Alexander S Ecker
Despite substantial progress in object detection and few-shot learning, detectin g objects based on a single example - one-shot object detection - remains a chal lenge. A central problem is the generalization gap: Object categories used durin g training are detected much more reliably than novel ones. We here show that th is generalization gap can be nearly closed by increasing the number of object ca tegories used during training. Doing so allows us to beat the state-of-the-art o n COCO by 5.4 %AP50 (from 22.0 to 27.5) and improve generalization from seen to unseen classes from 45% to 89%. We verify that the effect is caused by the numbe r of categories and not the amount of data and that it holds for different model s, backbones and datasets. This result suggests that the key to strong few-shot detection models may not lie in sophisticated metric learning approaches, but in stead 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 fut ure data annotation efforts to focus on wider datasets with a broader set of cat egories rather than gathering more samples per category.
****

Reducing the Computational Cost of Deep Generative Models with Binary Neural Net works
Thomas Bird,Friso Kingma,David Barber
Deep generative models provide a powerful set of tools to understand real-world data. But as these models improve, they increase in size and complexity, so thei r computational cost in memory and execution time grows. Using binary weights in neural networks is one method which has shown promise in reducing this cost. Ho wever, whether binary neural networks can be used in generative models is an ope n problem. In this work we show, for the first time, that we can successfully tr ain generative models which utilize binary neural networks. This reduces the com putational cost of the models massively. We develop a new class of binary weight normalization, and provide insights for architecture designs of these binarized generative models. We demonstrate that two state-of-the-art deep generative mod els, the ResNet VAE and Flow++ models, can be binarized effectively using these techniques. We train binary models that achieve loss values close to those of th e regular models but are 90%-94% smaller in size, and also allow significant spe ed-ups in execution time.

****
MixSize: Training Convnets With Mixed Image Sizes for Improved Accuracy, Speed a nd Scale Resiliency
Elad Hoffer,Berry Weinstein,Itay Hubara,Tal Ben-Nun,Torsten Hoefler,Daniel Soudr y
Convolutional neural networks (CNNs) are commonly trained using a fixed spatial image size predetermined for a given model. Although trained on images of a spec ific size, it is well established that CNNs can be used to evaluate a wide range of image sizes at test time, by adjusting the size of intermediate feature maps . \\
In this work, we describe and evaluate a novel mixed-size training regime that m ixes several image sizes at training time. We demonstrate that models trained us ing our method are more resilient to image size changes and generalize well even on small images. This allows faster inference by using smaller images at test t ime. For instance, we receive a 76.43% top-1 accuracy using ResNet50 with an image size of 160, which matches the accuracy of the baseline model with $2 \t imes$ fewer computations.
Furthermore, for a given image size used at test time, we show this method can b e exploited either to accelerate training or the final test accuracy. For exampl e, we are able to reach a 79.27% accuracy with a model evaluated at a 288 s patial size for a relative improvement of 14% over the baseline.
Our PyTorch implementation and pre-trained models are publicly available\footnot e{\url{https://github.com/paper-submissions/mix-match}
****
Towards Data Distillation for End-to-end Spoken Conversational Question Answerin g
Chenyu You,Nuo Chen,Fenglin Liu,Dongchao Yang,Zhiyang Xu,Yuexian Zou
In spoken question answering, QA systems are designed to answer questions from c ontiguous text spans within the related speech transcripts. However, the most na tural way that human seek or test their knowledge is via human conversations. Th erefore, we propose a new Spoken Conversational Question Answering task (SCQA), aiming at enabling QA systems to model complex dialogues flow given the speech u tterances and text corpora. In this task, our main objective is to build a QA sy stem to deal with conversational questions both in spoken and text forms, and to explore the plausibility of providing more cues in spoken documents with system s in information gathering. To this end, instead of adopting automatically gener ated speech transcripts with highly noisy data, we propose a novel unified data distillation approach, DDNet, which directly fuse audio-text features to reduce the misalignment between automatic speech recognition hypotheses and the referen ce transcriptions. In addition, to evaluate the capacity of QA systems in a dial ogue-style interaction, we assemble a Spoken Conversational Question Answering (Spoken-CoQA) dataset with more than 120k question-answer pairs. Experiments demo nstrate that our proposed method achieves superior performance in spoken convers ational question answering.

****
Structure Controllable Text Generation
Liming DENG,Long WANG,Binzhu WANG,Jiang Qian,Bojin Zhuang,Shaojun Wang,Jing Xiao
Controlling the presented forms (or structures) of generated text are as importa nt as controlling the generated contents during neural text generation. It helps to reduce the uncertainty and improve the interpretability of generated text. H owever, the structures and contents are entangled together and realized simultan eously during text generation, which is challenging for the structure controllin g. In this paper, we propose an efficient, straightforward generation framework to control the structure of generated text. A structure-aware transformer (SAT) is proposed to explicitly incorporate multiple types of multi-granularity struct ure information to guide the text generation with corresponding structure. The s tructure information is extracted from given sequence template by auxiliary mode l, and the type of structure for the given template can be learned, represented and imitated. Extensive experiments have been conducted on both Chinese lyrics c

orpus and English Penn Treebank dataset. Both automatic evaluation metrics and h
uman judgement demonstrate the superior capability of our model in controlling t
he structure of generated text, and the quality (like Fluency and Meaningfulnes
s) of the generated text is even better than the state-of-the-arts model.

****
Inductive Bias of Gradient Descent for Exponentially Weight Normalized Smooth Ho
mogeneous Neural Nets
Depen Morwani,Harish Guruprasad Ramaswamy
We analyze the inductive bias of gradient descent for weight normalized smooth h
omogeneous neural nets, when trained on exponential or cross-entropy loss. Our a
nalysis focuses on exponential weight normalization (EWN), which encourages weig
ht updates along the radial direction. This paper shows that the gradient flow p
ath with EWN is equivalent to gradient flow on standard networks with an adaptiv
e learning rate, and hence causes the weights to be updated in a way that prefer
s asymptotic relative sparsity. These results can be extended to hold for gradie
nt descent via an appropriate adaptive learning rate. The asymptotic convergence
 rate of the loss in this setting is given by $\Theta(\frac{1}{t(\log t)^2})$, a
nd is independent of the depth of the network. We contrast these results with th
e inductive bias of standard weight normalization (SWN) and unnormalized archite
ctures, and demonstrate their implications on synthetic data sets.Experimental r
esults on simple data sets and architectures support our claim on sparse EWN sol
utions, even with SGD. This demonstrates its potential applications in learning
prunable neural networks.
****
On the Universal Approximability and Complexity Bounds of Deep Learning in Hybri
d Quantum-Classical Computing
Weiwen Jiang,Yukun Ding,Yiyu Shi
With the continuously increasing number of quantum bits in quantum computers, th
ere are growing interests in exploring applications that can harvest the power o
f them. Recently, several attempts were made to implement neural networks, known
 to be computationally intensive, in hybrid quantum-classical scheme computing.
While encouraging results are shown, two fundamental questions need to be answer
ed: (1) whether neural networks in hybrid quantum-classical computing can levera
ge quantum power and meanwhile approximate any function within a given error bou
nd, i.e., universal approximability; (2) how do these neural networks compare wi
th ones on a classical computer in terms of representation power? This work shed
s light on these two questions from a theoretical perspective.
****
Minimal Geometry-Distortion Constraint for Unsupervised Image-to-Image Translati
on
Jiaxian Guo,Jiachen Li,Mingming Gong,Huan Fu,Kun Zhang,Dacheng Tao
Unsupervised image-to-image (I2I) translation, which aims to learn a domain mapp
ing function without paired data, is very challenging because the function is hi
ghly under-constrained. Despite the significant progress in constraining the map
ping function, current methods suffer from the \textit{geometry distortion} prob
lem: the geometry structure of the translated image is inconsistent with the inp
ut source image, which may cause the undesired distortions in the translated ima
ges. To remedy this issue, we propose a novel I2I translation constraint, call
ed \textit{Minimal Geometry-Distortion Constraint} (MGC), which promotes the con
sistency of geometry structures and reduce the unwanted distortions in translati
on by reducing the randomness of color transformation in the translation process
. To facilitate estimation and maximization of MGC, we propose an approximate re
presentation of mutual information called relative Squared-loss Mutual Informati
on (rSMI) that can be efficiently estimated analytically. We demonstrate the eff
ectiveness of our MGC by providing quantitative and qualitative comparisons with
 the state-of-the-art methods on several benchmark datasets.

****
MCM-aware Twin-least-square GAN for Hyperspectral Anomaly Detection

Jiaping Zhong,Weiying Xie,Jie Lei,Yunsong Li,Zan Li
Hyperspectral anomaly detection under high-dimensional data and interference of deteriorated bands without any prior information has been challenging and attracted close attention in the exploration of the unknown in real scenarios. However, some emerging methods based on generative adversarial network (GAN) suffer from the problems of gradient vanishing and training instability with struggling to strike a balance between performance and training sample limitations. In this work, aiming to remedy the drawbacks of existing methods, we present a novel multi-scale covariance map (MCM)-aware twin-least-square GAN (MTGAN). Instead of the widely used single-scale Gaussian hypothesis background estimation, in MTGAN, we introduce the MCM-aware strategy to construct multi-scale priors with precise second-order statistics, thereby implicitly bridging the spatial and spectral information. Thus, we reliably and adaptively represent the prior of HSI to change the priors-lack situation. Moreover, we impose the twin-least-square loss on GAN, which helps improve the generative ability and training stability in feature and image domains, overcoming the gradient vanishing problem. Finally, the network enforced with a new anomaly rejection loss establishes a pure and discriminative background estimation. Experiments demonstrate that the average detection accuracy of MTGAN reaches 0.99809, which is superior to the state-of-the-art algorithms.
****
Switching-Aligned-Words Data Augmentation for Neural Machine Translation
Fengshun Xiao,Zuchao Li,hai zhao
In neural machine translation (NMT), data augmentation methods such as back-translation make it possible to use extra monolingual data to help improve translation performance, while it needs extra training data and the in-domain monolingual data is not always available. In this paper, we present a novel data augmentation method for neural machine translation by using only the original training data without extra data. More accurately, we randomly replace words or mixup with their aligned alternatives in another language when training neural machine translation models. Since aligned word pairs appear in the same position of each other during training, it is helpful to form bilingual embeddings which are proved useful to provide a performance boost \citep{liu2019shared}. Experiments on both small and large scale datasets show that our method significantly outperforms the baseline models.
****
To Understand Representation of Layer-aware Sequence Encoders as Multi-order-graph
Sufeng Duan,hai zhao,Rui Wang
In this paper, we propose a unified explanation of representation for layer-aware neural sequence encoders, which regards the representation as a revisited multigraph called multi-order-graph (MoG), so that model encoding can be viewed as a processing to capture all subgraphs in MoG. The relationship reflected by Multi-order-graph, called n-order dependency, can present what existing simple directed graph explanation cannot present. Our proposed MoG explanation allows to precisely observe every step of the generation of representation, put diverse relationship such as syntax into a unifiedly depicted framework. Based on the proposed MoG explanation, we further propose a graph-based self-attention network empowered Graph-Transformer by enhancing the ability of capturing subgraph information over the current models. Graph-Transformer accommodates different subgraphs into different groups, which allows model to focus on salient subgraphs. Result of experiments on neural machine translation tasks show that the MoG-inspired model can yield effective performance improvement.
****
A Text GAN for Language Generation with Non-Autoregressive Generator
Fei Huang,Jian Guan,Pei Ke,Qihan Guo,Xiaoyan Zhu,Minlie Huang
Despite the great success of Generative Adversarial Networks (GANs) in generating high-quality images, GANs for text generation still face two major challenges: first, most text GANs are unstable in training mainly due to ineffective optimization of the generator, and they heavily rely on maximum likelihood pretraining

; second, most text GANs adopt autoregressive generators without latent variable s, which largely limits the ability to learn latent representations for natural language text. In this paper, we propose a novel text GAN, named NAGAN, which in corporates a non-autoregressive generator with latent variables. The non-autoreg ressive generator can be effectively trained with gradient-based methods and fre e of pretraining. The latent variables facilitate representation learning for te xt generation applications. Experiments show that our model is competitive compa ring with existing text GANs in unconditional text generation, and it outperform s existing methods on sentence manipulation in latent space and unsupervised tex t decipherment.
****

Large-width functional asymptotics for deep Gaussian neural networks
Daniele Bracale,Stefano Favaro,Sandra Fortini,Stefano Peluchetti
In this paper, we consider fully connected feed-forward deep neural networks whe re weights and biases are independent and identically distributed according to G aussian distributions. Extending previous results (Matthews et al., 2018a;b;Yang , 2019) we adopt a function-space perspective, i.e. we look at neural networks as infinite-dimensional random elements on the input space $\mathbb{R}^I$. Under suitable assumptions on the activation function we show that: i) a network defi nes a continuous Gaussian process on the input space $\mathbb{R}^I$; ii) a netwo rk with re-scaled weights converges weakly to a continuous Gaussian process in t he large-width limit; iii) the limiting Gaussian process has almost surely local ly γ-Hölder continuous paths, for $0 < \gamma <1$. Our results contribute to recent theoretical studies on the interplay between infinitely wide deep neu ral networks and Gaussian processes by establishing weak convergence in function -space with respect to a stronger metric.
****

Unbiased Learning with State-Conditioned Rewards in Adversarial Imitation Learni ng
Dong-Sig Han,Hyunseo Kim,Hyundo Lee,Je-Hwan Ryu,Byoung-Tak Zhang
Adversarial imitation learning has emerged as a general and scalable framework f or automatic reward acquisition. However, we point out that previous methods com monly exploited occupancy-dependent reward learning formulation—which hinders th e reconstruction of optimal decision as an energy-based model. Despite the theor etical justification, the occupancy measures tend to cause issues in practice be cause of high variance and low vulnerability to domain shifts. Another reported problem is termination biases induced by provided rewarding and regularization s chemes around terminal states. In order to deal with these issues, this work pre sents a novel algorithm called causal adversarial inverse reinforcement learning . Our formulation draws a strong connection between adversarial learning and ene rgy-based reinforcement learning; thus, the architecture is capable of recoverin g a reward function that induces a multi-modal policy. In experiments, we demons trate that our approach outperforms prior methods in challenging continuous cont rol tasks, even under significant variation in the environments.
****

MetaNorm: Learning to Normalize Few-Shot Batches Across Domains
Yingjun Du,Xiantong Zhen,Ling Shao,Cees G. M. Snoek
Batch normalization plays a crucial role when training deep neural networks. How ever, batch statistics become unstable with small batch sizes and are unreliable in the presence of distribution shifts. We propose MetaNorm, a simple yet effec tive meta-learning normalization. It tackles the aforementioned issues in a unif ied way by leveraging the meta-learning setting and learns to infer adaptive sta tistics for batch normalization. MetaNorm is generic, flexible and model-agnosti c, making it a simple plug-and-play module that is seamlessly embedded into exis ting meta-learning approaches. It can be efficiently implemented by lightweight hypernetworks with low computational cost. We verify its effectiveness by extens ive evaluation on representative tasks suffering from the small batch and domain shift problems: few-shot learning and domain generalization. We further introdu ce an even more challenging setting: few-shot domain generalization. Results dem onstrate that MetaNorm consistently achieves better, or at least competitive, ac

curacy compared to existing batch normalization methods.
****
Deep Networks from the Principle of Rate Reduction
Kwan Ho Ryan Chan,Yaodong Yu,Chong You,Haozhi Qi,John Wright,Yi Ma
This work attempts to interpret modern deep (convolutional) networks from the pr
inciples of rate reduction and (shift) invariant classification. We show that th
e basic iterative gradient ascent scheme for maximizing the rate reduction of le
arned features naturally leads to a deep network, one iteration per layer. The a
rchitectures, operators (linear or nonlinear), and parameters of the network are
 all explicitly constructed layer-by-layer in a forward propagation fashion. All
 components of this ``white box'' network have precise optimization, statistical
, and geometric interpretation. Our preliminary experiments indicate that such a
 network can already learn a good discriminative deep representation without any
 back propagation training. Moreover, all linear operators of the so-derived net
work naturally become multi-channel convolutions when we enforce classification
 to be rigorously shift-invariant. The derivation also indicates that such a conv
olutional network is significantly more efficient to learn and construct in the
spectral domain.
****
Non-decreasing Quantile Function Network with Efficient Exploration for Distribu
tional Reinforcement Learning
Fan Zhou,Zhoufan Zhu,Qi Kuang,Liwen Zhang
Although distributional reinforcement learning (DRL) has been widely examined in
 the past few years, there are two open questions people are still trying to add
ress. One is how to ensure the validity of the learned quantile function, the ot
her is how to efficiently utilize the distribution information. This paper attem
pts to provide some new perspectives to encourage the future in-depth studies in
 these two fields. We first propose a non-decreasing quantile function network (
NDQFN) to guarantee the monotonicity of the obtained quantile estimates and then
 design a general exploration framework called distributional prediction error (
DPE) for DRL which utilizes the entire distribution of the quantile function. In
 this paper, we not only discuss the theoretical necessity of our method but als
o show the performance gain it achieves in practice by comparing with some compe
titors on Atari 2600 Games especially in some hard-explored games.
****
Explicit homography estimation improves contrastive self-supervised learning
David Torpey,Richard Klein
The typical contrastive self-supervised algorithm uses a similarity measure in l
atent space as the supervision signal by contrasting positive and negative image
s directly or indirectly. Although the utility of self-supervised algorithms has
 improved recently, there are still bottlenecks hindering their widespread use,
such as the compute needed. In this paper, we propose a module that serves as an
 additional objective in the self-supervised contrastive learning paradigm. We s
how how the inclusion of this module to regress the parameters of an affine tran
sformation or homography, in addition to the original contrastive objective, imp
roves both performance and rate of learning. Importantly, we ensure that this mo
dule does not enforce invariance to the various components of the affine transfo
rm, as this is not always ideal. We demonstrate the effectiveness of the additio
nal objective on two recent, popular self-supervised algorithms. We perform an e
xtensive experimental analysis of the proposed method and show an improvement in
 performance for all considered datasets. Further, we find that although both th
e general homography and affine transformation are sufficient to improve perform
ance and convergence, the affine transformation performs better in all cases.
****
Variational Multi-Task Learning
Jiayi Shen,Xiantong Zhen,Marcel Worring,Ling Shao
Multi-task learning aims to improve the overall performance of a set of tasks by
 leveraging their relatedness. When training data is limited using priors is piv
otal, but currently this is done in ad-hoc ways. In this paper, we develop varia
tional multi-task learning - VMTL, a general probabilistic inference framework f

or simultaneously learning multiple related tasks. We cast multi-task learning as a variational Bayesian inference problem, which enables task relatedness to be explored in a principled way by specifying priors. We introduce Gumbel-softmax priors to condition the prior of each task on related tasks. Each prior is represented as a mixture of variational posteriors of other related tasks and the mixing weights are learned in a data-driven manner for each individual task. The posteriors over representations and classifiers are inferred jointly for all tasks and individual tasks are able to improve their performance by using the shared inductive bias. Experimental results demonstrate that VMTL is able to tackle challenging multi-task learning with limited training data well, and it achieves state-of-the-art performance on four benchmarks, consistently surpassing previous methods.
****
Probabilistic Mixture-of-Experts for Efficient Deep Reinforcement Learning
Jie Ren,Yewen Li,Zihan Ding,Wei Pan,Hao Dong
Deep reinforcement learning (DRL) has successfully solved various problems recently, typically with a unimodal policy representation. However, grasping the decomposable and hierarchical structures within a complex task can be essential for further improving its learning efficiency and performance, which may lead to a multimodal policy or a mixture-of-experts (MOE). To our best knowledge, present DRL algorithms for general utility do not deploy MOE methods as policy function approximators due to the lack of differentiability, or without explicit probabilistic representation. In this work, we propose a differentiable probabilistic mixture-of-experts (PMOE) embedded in the end-to-end training scheme for generic off-policy and on-policy algorithms using stochastic policies, e.g., Soft Actor-Critic (SAC) and Proximal Policy Optimisation (PPO). Experimental results testify the advantageous performance of our method over unimodal polices and three different MOE methods, as well as a method of option frameworks, based on two types of DRL algorithms. We also demonstrate the distinguishable primitives learned with PMOE in different environments.
****
Representation Balancing Offline Model-based Reinforcement Learning
Byung-Jun Lee,Jongmin Lee,Kee-Eung Kim
One of the main challenges in offline and off-policy reinforcement learning is to cope with the distribution shift that arises from the mismatch between the target policy and the data collection policy. In this paper, we focus on a model-based approach, particularly on learning the representation for a robust model of the environment under the distribution shift, which has been first studied by Representation Balancing MDP (RepBM). Although this prior work has shown promising results, there are a number of shortcomings that still hinder its applicability to practical tasks. In particular, we address the curse of horizon exhibited by RepBM, rejecting most of the pre-collected data in long-term tasks. We present a new objective for model learning motivated by recent advances in the estimation of stationary distribution corrections. This effectively overcomes the aforementioned limitation of RepBM, as well as naturally extending to continuous action spaces and stochastic policies. We also present an offline model-based policy optimization using this new objective, yielding the state-of-the-art performance in a representative set of benchmark offline RL tasks.
****
Local Convergence Analysis of Gradient Descent Ascent with Finite Timescale Separation
Tanner Fiez,Lillian J Ratliff
We study the role that a finite timescale separation parameter τ has on gradient descent-ascent in non-convex, non-concave zero-sum games where the learning rate of player 1 is denoted by γ_1 and the learning rate of player 2 is defined to be $\gamma_2=\tau\gamma_1$. We provide a non-asymptotic construction of the finite timescale separation parameter $\tau^{\ast}$ such that gradient descent-ascent locally converges to $x^{\ast}$ for all $\tau \in (\tau^{\ast}, \infty)$ if and only if it is a strict local minmax equilibrium. Moreover, we provide explicit local convergence rates given the finite timescale separation. The

convergence results we present are complemented by a non-convergence result: gi
ven a critical point $x^{\ast}$ that is not a strict local minmax equilibrium, w
e present a non-asymptotic construction of a finite timescale separation $\tau_{
0}$ such that gradient descent-ascent with timescale separation $\tau\in (\tau_0
, \infty)$ does not converge to $x^{\ast}$. Finally, we extend the results to gr
adient penalty regularization methods for generative adversarial networks and em
pirically demonstrate on CIFAR-10 and CelebA the significant impact timescale se
paration has on training performance.
****
Visual Explanation using Attention Mechanism in Actor-Critic-based Deep Reinforc
ement Learning
Hidenori Itaya,Tsubasa Hirakawa,Takayoshi Yamashita,Hironobu Fujiyoshi,Komei Sug
iura
Deep reinforcement learning (DRL) has great potential for acquiring the optimal
action in complex environments such as games and robot control. However, it is d
ifficult to analyze the decision-making of the agent, i.e., the reasons it selec
ts the action acquired by learning. In this work, we propose Mask-Attention A3C
(Mask A3C) that introduced an attention mechanism into Asynchronous Advantage Ac
tor-Critic (A3C) which is an actor-critic-based DRL method, and can analyze deci
sion making of agent in DRL. A3C consists of a feature extractor that extracts f
eatures from an image, a policy branch that outputs the policy, value branch tha
t outputs the state value. In our method, we focus on the policy branch and valu
e branch and introduce an attention mechanism to each. In the attention mechanis
m, mask processing is performed on the feature maps of each branch using mask-at
tention that expresses the judgment reason for the policy and state value with a
heat map. We visualized mask-attention maps for games on the Atari 2600 and fou
nd we could easily analyze the reasons behind an agent’s decision-making in vari
ous game tasks. Furthermore, experimental results showed that higher performance
of the agent could be achieved by introducing the attention mechanism.

****
FedMes: Speeding Up Federated Learning with Multiple Edge Servers
Dong-Jun Han,Minseok Choi,Jungwuk Park,Jaekyun Moon
We consider federated learning with multiple wireless edge servers having their
own local coverages. We focus on speeding up training in this increasingly pract
ical setup. Our key idea is to utilize the devices located in the overlapping ar
eas between the coverage of edge servers; in the model-downloading stage, the de
vices in the overlapping areas receive multiple models from different edge serve
rs, take the average of the received models, and then update the model with thei
r local data. These devices send their updated model to multiple edge servers by
broadcasting, which acts as bridges for sharing the trained models between serv
ers. Even when some edge servers are given biased datasets within their coverage
s, their training processes can be assisted by coverages of adjacent servers, th
rough the devices in the overlapping regions. As a result, the proposed scheme d
oes not require costly communications with the central cloud server (located at
the higher tier of edge servers) for model synchronization, significantly reduci
ng the overall training time compared to the conventional cloud-based federated
learning systems. Extensive experimental results show remarkable performance gai
ns of our scheme compared to existing methods.
****
Sself: Robust Federated Learning against Stragglers and Adversaries
Jungwuk Park,Dong-Jun Han,Minseok Choi,Jaekyun Moon
While federated learning allows efficient model training with local data at edge
devices, two major issues that need to be resolved are: slow devices known as s
tragglers and malicious attacks launched by adversaries. While the presence of b
oth stragglers and adversaries raises serious concerns for the deployment of pra
ctical federated learning systems, no known schemes or known combinations of sch
emes, to our best knowledge, effectively address these two issues at the same ti
me. In this work, we propose Sself, a semi-synchronous entropy and loss based fi
ltering/averaging, to tackle both stragglers and adversaries simultaneously. The

stragglers are handled by exploiting different staleness (arrival delay) information when combining locally updated models during periodic global aggregation. Various adversarial attacks are tackled by utilizing a small amount of public data collected at the server in each aggregation step, to first filter out the model-poisoned devices using computed entropies, and then perform weighted averaging based on the estimated losses to combat data poisoning and backdoor attacks. A theoretical convergence bound is established to provide insights on the convergence of Sself. Extensive experimental results show that Sself outperforms various combinations of existing methods aiming to handle stragglers/adversaries.
****

Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-Learning
Dong Bok Lee,Dongchan Min,Seanie Lee,Sung Ju Hwang
Unsupervised learning aims to learn meaningful representations from unlabeled data which can captures its intrinsic structure, that can be transferred to downstream tasks. Meta-learning, whose objective is to learn to generalize across tasks such that the learned model can rapidly adapt to a novel task, shares the spirit of unsupervised learning in that the both seek to learn more effective and efficient learning procedure than learning from scratch. The fundamental difference of the two is that the most meta-learning approaches are supervised, assuming full access to the labels. However, acquiring labeled dataset for meta-training not only is costly as it requires human efforts in labeling but also limits its applications to pre-defined task distributions. In this paper, we propose a principled unsupervised meta-learning model, namely Meta-GMVAE, based on Variational Autoencoder (VAE) and set-level variational inference. Moreover, we introduce a mixture of Gaussian (GMM) prior, assuming that each modality represents each class-concept in a randomly sampled episode, which we optimize with Expectation-Maximization (EM). Then, the learned model can be used for downstream few-shot classification tasks, where we obtain task-specific parameters by performing semi-supervised EM on the latent representations of the support and query set, and predict labels of the query set by computing aggregated posteriors. We validate our model on Omniglot and Mini-ImageNet datasets by evaluating its performance on downstream few-shot classification tasks. The results show that our model obtain impressive performance gains over existing unsupervised meta-learning baselines, even outperforming supervised MAML on a certain setting.
****

Defuse: Debugging Classifiers Through Distilling Unrestricted Adversarial Examples
Dylan Z Slack,Nathalie Rauschmayr,Krishnaram Kenthapadi
With the greater proliferation of machine learning models, the imperative of diagnosing and correcting bugs in models has become increasingly clear. As a route to better discover and fix model bugs, we propose failure scenarios: regions on the data manifold that are incorrectly classified by a model. We propose an end-to-end debugging framework called Defuse to use these regions for fixing faulty classifier predictions. The Defuse framework works in three steps. First, Defuse identifies many unrestricted adversarial examples--naturally occurring instances that are misclassified--using a generative model. Next, the procedure distills the misclassified data using clustering into failure scenarios. Last, the method corrects model behavior on the distilled scenarios through an optimization based approach. We illustrate the utility of our framework on a variety of image data sets. We find that Defuse identifies and resolves concerning predictions while maintaining model generalization.
****

XLA: A Robust Unsupervised Data Augmentation Framework for Cross-Lingual NLP
M Saiful Bari,Tasnim Mohiuddin,Shafiq Joty
Transfer learning has yielded state-of-the-art (SoTA) results in many supervised NLP tasks. However, annotated data for every target task in every target language is rare, especially for low-resource languages. We propose XLA, a novel data augmentation framework for self-supervised learning in zero-resource transfer learning scenarios. In particular, XLA aims to solve cross-lingual adaptation problems from a source language task distribution to an unknown target language tas

k distribution, assuming no training label in the target language task. At its c
ore, XLA performs simultaneous self-training with data augmentation and unsuperv
ised sample selection. To show its effectiveness, we conduct extensive experimen
ts on zero-resource cross-lingual transfer tasks for Named Entity Recognition (N
ER), Natural Language Inference (NLI) and paraphrase identification on Paraphras
e Adversaries from Word Scrambling (PAWS). XLA achieves SoTA results in all the
tasks, outperforming the baselines by a good margin. With an in-depth framework
dissection, we demonstrate the cumulative contributions of different components
to XLA's success.
****

Once Quantized for All: Progressively Searching for Quantized Compact Models
Mingzhu Shen,Feng Liang,Chuming Li,Chen Lin,Ming Sun,Junjie Yan,Wanli Ouyang
Automatic search of Quantized Neural Networks (QNN) has attracted a lot of atten
tion. However, the existing quantization-aware Neural Architecture Search (NAS)
approaches inherit a two-stage search-retrain schema, which is not only time-con
suming but also adversely affected by the unreliable ranking of architectures du
ring the search. To avoid the undesirable effect of the search-retrain schema, w
e present Once Quantized for All (OQA), a novel framework that searches for quan
tized compact models and deploys their quantized weights at the same time withou
t additional post-process. While supporting a huge architecture search space, ou
r OQA can produce a series of quantized compact models under ultra-low bit-width
s(e.g. 4/3/2 bit). A progressive bit inheritance procedure is introduced to supp
ort ultra-low bit-width. Our searched model family, OQANets, achieves a new stat
e-of-the-art (SOTA) on quantized compact models compared with various quantizati
on methods and bit-widths. In particular, OQA2bit-L achieves 64.0\% ImageNet Top
-1 accuracy, outperforming its 2 bit counterpart EfficientNet-B0@QKD by a large
margin of 14\% using 30\% less computation cost.
****

Rewriting by Generating: Learn Heuristics for Large-scale Vehicle Routing Proble
ms
Hansen Wang,Zefang Zong,Tong Xia,Shuyu Luo,Meng Zheng,Depeng Jin,Yong Li
 The large-scale vehicle routing problems are defined based on the classical VRP
s with thousands of customers. It is of great importance to find an efficient an
d high-quality solution for real-world applications. However, existing algorithm
s for VRPs including non-learning heuristics and RL-based methods, only perform
well on small-scale instances with usually no more than a hundred customers. The
y are unable to solve large-scale VRPs due to either high computation cost or ex
plosive solution space that results in model divergence.
Inspired by the classical idea of Divide-and-Conquer, we present a novel Rewriti
ng-by-Generating(RBG) framework with hierarchical RL agents to solve large-scale
 VRPs. RBG consists of a rewriter agent that refines the customer division globa
lly and an elementary generator to infer regional solutions locally. Extensive e
xperiments demonstrate the effectiveness and efficiency of our proposed RBG fram
ework. It outperforms LKH3, the state-of-the-art method for CVRPs, by 2.43% w
hen customer number $N=2000$ and shortens the inference time by about 100 times.
****

The Importance of Pessimism in Fixed-Dataset Policy Optimization
Jacob Buckman,Carles Gelada,Marc G Bellemare
We study worst-case guarantees on the expected return of fixed-dataset policy op
timization algorithms. Our core contribution is a unified conceptual and mathema
tical framework for the study of algorithms in this regime. This analysis reveal
s that for naive approaches, the possibility of erroneous value overestimation l
eads to a difficult-to-satisfy requirement: in order to guarantee that we select
 a policy which is near-optimal, we may need the dataset to be informative of th
e value of every policy. To avoid this, algorithms can follow the pessimism prin
ciple, which states that we should choose the policy which acts optimally in the
 worst possible world. We show why pessimistic algorithms can achieve good perfo
rmance even when the dataset is not informative of every policy, and derive fami
lies of algorithms which follow this principle. These theoretical findings are v
alidated by experiments on a tabular gridworld, and deep learning experiments on

four MinAtar environments.
****
Uncertainty Estimation and Calibration with Finite-State Probabilistic RNNs
Cheng Wang,Carolin Lawrence,Mathias Niepert
Uncertainty quantification is crucial for building reliable and trustable machine learning systems. We propose to estimate uncertainty in recurrent neural networks (RNNs) via stochastic discrete state transitions over recurrent timesteps. The uncertainty of the model can be quantified by running a prediction several times, each time sampling from the recurrent state transition distribution, leading to potentially different results if the model is uncertain. Alongside uncertainty quantification, our proposed method offers several advantages in different settings. The proposed method can (1) learn deterministic and probabilistic automata from data, (2) learn well-calibrated models on real-world classification tasks, (3) improve the performance of out-of-distribution detection, and (4) control the exploration-exploitation trade-off in reinforcement learning. An implementation is available.
****
Hybrid-Regressive Neural Machine Translation
Qiang Wang,Heng Yu,Shaohui Kuang,Weihua Luo
Although the non-autoregressive translation model based on iterative refinement has achieved comparable performance to the autoregressive counterparts with faster decoding, we empirically found that such aggressive iterations make the acceleration rely heavily on small batch size (e.g., 1) and computing device (e.g., GPU).
By designing synthetic experiments, we highlight that iteration times can be significantly reduced when providing a good (partial) target context.
Inspired by this, we propose a two-stage translation prototype -- Hybrid-Regressive Translation (HRT). HRT first jumpily generates a discontinuous sequence by autoregression (e.g., make a prediction every k tokens, k>1). Then, with the help of the partially deterministic target context, HRT fills all the previously skipped tokens with one iteration in a non-autoregressive way.
The experimental results on WMT'16 En-Ro and WMT'14 En-De show that our model outperforms the state-of-the-art non-autoregressive models with multiple iterations, even autoregressive models. Moreover, compared with autoregressive models, HRT can be steadily accelerated 1.5 times regardless of batch size and device.
****
Empirical or Invariant Risk Minimization? A Sample Complexity Perspective
Kartik Ahuja,Jun Wang,Amit Dhurandhar,Karthikeyan Shanmugam,Kush R. Varshney
Recently, invariant risk minimization (IRM) was proposed as a promising solution to address out-of-distribution (OOD) generalization. However, it is unclear when IRM should be preferred over the widely-employed empirical risk minimization (ERM) framework. In this work, we analyze both these frameworks from the perspective of sample complexity, thus taking a firm step towards answering this important question. We find that depending on the type of data generation mechanism, the two approaches might have very different finite sample and asymptotic behavior. For example, in the covariate shift setting we see that the two approaches not only arrive at the same asymptotic solution, but also have similar finite sample behavior with no clear winner. For other distribution shifts such as those involving confounders or anti-causal variables, however, the two approaches arrive at different asymptotic solutions where IRM is guaranteed to be close to the desired OOD solutions in the finite sample regime, while ERM is biased even asymptotically. We further investigate how different factors --- the number of environments, complexity of the model, and IRM penalty weight --- impact the sample complexity of IRM in relation to its distance from the OOD solutions.
****
Efficiently Troubleshooting Image Segmentation Models with Human-In-The-Loop
Haotao Wang,Tianlong Chen,Zhangyang Wang,Kede Ma
Image segmentation lays the foundation for many high-stakes vision applications such as autonomous driving and medical image analysis. It is, therefore, of great importance to not only improve the accuracy of segmentation models on well-est

ablished benchmarks, but also enhance their robustness in the real world so as t
o avoid sparse but fatal failures. In this paper, instead of chasing state-of-th
e-art performance on existing benchmarks, we turn our attention to a new challen
ging problem: how to efficiently expose failures of ``top-performing'' segmentat
ion models in the real world and how to leverage such counterexamples to rectify
 the models. To achieve this with minimal human labelling effort, we first autom
atically sample a small set of images that are likely to falsify the target mode
l from a large corpus of web images via the maximum discrepancy competition prin
ciple. We then propose a weakly labelling strategy to further reduce the number
of false positives, before time-consuming pixel-level labelling by humans. Final
ly, we fine-tune the model to harness the identified failures, and repeat the wh
ole process, resulting in an efficient and progressive framework for troubleshoo
ting segmentation models. We demonstrate the feasibility of our framework using
the semantic segmentation task in PASCAL VOC, and find that the fine-tuned model
 exhibits significantly improved generalization when applied to real-world image
s with greater content diversity. All experimental codes will be publicly releas
ed upon acceptance.
****

Weakly Supervised Scene Graph Grounding
Yizhou Zhang,Zhaoheng Zheng,Yan Liu
 Recent researches have achieved substantial advances in learning structured rep
resentations from images. However, current methods rely heavily on the annotated
 mapping between the nodes of scene graphs and object bounding boxes inside imag
es. Here, we explore the problem of learning the mapping between scene graph nod
es and visual objects under weak supervision. Our proposed method learns a metri
c among visual objects and scene graph nodes by incorporating information from b
oth object features and relational features. Extensive experiments on Visual Gen
ome (VG) and Visual Relation Detection (VRD) datasets verify that our model post
 an improvement on scene graph grounding task over current state-of-the-art appr
oaches. Further experiments on scene graph parsing task verify the grounding fou
nd by our model can reinforce the performance of the existing method.
****

Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrie
val
Lee Xiong,Chenyan Xiong,Ye Li,Kwok-Fung Tang,Jialin Liu,Paul N. Bennett,Junaid A
hmed,Arnold Overwijk
Conducting text retrieval in a learned dense representation space has many intri
guing advantages. Yet dense retrieval (DR) often underperforms word-based sparse
 retrieval. In this paper, we first theoretically show the bottleneck of dense r
etrieval is the domination of uninformative negatives sampled in mini-batch trai
ning, which yield diminishing gradient norms, large gradient variances, and slow
 convergence. We then propose Approximate nearest neighbor Negative Contrastive
Learning (ANCE), which selects hard training negatives globally from the entire
corpus. Our experiments demonstrate the effectiveness of ANCE on web search, que
stion answering, and in a commercial search engine, showing ANCE dot-product ret
rieval nearly matches the accuracy of BERT-based cascade IR pipeline. We also em
pirically validate our theory that negative sampling with ANCE better approximat
es the oracle importance sampling procedure and improves learning convergence.
****

Deep Learning Solution of the Eigenvalue Problem for Differential Operators
Ido Ben-Shaul,Leah Bar,Nir Sochen
Solving the eigenvalue problem for differential operators is a common problem in
 many scientific fields. Classical numerical methods rely on intricate domain di
scretization, and yield non-analytic or non-smooth approximations. We introduce
a novel Neural Network (NN)-based solver for the eigenvalue problem of different
ial self-adjoint operators where the eigenpairs are learned in an unsupervised e
nd-to-end fashion. We propose three different training procedures, for solving i
ncreasingly challenging tasks towards the general eigenvalue problem.

The proposed solver is able to find the M smallest eigenpairs for a general diff

erential operator. We demonstrate the method on the Laplacian operator which is of particular interest in image processing, computer vision, shape analysis among many other applications.
Unlike other numerical methods such as finite differences, the partial derivatives of the network approximation of the eigenfunction can be analytically calculated to any order. Therefore, the proposed framework enables the solution of higher order operators and on free shape domain or even on a manifold. Non-linear operators can be investigated by this approach as well.
****
Enhancing Certified Robustness of Smoothed Classifiers via Weighted Model Ensembling
Chizhou Liu,Yunzhen Feng,Ranran Wang,Bin Dong
Randomized smoothing has achieved state-of-the-art certified robustness against l_2-norm adversarial attacks. However, it is not wholly resolved on how to find the optimal base classifier for randomized smoothing. In this work, we employ a Smoothed WEighted ENsembling (SWEEN) scheme to improve the performance of randomized smoothed classifiers. We show the ensembling generality that SWEEN can help achieve optimal certified robustness. Furthermore, theoretical analysis proves that the optimal SWEEN model can be obtained from training under mild assumptions. We also develop an adaptive prediction algorithm to reduce
the prediction and certification cost of SWEEN models. Extensive experiments show that SWEEN models outperform the upper envelope of their corresponding candidate models by a large margin. Moreover, SWEEN models constructed using a few small models can achieve comparable performance to a single large model with a notable reduction in training time.
****
Contextual Image Parsing via Panoptic Segment Sorting
Jyh-Jing Hwang,Tsung-Wei Ke,Stella Yu
Visual context is versatile and hard to describe or label precisely. We aim to leverage the densely labeled task, image parsing, a.k.a panoptic segmentation, to learn a model that encodes and discovers object-centric context. Most existing approaches based on deep learning tackle image parsing via fusion of pixel-wise classification and instance masks from two sub-networks. Such approaches isolate things from stuff and fuse the semantic and instance masks in the later stage. To encode object-centric context inherently, we propose a metric learning framework, Panoptic Segment Sorting, that is directly trained with stuff and things jointly. Our key insight is to make the panoptic embeddings separate every instance so that the model automatically learns to leverage visual context as many instances across different images appear similar. We show that the context of our model's retrieved instances is more consistent relatively by 13.7%, further demonstrating its ability to discover novel context unsupervisedly. Our overall framework also achieves competitive performance across standard panoptic segmentation metrics amongst the state-of-the-art methods on two large datasets, Cityscapes and PASCAL VOC. These promising results suggest that pixel-wise embeddings can not only inject new understanding into panoptic segmentation but potentially serve for other tasks such as modeling instance relationships.
****
Cross-Domain Few-Shot Learning by Representation Fusion
Thomas Adler,Johannes Brandstetter,Michael Widrich,Andreas Mayr,David Kreil,Michael K Kopp,Günter Klambauer,Sepp Hochreiter
In order to quickly adapt to new data, few-shot learning aims at learning from few examples, often by using already acquired knowledge. The new data often differs from the previously seen data due to a domain shift, that is, a change of the input-target distribution. While several methods perform well on small domain shifts like new target classes with similar inputs, larger domain shifts are still challenging. Large domain shifts may result in abstract concepts that are not shared between the original and the new domain. However, low-level concepts like edges in images might still be shared and useful. For cross-domain few-shot learning, we suggest representation fusion to unify different abstraction levels of a deep neural network into one representation. We propose Cross-domain Hebbian

Ensemble Few-shot learning (CHEF), which consists of representation fusion by an ensemble of Hebbian learners acting on different layers of a deep neural network that was trained on the original domain. On the few-shot datasets miniImagenet and tieredImagenet, where the domain shift is small, CHEF is competitive with state-of-the-art methods. On cross-domain few-shot benchmark challenges with larger domain shifts, CHEF obtains state-of-the-art results in all categories. We further apply CHEF on a real-world cross-domain application in drug discovery. We consider a domain shift from bioactive molecules to environmental chemicals and drugs with twelve associated toxicity prediction tasks. On these tasks that are highly relevant for computational drug discovery, CHEF significantly outperforms all its competitors.
****
Implicit Convex Regularizers of CNN Architectures: Convex Optimization of Two- and Three-Layer Networks in Polynomial Time
Tolga Ergen,Mert Pilanci
We study training of Convolutional Neural Networks (CNNs) with ReLU activations and introduce exact convex optimization formulations with a polynomial complexity with respect to the number of data samples, the number of neurons, and data dimension. More specifically, we develop a convex analytic framework utilizing semi-infinite duality to obtain equivalent convex optimization problems for several two- and three-layer CNN architectures. We first prove that two-layer CNNs can be globally optimized via an ℓ_2 norm regularized convex program. We then show that multi-layer circular CNN training problems with a single ReLU layer are equivalent to an ℓ_1 regularized convex program that encourages sparsity in the spectral domain. We also extend these results to three-layer CNNs with two ReLU layers. Furthermore, we present extensions of our approach to different pooling methods, which elucidates the implicit architectural bias as convex regularizers.
****
Counterfactual Self-Training
Ruijiang Gao,Max Biggs,Wei Sun,Ligong Han
Unlike traditional supervised learning, in many settings only partial feedback is available. We may only observe outcomes for the chosen actions, but not the counterfactual outcomes associated with other alternatives. Such settings encompass a wide variety of applications including pricing, online marketing and precision medicine. A key challenge is that observational data are influenced by historical policies deployed in the system, yielding a biased data distribution. We approach this task as a domain adaptation problem and propose a self-training algorithm which imputes outcomes for the unseen actions in the observational data to simulate a randomized trial. We offer a theoretical motivation for this approach by providing an upper bound on the generalization error defined on a randomized trial under the self-training objective. We empirically demonstrate the effectiveness of the proposed algorithms on both synthetic and real datasets.
****
EXPLORING VULNERABILITIES OF BERT-BASED APIS
Xuanli He,Lingjuan Lyu,Lichao Sun,Xiaojun Chang,Jun Zhao
Natural language processing (NLP) tasks, ranging from text classification to text
generation, have been revolutionised by pretrained BERT models. This allows
corporations to easily build powerful APIs by encapsulating fine-tuned BERT
models. These BERT-based APIs are often designed to not only provide reliable
service but also protect intellectual properties or privacy-sensitive information of
the training data. However, a series of privacy and robustness issues may still exist
when a fine-tuned BERT model is deployed as a service. In this work, we first
present an effective model extraction attack, where the adversary can practically
steal a BERT-based API (the target/victim model). We then demonstrate: (1)
how the extracted model can be further exploited to develop effective attribute

inference attack to expose sensitive information of the training data of the victim
model; (2) how the extracted model can lead to highly transferable adversarial attacks against the victim model. Extensive experiments on multiple benchmark datasets under various realistic settings validate the potential privacy and adversarial
vulnerabilities of BERT-based APIs.
****
Revealing the Structure of Deep Neural Networks via Convex Duality
Tolga Ergen,Mert Pilanci
We study regularized deep neural networks (DNNs) and introduce a convex analytic framework to characterize the structure of the hidden layers. We show that a set of optimal hidden layer weights for a norm regularized DNN training problem can be explicitly found as the extreme points of a convex set. For the special case of deep linear networks with K outputs, we prove that each optimal weight matrix is rank-K and aligns with the previous layers via duality. More importantly, we apply the same characterization to deep ReLU networks with whitened data and prove the same weight alignment holds. As a corollary, we prove that norm regularized deep ReLU networks yield spline interpolation for one-dimensional datasets which was previously known only for two-layer networks. Furthermore, we provide closed-form solutions for the optimal layer weights when data is rank-one or whitened. We then verify our theory via numerical experiments.
****
Learning Robust Models by Countering Spurious Correlations
Haohan Wang,Zeyi Huang,Eric Xing
Machine learning has demonstrated remarkable prediction accuracy over i.i.d data, but the accuracy often drops when tested with data from another distribution. One reason behind this accuracy drop is the reliance of models on the features that are only associated with the label in the training distribution, but not the test distribution. This problem is usually known as spurious correlation, confounding factors, or dataset bias. In this paper, we formally study the generalization error bound for this setup with the knowledge of how the spurious features are associated with the label. We also compare our analysis to the widely-accepted domain adaptation error bound and show that our bound can be tighter, with more assumptions that we consider realistic. Further, our analysis naturally offers a set of solutions for this problem, linked to established solutions in various topics about robustness in general, and these solutions all require some understandings of how the spurious features are associated with the label. Finally, we also briefly discuss a method that does not require such an understanding.
****
FairBatch: Batch Selection for Model Fairness
Yuji Roh,Kangwook Lee,Steven Euijong Whang,Changho Suh
Training a fair machine learning model is essential to prevent demographic disparity. Existing techniques for improving model fairness require broad changes in either data preprocessing or model training, rendering themselves difficult-to-adopt for potentially already complex machine learning systems. We address this problem via the lens of bilevel optimization. While keeping the standard training algorithm as an inner optimizer, we incorporate an outer optimizer so as to equip the inner problem with an additional functionality: Adaptively selecting mini batch sizes for the purpose of improving model fairness. Our batch selection algorithm, which we call FairBatch, implements this optimization and supports prominent fairness measures: equal opportunity, equalized odds, and demographic parity. FairBatch comes with a significant implementation benefit -- it does not require any modification to data preprocessing or model training. For instance, a single-line change of PyTorch code for replacing batch selection part of model training suffices to employ FairBatch. Our experiments conducted both on synthetic and benchmark real data demonstrate that FairBatch can provide such functionalities while achieving comparable (or even greater) performances against the state of the arts. Furthermore, FairBatch can readily improve fairness of any pre-trained model simply via fine-tuning. It is also compatible with existing batch sel

ection techniques intended for different purposes, such as faster convergence, t
hus gracefully achieving multiple purposes.
****

On the Consistency Loss for Leveraging Augmented Data to Learn Robust and Invari
ant Representations

Haohan Wang,Zeyi Huang,Xindi Wu,Eric Xing

Data augmentation is one of the most popular techniques for improving the robust
ness of neural networks. In addition to directly training the model with origina
l samples and augmented samples, a torrent of methods regularizing the distance
between embeddings/representations of the original samples and their augmented c
ounterparts have been introduced. In this paper, we explore these various regula
rization choices, seeking to provide a general understanding of how we should re
gularize the embeddings. Our analysis suggests how the ideal choices of regulari
zation correspond to various assumptions. With an invariance test, we show that
regularization is important if the model is to be used in a broader context than
 the in-lab setting because non-regularized approaches are limited in learning t
he concept of invariance, despite equally high accuracy. Finally, we also show t
hat the generic approach we identified (squared ℓ_2 norm regularized augmen
tation) performs better than several recent methods, which are each specially de
signed for one task and significantly more complicated than ours, over three dif
ferent tasks.
****

Privacy Preserving Recalibration under Domain Shift

Rachel Luo,Shengjia Zhao,Jiaming Song,Jonathan Kuck,Stefano Ermon,Silvio Savares
e

Classifiers deployed in high-stakes applications must output calibrated confiden
ce scores, i.e. their predicted probabilities should reflect empirical frequenci
es. Typically this is achieved with recalibration algorithms that adjust probabi
lity estimates based on the real-world data; however, existing algorithms are no
t applicable in real-world situations where the test data follows a different di
stribution from the training data, and privacy preservation is paramount (e.g. p
rotecting patient records). We introduce a framework that provides abstractions
for performing recalibration under differential privacy constraints. This framew
ork allows us to adapt existing recalibration algorithms to satisfy differential
 privacy while remaining effective for domain-shift situations. Guided by our fr
amework, we also design a novel recalibration algorithm, accuracy temperature sc
aling, that is tailored to the requirements of differential privacy. In an exten
sive empirical study, we find that our algorithm improves calibration on domain-
shift benchmarks under the constraints of differential privacy. On the 15 highes
t severity perturbations of the ImageNet-C dataset, our method achieves a median
 ECE of 0.029, over 2x better than the next best recalibration method and almost
 5x better than without recalibration.
****

Identifying Coarse-grained Independent Causal Mechanisms with Self-supervision

Xiaoyang Wang,Klara Nahrstedt,Oluwasanmi O Koyejo

Current approaches for learning disentangled representations assume that indepen
dent latent variables generate the data through a single data generation process
. In contrast, this manuscript considers independent causal mechanisms (ICM), wh
ich, unlike disentangled representations, directly model multiple data generatio
n processes (mechanisms) in a coarse granularity. In this work, we aim to learn
a model that disentangles each mechanism and approximates the ground-truth mecha
nisms from observational data. We outline sufficient conditions under which the
mechanisms can be learned using a single self-supervised generative model with a
n unconventional mixture prior, simplifying previous methods. Moreover, we prove
 the identifiability of our model w.r.t. the mechanisms in the self-supervised s
cenario. We compare our approach to disentangled representations on various down
stream tasks, showing that our approach is more robust to intervention, covarian
t shift, and noise due to the disentanglement between the data generation proces
ses.
****

BDS-GCN: Efficient Full-Graph Training of Graph Convolutional Nets with Partitio
n-Parallelism and Boundary Sampling
Cheng Wan,Youjie Li,Nam Sung Kim,Yingyan Lin
Graph Convolutional Networks (GCNs) have emerged as the state-of-the-art model f
or graph-based learning tasks. However, it is still challenging to train GCNs at
 scale, limiting their applications to real-world large graphs and hindering the
 exploration of deeper and more sophisticated GCN architectures. While it can be
 natural to leverage graph partition and distributed training for tackling this
challenge, this direction has only been slightly touched on previously due to th
e unique challenge posed by the GCN structures, especially the excessive amount
of boundary nodes in each partitioned subgraph, which can easily explode the req
uired memory and communications for distributed training of GCNs. To this end, w
e propose BDS-GCN, a method that adopts unbiased boundary sampling strategy to e
nable efficient and scalable distributed GCN training while maintaining the full
-graph accuracy. Empirical evaluations and ablation studies validate the effecti
veness of the proposed BDS-GCN, e.g., boosting the throughput by up-to 500% and
reducing the memory usage by up-to 58% for distributed GCN training, while achie
ving the same accuracy, as compared with the state-of-the-art methods. We believ
e our BDS-GCN would open up a new paradigm for enabling GCN training at scale. A
ll code will be released publicly upon acceptance.
****

An Unsupervised Deep Learning Approach for Real-World Image Denoising
Dihan Zheng,Sia Huat Tan,Xiaowen Zhang,Zuoqiang Shi,Kaisheng Ma,Chenglong Bao
Designing an unsupervised image denoising approach in practical applications is
a challenging task due to the complicated data acquisition process. In the real-
world case, the noise distribution is so complex that the simplified additive wh
ite Gaussian (AWGN) assumption rarely holds, which significantly deteriorates th
e Gaussian denoisers' performance. To address this problem, we apply a deep neur
al network that maps the noisy image into a latent space in which the AWGN assum
ption holds, and thus any existing Gaussian denoiser is applicable. More specifi
cally, the proposed neural network consists of the encoder-decoder structure and
 approximates the likelihood term in the Bayesian framework. Together with a Gau
ssian denoiser, the neural network can be trained with the input image itself an
d does not require any pre-training in other datasets. Extensive experiments on
real-world noisy image datasets have shown that the combination of neural networ
ks and Gaussian denoisers improves the performance of the original Gaussian deno
isers by a large margin. In particular, the neural network+BM3D method significa
ntly outperforms other unsupervised denoising approaches and is competitive with
 supervised networks such as DnCNN, FFDNet, and CBDNet.
****

When Optimizing f-Divergence is Robust with Label Noise
Jiaheng Wei,Yang Liu
We show when maximizing a properly defined f-divergence measure with respect t
o a classifier's predictions and the supervised labels is robust with label nois
e. Leveraging its variational form, we derive a nice decoupling property for a f
amily of f-divergence measures when label noise presents, where the divergence
 is shown to be a linear combination of the variational difference defined on th
e clean distribution and a bias term introduced due to the noise. The above deri
vation helps us analyze the robustness of different f-divergence functions. Wi
th established robustness, this family of f-divergence functions arises as use
ful metrics for the problem of learning with noisy labels, which do not require
the specification of the labels' noise rate. When they are possibly not robust,
we propose fixes to make them so. In addition to the analytical results, we pres
ent thorough experimental evidence. Our code is available at https://github.com/
UCSC-REAL/Robust-f-divergence-measures.
****

Communication-Computation Efficient Secure Aggregation for Federated Learning
Beongjun Choi,Jy-yong Sohn,Dong-Jun Han,Jaekyun Moon
Federated learning has been spotlighted as a way to train neural network models
using data distributed over multiple clients without a need to share private dat

a. Unfortunately, however, it has been shown that data privacy could not be full
y guaranteed as adversaries may be able to extract certain information on local
data from the model parameters transmitted during federated learning. A recent s
olution based on the secure aggregation primitive enables privacy-preserving fed
erated learning, but at the expense of significant extra communication/computati
onal resources. In this paper, we propose communication-computation efficient se
cure aggregation which reduces the amount of communication/computational resourc
es at least by a factor of $\sqrt{n/ \log n}$ relative to the existing secure so
lution without sacrificing data privacy, where n is the number of clients. The
 key idea behind the suggested scheme is to design the topology of the secret-sh
aring nodes (denoted by the assignment graph G) as sparse random graphs instea
d of the complete graph corresponding to the existing solution. We first obtain
a sufficient condition on G to guarantee reliable and private federated learni
ng. Afterwards, we suggest using the Erd\H{o}s-Rényi graph as G, and provide t
heoretical guarantees on the reliability/privacy of the proposed scheme. Through
 extensive real-world experiments, we demonstrate that our scheme, using only 50
% of the resources required in the conventional scheme, maintains virtually the
same levels of reliability and data privacy in practical federated learning syst
ems.
****

Energy-Based Models for Continual Learning
Shuang Li,Yilun Du,Gido Martijn van de Ven,Antonio Torralba,Igor Mordatch
We motivate Energy-Based Models (EBMs) as a promising model class for continual
learning problems. Instead of tackling continual learning via the use of externa
l memory, growing models, or regularization, EBMs have a natural way to support
a dynamically-growing number of tasks and classes and less interference with old
 tasks. We show that EBMs are adaptable to a more general continual learning set
ting where the data distribution changes without the notion of explicitly deline
ated tasks. We also find that EBMs outperform the baseline methods by a large ma
rgin on several continual learning benchmarks. These observations point towards
EBMs as a class of models naturally inclined towards the continual learning regi
me.
****

PhraseTransformer: Self-Attention using Local Context for Semantic Parsing
Phuong Minh Nguyen,Vu Tran,Minh Le Nguyen
Semantic parsing is a challenging task whose purpose is to convert a natural lan
guage utterance to machine-understandable information representation. Recently,
solutions using Neural Machine Translation have achieved many promising results,
 especially Transformer because of the ability to learn long-range word dependen
cies. However, the one drawback of adapting the original Transformer to the sema
ntic parsing is the lack of detail in expressing the information of sentences. T
herefore, this work proposes a PhraseTransformer architecture that is capable of
 a more detailed meaning representation by learning the phrase dependencies in t
he sentence. The main idea is to incorporate Long Short-Term Memory (LSTM) into
the Self-Attention mechanism of the original Transformer to capture more local c
ontext of phrases. Experimental results show that the proposed model captures th
e detailed meaning better than Transformer, raises local context awareness and a
chieves strong competitive performance on Geo, MSParS datasets, and leads to SOT
A performance on Atis dataset in methods using Neural Network.
****

Meta-learning Symmetries by Reparameterization
Allan Zhou,Tom Knowles,Chelsea Finn
Many successful deep learning architectures are equivariant to certain transform
ations in order to conserve parameters and improve generalization: most famously
, convolution layers are equivariant to shifts of the input. This approach only
works when practitioners know the symmetries of the task and can manually constr
uct an architecture with the corresponding equivariances. Our goal is an approac
h for learning equivariances from data, without needing to design custom task-sp
ecific architectures. We present a method for learning and encoding equivariance
s into networks by learning corresponding parameter sharing patterns from data.

Our method can provably represent equivariance-inducing parameter sharing for an y finite group of symmetry transformations. Our experiments suggest that it can automatically learn to encode equivariances to common transformations used in im age processing tasks.
****

A Geometric Analysis of Deep Generative Image Models and Its Applications
Binxu Wang,Carlos R Ponce
Generative adversarial networks (GANs) have emerged as a powerful unsupervised m ethod to model the statistical patterns of real-world data sets, such as natural images. These networks are trained to map random inputs in their latent space t o new samples representative of the learned data. However, the structure of the latent space is hard to intuit due to its high dimensionality and the non-linear ity of the generator, which limits the usefulness of the models. Understanding t he latent space requires a way to identify input codes for existing real-world i mages (inversion), and a way to identify directions with known image transformat ions (interpretability). Here, we use a geometric framework to address both issu es simultaneously. We develop an architecture-agnostic method to compute the Rie mannian metric of the image manifold created by GANs. The eigen-decomposition of the metric isolates axes that account for different levels of image variability . An empirical analysis of several pretrained GANs shows that image variation ar ound each position is concentrated along surprisingly few major axes (the space is highly anisotropic) and the directions that create this large variation are s imilar at different positions in the space (the space is homogeneous). We show t hat many of the top eigenvectors correspond to interpretable transforms in the i mage space, with a substantial part of eigenspace corresponding to minor transfo rms which could be compressed out. This geometric understanding unifies key prev ious results related to GAN interpretability. We show that the use of this metri c allows for more efficient optimization in the latent space (e.g. GAN inversion) and facilitates unsupervised discovery of interpretable axes. Our results illu strate that defining the geometry of the GAN image manifold can serve as a gener al framework for understanding GANs.
****

SoGCN: Second-Order Graph Convolutional Networks
Peihao Wang,Yuehao Wang,Hua Lin,Jianbo Shi
We introduce a second-order graph convolution (SoGC), a maximally localized kern el, that can express a polynomial spectral filter with arbitrary coefficients. We contrast our SoGC with vanilla GCN, first-order (one-hop) aggregation, and hi gher-order (multi-hop) aggregation by analyzing graph convolutional layers via g eneralized filter space. We argue that SoGC is a simple design capable of formi ng the basic building block of graph convolution, playing the same role as $3 \t imes 3$ kernels in CNNs. We build purely topological Second-Order Graph Convol utional Networks (SoGCN) and demonstrate that SoGCN consistently achieves state- of-the-art performance on the latest benchmark. Moreover, we introduce the Gated Recurrent Unit (GRU) to spectral GCNs. This explorative attempt further improve s our experimental results.
****

Unsupervised Class-Incremental Learning through Confusion
Shivam Khare,Kun Cao,James Matthew Rehg
While many works on Continual Learning have shown promising results for mitigati ng catastrophic forgetting, they have relied on supervised training. To successf ully learn in a label-agnostic incremental setting, a model must distinguish bet ween learned and novel classes to properly include samples for training. We intr oduce a novelty detection method that leverages network confusion caused by trai ning incoming data as a new class. We found that incorporating a class-imbalance during this detection method substantially enhances performance. The effectiven ess of our approach is demonstrated across a set of common image classification benchmarks: MNIST, SVHN, CIFAR-10, and CIFAR-100.
****

What Makes Instance Discrimination Good for Transfer Learning?
Nanxuan Zhao,Zhirong Wu,Rynson W. H. Lau,Stephen Lin

Contrastive visual pretraining based on the instance discrimination pretext task has made significant progress. Notably, recent work on unsupervised pretraining has shown to surpass the supervised counterpart for finetuning downstream applications such as object detection and segmentation. It comes as a surprise that image annotations would be better left unused for transfer learning. In this work, we investigate the following problems: What makes instance discrimination pretraining good for transfer learning? What knowledge is actually learned and transferred from these models? From this understanding of instance discrimination, how can we better exploit human annotation labels for pretraining? Our findings are threefold. First, what truly matters for the transfer is low-level and mid-level representations, not high-level representations. Second, the intra-category invariance enforced by the traditional supervised model weakens transferability by increasing task misalignment. Finally, supervised pretraining can be strengthened by following an exemplar-based approach without explicit constraints among the instances within the same category.
****

Natural World Distribution via Adaptive Confusion Energy Regularization
Yen-Chi Hsu,Cheng-Yao Hong,Wan-Cyuan Fan,Ding-Jie Chen,Ming-Sui Lee,davi geiger,Tyng-Luh Liu
We introduce a novel and adaptive batch-wise regularization based on the proposed Batch Confusion Norm (BCN) to flexibly address the natural world distribution which usually involves fine-grained and long-tailed properties at the same time. The Fine-Grained Visual Classification (FGVC) problem is notably characterized by two intriguing properties, significant inter-class similarity and intra-class variations, which cause learning an effective FGVC classifier a challenging task. Existing techniques attempt to capture the discriminative parts by their modified attention mechanism. The long-tailed distribution of visual classification poses a great challenge for handling the class imbalance problem. Most of existing solutions usually focus on the class-balancing strategies, classifier normalization, or alleviating the negative gradient of tailed categories. Depart from the conventional approaches, we propose to tackle both problems simultaneously with the adaptive confusion concept. When inter-class similarity prevails in a batch, the BCN term can alleviate possible overfitting due to exploring image features of fine details. On the other hand, when inter-class similarity is not an issue, the class predictions from different samples would unavoidably yield a substantial BCN loss, and prompt the network learning to further reduce the cross-entropy loss. More importantly, extending the existing confusion energy-based framework to account for long-tailed scenario, BCN can learn to exert proper distribution of confusion strength over tailed and head categories to improve classification performance. While the resulting FGVC model by the BCN technique is effective, the performance can be consistently boosted by incorporating extra attention mechanism. In our experiments, we have obtained state-of-the-art results on several benchmark FGVC datasets, and also demonstrated that our approach is competitive on the popular natural world distribution dataset, iNaturalist2018.
****

Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval
Wenhan Xiong,Xiang Li,Srini Iyer,Jingfei Du,Patrick Lewis,William Yang Wang,Yashar Mehdad,Scott Yih,Sebastian Riedel,Douwe Kiela,Barlas Oguz
We propose a simple and efficient multi-hop dense retrieval approach for answering complex open-domain questions, which achieves state-of-the-art performance on two multi-hop datasets, HotpotQA and multi-evidence FEVER. Contrary to previous work, our method does not require access to any corpus-specific information, such as inter-document hyperlinks or human-annotated entity markers, and can be applied to any unstructured text corpus. Our system also yields a much better efficiency-accuracy trade-off, matching the best published accuracy on HotpotQA while being 10 times faster at inference time.
****

Weakly Supervised Neuro-Symbolic Module Networks for Numerical Reasoning
Amrita Saha,Shafiq Joty,Steven Hoi
Neural Module Networks (NMNs) have been quite successful in incorporating explic

it reasoning as learnable modules in various question answering tasks, including the most generic form of numerical reasoning over text in Machine Reading Comprehension (MRC). However, to achieve this, contemporary NMNs need strong supervision in executing the query as a specialized program over the reasoning modules and fail to generalize to more open-ended settings where such supervision is not readily available. In this work, we propose Weakly Supervised Neuro-Symbolic Module Network (WNSMN) trained with answers as the sole supervision for numerical reasoning based MRC. It learns to execute a noisy heuristic program obtained from dependency parsing of the query as discrete actions over both neural and symbolic reasoning modules and trains it end-to-end in a reinforcement learning framework with discrete reward from answer matching. On the numerical-answer subset of the DROP dataset, WNSMN outperforms NMN by 32% and the reasoning-free language model GenBERT by 8% in exact match accuracy when trained under comparable weak supervised settings. This showcases the effectiveness and generalizability of modular networks that can handle explicit discrete reasoning over noisy programs in an end-to-end manner.

****

Oblivious Sketching-based Central Path Method for Solving Linear Programming Problems

Zhao Song,Zheng Yu

In this work, we propose a sketching-based central path method for solving linear programmings, whose running time matches the state of art results [Cohen, Lee, Song STOC 19; Lee, Song, Zhang COLT 19]. Our method opens up the iterations of the central path method and deploys an "iterate and sketch" approach towards the problem by introducing a new coordinate-wise embedding technique, which may be of independent interest. Compare to previous methods, the work [Cohen, Lee, Song STOC 19] enjoys feasibility while being non-oblivious, and [Lee, Song, Zhang COLT 19] is oblivious but infeasible, and relies on dense sketching matrices such as subsampled randomized Hadamard/Fourier transform matrices. Our method enjoys the benefits of being both oblivious and feasible, and can use sparse sketching matrix [Nelson, Nguyen FOCS 13] to speed up the online matrix-vector multiplication. Our framework for solving LP naturally generalizes to a broader class of convex optimization problems including empirical risk minimization.

****

Intriguing class-wise properties of adversarial training

Qi Tian,Kun Kuang,Fei Wu,Yisen Wang

Adversarial training is one of the most effective approaches to improve model robustness against adversarial examples. However, previous works mainly focus on the overall robustness of the model, and the in-depth analysis on the role of each class involved in adversarial training is still missing. In this paper, we provide the first detailed class-wise diagnosis of adversarial training on six widely used datasets, $\textit{i.e.}$, MNIST, CIFAR-10, CIFAR-100, SVHN, STL-10 and ImageNet. Surprisingly, we find that there are $\textit{remarkable robustness discrepancies among classes}$, demonstrating the following intriguing properties: 1) Many examples from a certain class could only be maliciously attacked to some specific semantic-similar classes, and these examples will not exist adversarial counterparts in bounded ϵ-ball if we re-train the model without those specific classes; 2) The robustness of each class is positively correlated with its norm of classifier weight in deep neural networks; 3) Stronger attacks are usually more powerful for vulnerable classes. Finally, we propose an attack to better understand the defense mechanism of some state-of-the-art models from the class-wise perspective. We believe these findings can contribute to a more comprehensive understanding of adversarial training as well as further improvement of adversarial robustness.

****

Empirical Sufficiency Featuring Reward Delay Calibration

Yixuan Liu,Hu Wang,Xiaowei Wang,Xiaoyue Sun,Liuyue Jiang,Minhui Xue

Appropriate credit assignment for delay rewards is a fundamental challenge in various deep reinforcement learning tasks. To tackle this problem, we introduce a

delay reward calibration paradigm inspired from a classification perspective. We hypothesize that when an agent's behavior satisfies an equivalent sufficient condition to be awarded, well-represented state vectors should share similarities. To this end, we define an empirical sufficient distribution, where the state vectors within the distribution will lead agents to environmental reward signals in consequent steps. Therefore, an overfitting classifier is established to handle the distribution and generate calibrated rewards. We examine the correctness of sufficient state extraction by tracking the real-time extraction and building hybrid different reward functions in environments with different levels of awarding latency. The results demonstrate that the classifier could generate timely and accurate calibrated rewards, and the rewards could make the training more efficient. Finally, we find that the sufficient states extracted by our model resonate with observations of human cognition.
****

On Dyadic Fairness: Exploring and Mitigating Bias in Graph Connections
Peizhao Li,Yifei Wang,Han Zhao,Pengyu Hong,Hongfu Liu
Disparate impact has raised serious concerns in machine learning applications and its societal impacts. In response to the need of mitigating discrimination, fairness has been regarded as a crucial property in algorithmic design. In this work, we study the problem of disparate impact on graph-structured data. Specifically, we focus on dyadic fairness, which articulates a fairness concept that a predictive relationship between two instances should be independent of the sensitive attributes. Based on this, we theoretically relate the graph connections to dyadic fairness on link predictive scores in learning graph neural networks, and reveal that regulating weights on existing edges in a graph contributes to dyadic fairness conditionally. Subsequently, we propose our algorithm, \textbf{FairAdj}, to empirically learn a fair adjacency matrix with proper graph structural constraints for fair link prediction, and in the meanwhile preserve predictive accuracy as much as possible. Empirical validation demonstrates that our method delivers effective dyadic fairness in terms of various statistics, and at the same time enjoys a favorable fairness-utility tradeoff.
****

A Framework For Differentiable Discovery Of Graph Algorithms
Hanjun Dai,Xinshi Chen,Yu Li,Xin Gao,Le Song
Recently there is a surge of interests in using graph neural networks (GNNs) to learn algorithms. However, these works focus more on imitating existing algorithms, and are limited in two important aspects: the search space for algorithms is too small and the learned GNN models are not interpretable. To address these issues, we propose a novel framework which enlarge the search space using cheap global information from tree decomposition of the graphs, and can explain the structures of the graph leading to the decision of learned algorithms. We apply our framework to three NP-complete problems on graphs and show that the framework is able to discover effective and explainable algorithms.
****

Spatio-Temporal Graph Scattering Transform
Chao Pan,Siheng Chen,Antonio Ortega
Although spatio-temporal graph neural networks have achieved great empirical success in handling multiple correlated time series, they may be impractical in some real-world scenarios due to a lack of sufficient high-quality training data. Furthermore, spatio-temporal graph neural networks lack theoretical interpretation. To address these issues, we put forth a novel mathematically designed framework to analyze spatio-temporal data. Our proposed spatio-temporal graph scattering transform (ST-GST) extends traditional scattering transform to the spatio-temporal domain. It performs iterative applications of spatio-temporal graph wavelets and nonlinear activation functions, which can be viewed as a forward pass of spatio-temporal graph convolutional networks without training. Since all the filter coefficients in ST-GST are mathematically designed, it is promising for the real-world scenarios with limited training data, and also allows for a theoretical analysis, which shows that the proposed ST-GST is stable to small perturbations of input signals and structures. Finally, our experiments show that i) ST-GS

T outperforms spatio-temporal graph convolutional networks by an increase of 35% in accuracy for MSR Action3D dataset; ii) it is better and computationally more efficient to design the transform based on separable spatio-temporal graphs than the joint ones; and iii) nonlinearity in ST-GST is critical to empirical performance.
****
Class2Simi: A New Perspective on Learning with Label Noise
Songhua Wu,Xiaobo Xia,Tongliang Liu,Bo Han,Mingming Gong,Nannan Wang,Haifeng Liu,Gang Niu
Label noise is ubiquitous in the era of big data. Deep learning algorithms can easily fit the noise and thus cannot generalize well without properly modeling the noise. In this paper, we propose a new perspective on dealing with label noise called ``\textit{Class2Simi}''. Specifically, we transform the training examples with noisy class labels into pairs of examples with noisy similarity labels, and propose a deep learning framework to learn robust classifiers with the noisy similarity labels. Note that a class label shows the class that an instance belongs to; while a similarity label indicates whether or not two instances belong to the same class. It is worthwhile to perform the transformation: We prove that the noise rate for the noisy similarity labels is lower than that of the noisy class labels, because similarity labels themselves are robust to noise. For example, given two instances, even if both of their class labels are incorrect, their similarity label could be correct. Due to the lower noise rate, Class2Simi achieves remarkably better classification accuracy than its baselines that directly deals with the noisy class labels.
****
Optimization Variance: Exploring Generalization Properties of DNNs
Xiao Zhang,Dongrui Wu,Haoyi Xiong,Bo Dai
Unlike the conventional wisdom in statistical learning theory, the test error of a deep neural network (DNN) often demonstrates double descent: as the model complexity increases, it first follows a classical U-shaped curve and then shows a second descent. Through bias-variance decomposition, recent studies revealed that the bell-shaped variance is the major cause of model-wise double descent (when the DNN is widened gradually). This paper investigates epoch-wise double descent, i.e., the test error of a DNN also shows double descent as the number of training epochs increases. Specifically, we extend the bias-variance analysis to epoch-wise double descent, and reveal that the variance also contributes the most to the zero-one loss, as in model-wise double descent. Inspired by this result, we propose a novel metric called optimization variance to measure the diversity of model updates caused by the stochastic gradients of random training batches drawn in the same iteration. This metric can be estimated using samples from the training set only but correlates well with the test error. The proposed optimization variance can be used to predict the generalization ability of a DNN, and hence early stopping can be achieved without using any validation set.
****
Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels
Denis Yarats,Ilya Kostrikov,Rob Fergus
We propose a simple data augmentation technique that can be applied to standard model-free reinforcement learning algorithms, enabling robust learning directly from pixels without the need for auxiliary losses or pre-training. The approach leverages input perturbations commonly used in computer vision tasks to transform input examples, as well as regularizing the value function and policy. Existing model-free approaches, such as Soft Actor-Critic (SAC), are not able to train deep networks effectively from image pixels. However, the addition of our augmentation method dramatically improves SAC's performance, enabling it to reach state-of-the-art performance on the DeepMind control suite, surpassing model-based (Hafner et al., 2019; Lee et al., 2019; Hafner et al., 2018) methods and recently proposed contrastive learning (Srinivas et al., 2020). Our approach, which we dub DrQ: Data-regularized Q, can be combined with any model-free reinforcement learning algorithm. We further demonstrate this by applying it to DQN and signi

ficantly improve its data-efficiency on the Atari 100k benchmark.
****
KETG: A Knowledge Enhanced Text Generation Framework
Yan Cui,Xi Chen,Jiang Qian,Bojin Zhuang,Shaojun Wang,Jing Xiao
Embedding logical knowledge information into text generation is a challenging NL P task. In this paper, we propose a knowledge enhanced text generation (KETG) fr amework, which incorporates both the knowledge and associated text corpus to add ress logicality and diversity in text generation. Specifically, we validate our framework on rhetorical text generation from our newly built rhetoric knowledge graph. Experiments show that our framework outperforms baseline models such as T ransformer and GPT-2, on rhetorical type control, semantic comprehensibility and diversity.
****
MODALS: Modality-agnostic Automated Data Augmentation in the Latent Space
Tsz-Him Cheung,Dit-Yan Yeung
Data augmentation is an efficient way to expand a training dataset by creating a dditional artificial data. While data augmentation is found to be effective in i mproving the generalization capabilities of models for various machine learning tasks, the underlying augmentation methods are usually manually designed and car efully evaluated for each data modality separately, like image processing functi ons for image data and word-replacing rules for text data. In this work, we prop ose an automated data augmentation approach called MODALS (Modality-agnostic Aut omated Data Augmentation in the Latent Space) to augment data for any modality i n a generic way. MODALS exploits automated data augmentation to fine-tune four u niversal data transformation operations in the latent space to adapt the transfo rm to data of different modalities. Through comprehensive experiments, we demons trate the effectiveness of MODALS on multiple datasets for text, tabular, time-s eries and image modalities.
****
ALFWorld: Aligning Text and Embodied Environments for Interactive Learning
Mohit Shridhar,Xingdi Yuan,Marc-Alexandre Cote,Yonatan Bisk,Adam Trischler,Matth ew Hausknecht
Given a simple request like Put a washed apple in the kitchen fridge, humans can reason in purely abstract terms by imagining action sequences and scoring their likelihood of success, prototypicality, and efficiency, all without moving a mu scle. Once we see the kitchen in question, we can update our abstract plans to f it the scene. Embodied agents require the same abilities, but existing work does not yet provide the infrastructure necessary for both reasoning abstractly and executing concretely. We address this limitation by introducing ALFWorld, a simu lator that enables agents to learn abstract, text-based policies in TextWorld (C ôté et al., 2018) and then execute goals from the ALFRED benchmark (Shridhar et al., 2020) in a rich visual environment. ALFWorld enables the creation of a new BUTLER agent whose abstract knowledge, learned in TextWorld, corresponds directl y to concrete, visually grounded actions. In turn, as we demonstrate empirically , this fosters better agent generalization than training only in the visually gr ounded environment. BUTLER's simple, modular design factors the problem to allow researchers to focus on models for improving every piece of the pipeline (langu age understanding, planning, navigation, and visual scene understanding).
****
Latent Skill Planning for Exploration and Transfer
Kevin Xie,Homanga Bharadhwaj,Danijar Hafner,Animesh Garg,Florian Shkurti
To quickly solve new tasks in complex environments, intelligent agents need to b uild up reusable knowledge. For example, a learned world model captures knowledg e about the environment that applies to new tasks. Similarly, skills capture gen eral behaviors that can apply to new tasks. In this paper, we investigate how th ese two approaches can be integrated into a single reinforcement learning agent. Specifically, we leverage the idea of partial amortization for fast adaptation at test time. For this, actions are produced by a policy that is learned over ti me while the skills it conditions on are chosen using online planning. We demons trate the benefits of our design decisions across a suite of challenging locomot

ion tasks and demonstrate improved sample efficiency in single tasks as well as in transfer from one task to another, as compared to competitive baselines. Videos are available at: https://sites.google.com/view/latent-skill-planning/
****

Multi-Source Unsupervised Hyperparameter Optimization
Masahiro Nomura,Yuta Saito
How can we conduct efficient hyperparameter optimization for a completely new task? In this work, we consider a novel setting, where we search for the optimal hyperparameters for a target task of interest using only unlabeled target task and ‘somewhat relevant’ source task datasets. In this setting, it is essential to estimate the ground-truth target task objective using only the available information. We propose estimators to unbiasedly approximate the ground-truth with a desirable variance property. Building on these estimators, we provide a general and tractable hyperparameter optimization procedure for our setting. The experimental evaluations demonstrate that the proposed framework broadens the applications of automated hyperparameter optimization.
****

Deep Retrieval: An End-to-End Structure Model for Large-Scale Recommendations
Weihao Gao,Xiangjun Fan,Jiankai Sun,Kai Jia,Wenzhi Xiao,Chong Wang,Xiaobing Liu
One of the core problems in large-scale recommendations is to retrieve top relevant candidates accurately and efficiently, preferably in sub-linear time. Previous approaches are mostly based on a two-step procedure: first learn an inner-product model and then use maximum inner product search (MIPS) algorithms to search top candidates, leading to potential loss of retrieval accuracy. In this paper, we present Deep Retrieval (DR), an end-to-end learnable structure model for large-scale recommendations. DR encodes all candidates into a discrete latent space. Those latent codes for the candidates are model parameters and to be learnt together with other neural network parameters to maximize the same objective function. With the model learnt, a beam search over the latent codes is performed to retrieve the top candidates. Empirically, we showed that DR, with sub-linear computational complexity, can achieve almost the same accuracy as the brute-force baseline.
****

Evaluating the Disentanglement of Deep Generative Models through Manifold Topology
Sharon Zhou,Eric Zelikman,Fred Lu,Andrew Y. Ng,Gunnar E. Carlsson,Stefano Ermon
Learning disentangled representations is regarded as a fundamental task for improving the generalization, robustness, and interpretability of generative models. However, measuring disentanglement has been challenging and inconsistent, often dependent on an ad-hoc external model or specific to a certain dataset. To address this, we present a method for quantifying disentanglement that only uses the generative model, by measuring the topological similarity of conditional submanifolds in the learned representation. This method showcases both unsupervised and supervised variants. To illustrate the effectiveness and applicability of our method, we empirically evaluate several state-of-the-art models across multiple datasets. We find that our method ranks models similarly to existing methods. We make our code publicly available at https://github.com/stanfordmlgroup/disentanglement.
****

Cooperating RPN's Improve Few-Shot Object Detection
Weilin Zhang,Yu-Xiong Wang,David Forsyth
Learning to detect an object in an image from very few training examples - few-shot object detection - is challenging, because the classifier that sees proposal boxes has very little training data. A particularly challenging training regime occurs when there are one or two training examples. In this case, if the region proposal network (RPN) misses even one high intersection-over-union (IOU) training box, the classifier's model of how object appearance varies can be severely impacted. We use multiple distinct yet cooperating RPN's. Our RPN's are trained to be different, but not too different; doing so yields significant performance improvements over state of the art for COCO and PASCAL VOC in the very few-sho

t setting. This effect appears to be independent of the choice of classifier or dataset.

****

Combining Ensembles and Data Augmentation Can Harm Your Calibration

Yeming Wen,Ghassen Jerfel,Rafael Muller,Michael W Dusenberry,Jasper Snoek,Balaji Lakshminarayanan,Dustin Tran

Ensemble methods which average over multiple neural network predictions are a simple approach to improve a model's calibration and robustness. Similarly, data augmentation techniques, which encode prior information in the form of invariant feature transformations, are effective for improving calibration and robustness. In this paper, we show a surprising pathology: combining ensembles and data augmentation can harm model calibration. This leads to a trade-off in practice, whereby improved accuracy by combining the two techniques comes at the expense of calibration. On the other hand, selecting only one of the techniques ensures good uncertainty estimates at the expense of accuracy. We investigate this pathology and identify a compounding under-confidence among methods which marginalize over sets of weights and data augmentation techniques which soften labels. Finally, we propose a simple correction, achieving the best of both worlds with significant accuracy and calibration gains over using only ensembles or data augmentation individually. Applying the correction produces new state-of-the art in uncertainty calibration and robustness across CIFAR-10, CIFAR-100, and ImageNet.

****

Compute- and Memory-Efficient Reinforcement Learning with Latent Experience Replay

Lili Chen,Kimin Lee,Aravind Srinivas,Pieter Abbeel

Recent advances in off-policy deep reinforcement learning (RL) have led to impressive success in complex tasks from visual observations. Experience replay improves sample-efficiency by reusing experiences from the past, and convolutional neural networks (CNNs) process high-dimensional inputs effectively. However, such techniques demand high memory and computational bandwidth. In this paper, we present Latent Vector Experience Replay (LeVER), a simple modification of existing off-policy RL methods, to address these computational and memory requirements without sacrificing the performance of RL agents. To reduce the computational overhead of gradient updates in CNNs, we freeze the lower layers of CNN encoders early in training due to early convergence of their parameters. Additionally, we reduce memory requirements by storing the low-dimensional latent vectors for experience replay instead of high-dimensional images, enabling an adaptive increase in the replay buffer capacity, a useful technique in constrained-memory settings. In our experiments, we show that LeVER does not degrade the performance of RL agents while significantly saving computation and memory across a diverse set of DeepMind Control environments and Atari games. Finally, we show that LeVER is useful for computation-efficient transfer learning in RL because lower layers of CNNs extract generalizable features, which can be used for different tasks and domains.

****

Dynamic of Stochastic Gradient Descent with State-dependent Noise

Qi Meng,Shiqi Gong,Wei Chen,Zhi-Ming Ma,Tie-Yan Liu

Stochastic gradient descent (SGD) and its variants are mainstream methods to train deep neural networks. Since neural networks are non-convex, more and more works study the dynamic behavior of SGD and its impact to generalization, especially the escaping efficiency from local minima. However, these works make the over-simplified assumption that the distribution of gradient noise is state-independent, although it is state-dependent. In this work, we propose a novel power-law dynamic with state-dependent diffusion to approximate the dynamic of SGD. Then, we prove that the stationary distribution of power-law dynamic is heavy-tailed, which matches the existing empirical observations. Next, we study the escaping efficiency from local minimum of power-law dynamic and prove that the mean escaping time is in polynomial order of the barrier height of the basin, much faster than an exponential order of previous dynamics. It indicates that SGD can escape deep

sharp minima efficiently and tends to stop at flat minima that have lower generalization error. Finally, we conduct experiments to compare SGD and power-law dynamic, and the results verify our theoretical findings.
****

Dynamic Tensor Rematerialization
Marisa Kirisame,Steven Lyubomirsky,Altan Haan,Jennifer Brennan,Mike He,Jared Roesch,Tianqi Chen,Zachary Tatlock
Checkpointing enables the training of deep learning models under restricted memory budgets by freeing intermediate activations from memory and recomputing them on demand. Current checkpointing techniques statically plan these recomputations offline and assume static computation graphs. We demonstrate that a simple online algorithm can achieve comparable performance by introducing Dynamic Tensor Rematerialization (DTR), a greedy online algorithm for checkpointing that is extensible and general, is parameterized by eviction policy, and supports dynamic models. We prove that DTR can train an N-layer linear feedforward network on an $\Omega(\sqrt{N})$ memory budget with only $\mathcal{O}(N)$ tensor operations. DTR closely matches the performance of optimal static checkpointing in simulated experiments. We incorporate a DTR prototype into PyTorch merely by interposing on tensor allocations and operator calls and collecting lightweight metadata on tensors.
****

CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers
SHIYANG LI,Semih Yavuz,Kazuma Hashimoto,Jia Li,Tong Niu,Nazneen Rajani,Xifeng Yan,Yingbo Zhou,Caiming Xiong
Dialogue state trackers have made significant progress on benchmark datasets, but their generalization capability to novel and realistic scenarios beyond the held- out conversations is less understood. We propose controllable counterfactuals (COCO) to bridge this gap and evaluate dialogue state tracking (DST) models on novel scenarios, i.e., would the system successfully tackle the request if the user responded differently but still consistently with the dialogue flow? COCO leverages turn-level belief states as counterfactual conditionals to produce novel conversation scenarios in two steps: (i) counterfactual goal generation at turn- level by dropping and adding slots followed by replacing slot values, (ii) counterfactual conversation generation that is conditioned on (i) and consistent with the dialogue flow. Evaluating state-of-the-art DST models on MultiWOZ dataset with COCO-generated counterfactuals results in a significant performance drop of up to 30.8% (from 49.4% to 18.6%) in absolute joint goal accuracy. In comparison, widely used techniques like paraphrasing only affect the accuracy by at most 2%. Human evaluations show that COCO-generated conversations perfectly reflect the underlying user goal with more than 95% accuracy and are as human-like as the original conversations, further strengthening its reliability and promise to be adopted as part of the robustness evaluation of DST models.
****

Deep Neural Network Fingerprinting by Conferrable Adversarial Examples
Nils Lukas,Yuxuan Zhang,Florian Kerschbaum
In Machine Learning as a Service, a provider trains a deep neural network and gives many users access. The hosted (source) model is susceptible to model stealing attacks, where an adversary derives a surrogate model from API access to the source model. For post hoc detection of such attacks, the provider needs a robust method to determine whether a suspect model is a surrogate of their model. We propose a fingerprinting method for deep neural network classifiers that extracts a set of inputs from the source model so that only surrogates agree with the source model on the classification of such inputs. These inputs are a subclass of transferable adversarial examples which we call conferrable adversarial examples that exclusively transfer with a target label from a source model to its surrogates. We propose a new method to generate these conferrable adversarial examples. We present an extensive study on the irremovability of our fingerprint against fine-tuning, weight pruning, retraining, retraining with different architectures, three model extraction attacks from related work, transfer learning, adversarial training, and two new adaptive attacks. Our fingerprint is robust against di

stillation, related model extraction attacks, and even transfer learning when th
e attacker has no access to the model provider's dataset. Our fingerprint is the
 first method that reaches a ROC AUC of 1.0 in verifying surrogates, compared to
 a ROC AUC of 0.63 by previous fingerprints.
****

ROGA: Random Over-sampling Based on Genetic Algorithm
ZONGDA HAN,XIUQUAN QIAO,SHUBO ZHAN
When using machine learning to solve practical tasks, we often face the problem
 of class imbalance. Unbalanced classes will cause the model to generate preferen
ces during the learning process, thereby ignoring classes with fewer samples. Th
e oversampling algorithm achieves the purpose of balancing the difference in qua
ntity by generating a minority of samples. The quality of the artificial samples
 determines the impact of the oversampling algorithm on model training. Therefor
e, a challenge of the oversampling algorithm is how to find a suitable sample ge
neration space. However, too strong conditional constraints can make the generat
ed samples as non-noise points as possible, but at the same time they also limit
 the search space of the generated samples, which is not conducive to the discov
ery of better-quality new samples. Therefore, based on this problem, we propose
an oversampling algorithm ROGA based on genetic algorithm. Based on random sampl
ing, new samples are gradually generated and the samples that may become noise a
re filtered out. ROGA can ensure that the sample generation space is as wide as
possible, and it can also reduce the noise samples generated. By verifying on mu
ltiple datasets, ROGA can achieve a good result.
****

Controllable Pareto Multi-Task Learning
Xi Lin,Zhiyuan YANG,Qingfu Zhang,Sam Kwong
A multi-task learning (MTL) system aims at solving multiple related tasks at the
 same time. With a fixed model capacity, the tasks would be conflicted with eac
h other, and the system usually has to make a trade-off among learning all of th
em together. Multiple models with different preferences over tasks have to be tr
ained and stored for many real-world applications where the trade-off has to be
made online. This work proposes a novel controllable Pareto multi-task learning
framework, to enable the system to make real-time trade-off switch among differe
nt tasks with a single model. To be specific, we formulate the MTL as a preferen
ce-conditioned multiobjective optimization problem, for which there is a paramet
ric mapping from the preferences to the Pareto stationary solutions. A single hy
pernetwork-based multi-task neural network is built to learn all tasks with diff
erent trade-off preferences among them, where the hypernetwork generates the mod
el parameters conditioned on the preference. At the inference time, MTL practiti
oners can easily control the model performance based on different trade-off pref
erences in real-time. Experiments on different applications demonstrate that the
 proposed model is efficient for solving various multi-task learning problems.
****

On the Estimation Bias in Double Q-Learning
Zhizhou Ren,Guangxiang Zhu,Beining Han,Jianglun Chen,Chongjie Zhang
Double Q-learning is a classical method for reducing overestimation bias, which
is caused by taking maximum estimated values in the Bellman operator. Its varian
ts in the deep Q-learning paradigm have shown great promise in producing reliabl
e value prediction and improving learning performance. However, as shown by prio
r work, double Q-learning is not fully unbiased and still suffers from underesti
mation bias. In this paper, we show that such underestimation bias may lead to m
ultiple non-optimal fixed points under an approximated Bellman operation. To add
ress the concerns of converging to non-optimal stationary solutions, we propose
a simple and effective approach as a partial fix for underestimation bias in dou
ble Q-learning. This approach leverages real returns to bound the target value.
We extensively evaluate the proposed method in the Atari benchmark tasks and dem
onstrate its significant improvement over baseline algorithms.
****

Practical Marginalized Importance Sampling with the Successor Representation
Scott Fujimoto,David Meger,Doina Precup

Marginalized importance sampling (MIS), which measures the density ratio between the state-action occupancy of a target policy and that of a sampling distribution, is a promising approach for off-policy evaluation. However, current state-of-the-art MIS methods rely on complex optimization tricks and succeed mostly on simple toy problems. We bridge the gap between MIS and deep reinforcement learning by observing that the density ratio can be computed from the successor representation of the target policy. The successor representation can be trained through deep reinforcement learning methodology and decouples the reward optimization from the dynamics of the environment, making the resulting algorithm stable and applicable to high-dimensional domains. We evaluate the empirical performance of our approach on a variety of challenging Atari and MuJoCo environments.
****

Streaming Probabilistic Deep Tensor Factorization
shikai fang,Zheng Wang,Zhimeng pan,Ji Liu,Shandian Zhe
Despite the success of existing tensor factorization methods, most of them conduct a multilinear decomposition, and rarely exploit powerful modeling frameworks, like deep neural networks, to capture a variety of complicated interactions in data. More important, for highly expressive, deep factorization, we lack an effective approach to handle streaming data, which are ubiquitous in real-world applications. To address these issues, we propose SPIDER, a Streaming ProbabilistIc Deep tEnsoR factorization method. We first use Bayesian neural networks (NNs) to construct a deep tensor factorization model. We assign a spike-and-slab prior over the NN weights to encourage sparsity and prevent overfitting. We then use Taylor expansions and moment matching to approximate the posterior of the NN output and calculate the running model evidence, based on which we develop an efficient streaming posterior inference algorithm in the assumed-density-filtering and expectation propagation framework. Our algorithm provides responsive incremental updates for the posterior of the latent factors and NN weights upon receiving new tensor entries, and meanwhile select and inhibit redundant/useless weights. We show the advantages of our approach in four real-world applications.
****

Pareto Adversarial Robustness: Balancing Spatial Robustness and Sensitivity-based Robustness
Ke Sun,Mingjie Li,Zhouchen Lin
Adversarial robustness, mainly including sensitivity-based robustness and spatial robustness, plays an integral part in the robust generalization. In this paper, we endeavor to design strategies to achieve comprehensive adversarial robustness. To hit this target, firstly we investigate the less-studied spatial robustness and then integrate existing spatial robustness methods by incorporating both local and global spatial vulnerability into one spatial attack design. Based on this exploration, we further present a comprehensive relationship between natural accuracy, sensitivity-based and different spatial robustness, supported by the strong evidence from the perspective of representation. More importantly, in order to balance these mutual impact within different robustness into one unified framework, we incorporate the Pareto criterion into the adversarial robustness analysis, yielding a novel strategy towards comprehensive robustness called \textit{Pareto Adversarial Training}. The resulting Pareto front, the set of optimal solutions, provides the set of optimal balance among natural accuracy and different adversarial robustness, shedding light on solutions towards comprehensive robustness in the future. To the best of our knowledge, we are the first to consider comprehensive robustness via the multi-objective optimization.
****

Cut-and-Paste Neural Rendering
Anand Bhattad,David Forsyth
Cut-and-paste methods take an object from one image and insert it into another. Doing so often results in unrealistic looking images because the inserted object's shading is inconsistent with the target scene's shading. Existing reshading methods require a geometric and physical model of the inserted object, which is then rendered using environment parameters. Accurately constructing such a model only from a single image is beyond the current understanding of computer visio

n.

We describe an alternative procedure -- cut-and-paste neural rendering, to render the inserted fragment's shading field consistent with the target scene. We use a Deep Image Prior (DIP) as a neural renderer trained to render an image with consistent image decomposition inferences. The resulting rendering from DIP should have an albedo consistent with composite albedo; it should have a shading field that, outside the inserted fragment, is the same as the target scene's shading field;
and composite surface normals are consistent with the final rendering's shading field.
The result is a simple procedure that produces convincing and realistic shading. Moreover, our procedure does not require rendered images or image-decomposition from real images in the training or labeled annotations. In fact, our only use of simulated ground truth is our use of a pre-trained normal estimator. Qualitative results are strong, supported by a user study comparing against state-of-the-art image harmonization baseline.
****
Dissecting Hessian: Understanding Common Structure of Hessian in Neural Networks
Yikai Wu,Xingyu Zhu,Chenwei Wu,Annie N. Wang,Rong Ge
Hessian captures important properties of the deep neural network loss landscape. We observe that eigenvectors and eigenspaces of the layer-wise Hessian for neural network objective have several interesting structures -- top eigenspaces for different models have high overlap, and top eigenvectors form low rank matrices when they are reshaped into the same shape as the weight matrix of the corresponding layer. These structures, as well as the low rank structure of the Hessian observed in previous studies, can be explained by approximating the Hessian using Kronecker factorization. Our new understanding can also explain why some of these structures become weaker when the network is trained with batch normalization. Finally, we show that the Kronecker factorization can be combined with PAC-Bayes techniques to get better generalization bounds.
****
BiGCN: A Bi-directional Low-Pass Filtering Graph Neural Network
Zhixian Chen,Tengfei Ma,Zhihua Jin,Yangqiu Song,Yang Wang
Graph convolutional networks have achieved great success on graph-structured data. Many graph convolutional networks can be regarded as low-pass filters for graph signals. In this paper, we propose a new model, BiGCN, which represents a graph neural network as a bi-directional low-pass filter. Specifically, we not only consider the original graph structure information but also the latent correlation between features, thus BiGCN can filter the signals along with both the original graph and a latent feature-connection graph. Our model outperforms previous graph neural networks in the tasks of node classification and link prediction on benchmark datasets, especially when we add noise to the node features.
****
AdaGCN: Adaboosting Graph Convolutional Networks into Deep Models
Ke Sun,Zhanxing Zhu,Zhouchen Lin
The design of deep graph models still remains to be investigated and the crucial part is how to explore and exploit the knowledge from different hops of neighbors in an efficient way. In this paper, we propose a novel RNN-like deep graph neural network architecture by incorporating AdaBoost into the computation of network; and the proposed graph convolutional network called AdaGCN~(Adaboosting Graph Convolutional Network) has the ability to efficiently extract knowledge from high-order neighbors of current nodes and then integrates knowledge from different hops of neighbors into the network in an Adaboost way. Different from other graph neural networks that directly stack many graph convolution layers, AdaGCN shares the same base neural network architecture among all ``layers'' and is recursively optimized, which is similar to an RNN. Besides, We also theoretically established the connection between AdaGCN and existing graph convolutional methods, presenting the benefits of our proposal. Finally, extensive experiments demonstrate the consistent state-of-the-art prediction performance on graphs across di

fferent label rates and the computational advantage of our approach AdaGCN~\footnote{Code is available at \url{https://github.com/datake/AdaGCN}.}.
****
Wasserstein diffusion on graphs with missing attributes
Zhixian Chen,Tengfei Ma,Yangqiu Song,Yang Wang
Many real-world graphs are attributed graphs where nodes are associated with non-topological features. While attributes can be missing anywhere in an attributed graph, most of existing node representation learning approaches do not consider such incomplete information.
In this paper, we propose a general non-parametric framework to mitigate this problem. Starting from a decomposition of the attribute matrix, we transform node features into discrete distributions in a lower-dimensional space equipped with the Wasserstein metric. On this Wasserstein space, we propose Wasserstein graph diffusion to smooth the distributional representations of nodes with information from their local neighborhoods. This allows us to reduce the distortion caused by missing attributes and obtain integrated representations expressing information of both topology structures and attributes. We then pull the nodes back to the original space and produce corresponding point representations to facilitate various downstream tasks. To show the power of our representation method, we designed two algorithms based on it for node classification (with missing attributes) and matrix completion respectively, and demonstrate their effectiveness in experiments.
****
Learning What To Do by Simulating the Past
David Lindner,Rohin Shah,Pieter Abbeel,Anca Dragan
Since reward functions are hard to specify, recent work has focused on learning policies from human feedback. However, such approaches are impeded by the expense of acquiring such feedback. Recent work proposed that agents have access to a source of information that is effectively free: in any environment that humans have acted in, the state will already be optimized for human preferences, and thus an agent can extract information about what humans want from the state. Such learning is possible in principle, but requires simulating all possible past trajectories that could have led to the observed state. This is feasible in gridworlds, but how do we scale it to complex tasks? In this work, we show that by combining a learned feature encoder with learned inverse models, we can enable agents to simulate human actions backwards in time to infer what they must have done. The resulting algorithm is able to reproduce a specific skill in MuJoCo environments given a single state sampled from the optimal policy for that skill.
****
F^2ed-Learning: Good Fences Make Good Neighbors
Lun Wang,Qi Pang,Shuai Wang,Dawn Song
In this paper, we present F^2ed-Learning, the first federated learning protocol simultaneously defending against both semi-honest server and Byzantine malicious clients. Using a robust mean estimator called FilterL2, F^2ed-Learning is the first FL protocol with dimension-free estimation error against Byzantine malicious clients. Besides, F^2ed-Learning leverages secure aggregation to protect the clients from a semi-honest server who wants to infer the clients' information from the legitimate updates. The main challenge stems from the incompatibility between FilterL2 and secure aggregation. Specifically, to run FilterL2, the server needs to access individual updates from clients while secure aggregation hides those updates from it. We propose to split the clients into shards, securely aggregate each shard's updates and run FilterL2 on the updates from different shards. The evaluation shows that F^2ed-Learning consistently achieves optimal or sub-optimal performance under three attacks among five robust FL protocols. The code for evaluation is available in the supplementary material.
****
Classify and Generate Reciprocally: Simultaneous Positive-Unlabelled Learning and Conditional Generation with Extra Data
Bing Yu,Ke Sun,He Wang,Zhouchen Lin,Zhanxing Zhu
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 pro
vide a potential solution, it is extremely challenging to exploit them. In this
paper, we address this problem by leveraging Positive-Unlabeled~(PU) classificat
ion and conditional generation with extra unlabeled data \emph{simultaneously},
both of which aim to make full use of agnostic unlabeled data to improve classif
ication and generation performances. In particular, we present a novel training
framework to jointly target both PU classification and conditional generation wh
en exposing to extra data, especially out-of-distribution unlabeled data, by exp
loring the interplay between them: 1) enhancing the performance of PU classifier
s with the assistance of a novel Conditional Generative Adversarial Network~(CGA
N) that is robust to noisy labels, 2) leveraging extra data with predicted label
s from a PU classifier to help the generation. Our key contribution is a Classif
ier-Noise-Invariant Conditional GAN~(CNI-CGAN) that can learn the clean data dis
tribution from noisy labels predicted by a PU classifier. Theoretically, we prov
ed the optimal condition of CNI-CGAN and experimentally, we conducted extensive
evaluations on diverse datasets, verifying the simultaneous improvements on both
classification and generation.
****

CcGAN: Continuous Conditional Generative Adversarial Networks for Image Generati
on

Xin Ding,Yongwei Wang,Zuheng Xu,William J Welch,Z. Jane Wang

This work proposes the continuous conditional generative adversarial network (Cc
GAN), the first generative model for image generation conditional on continuous,
scalar conditions (termed regression labels). Existing conditional GANs (cGANs)
are mainly designed for categorical conditions (e.g., class labels); conditioni
ng on a continuous label is mathematically distinct and raises two fundamental p
roblems: (P1) Since there may be very few (even zero) real images for some regre
ssion labels, minimizing existing empirical versions of cGAN losses (a.k.a. empi
rical cGAN losses) often fails in practice; (P2) Since regression labels are sca
lar and infinitely many, conventional label input methods (e.g., combining a hid
den map of the generator/discriminator with a one-hot encoded label) are not app
licable. The proposed CcGAN solves the above problems, respectively, by (S1) ref
ormulating existing empirical cGAN losses to be appropriate for the continuous s
cenario; and (S2) proposing a novel method to incorporate regression labels into
the generator and the discriminator. The reformulation in (S1) leads to two nov
el empirical discriminator losses, termed the hard vicinal discriminator loss (H
VDL) and the soft vicinal discriminator loss (SVDL) respectively, and a novel em
pirical generator loss. The error bounds of a discriminator trained with HVDL an
d SVDL are derived under mild assumptions in this work. A new benchmark dataset,
RC-49, is also proposed for generative image modeling conditional on regression
labels. Our experiments on the Circular 2-D Gaussians, RC-49, and UTKFace datas
ets show that CcGAN is able to generate diverse, high-quality samples from the i
mage distribution conditional on a given regression label. Moreover, in these ex
periments, CcGAN substantially outperforms cGAN both visually and quantitatively
.
****

ANOCE: Analysis of Causal Effects with Multiple Mediators via Constrained Struct
ural Learning

Hengrui Cai,Rui Song,Wenbin Lu

In the era of causal revolution, identifying the causal effect of an exposure on
the outcome of interest is an important problem in many areas, such as epidemic
s, medicine, genetics, and economics. Under a general causal graph, the exposure
may have a direct effect on the outcome and also an indirect effect regulated b
y a set of mediators. An analysis of causal effects that interprets the causal m
echanism contributed through mediators is hence challenging but on demand. To th
e best of our knowledge, there are no feasible algorithms that give an exact dec
omposition of the indirect effect on the level of individual mediators, due to c
ommon interaction among mediators in the complex graph. In this paper, we establ
ish a new statistical framework to comprehensively characterize causal effects w
ith multiple mediators, namely, ANalysis Of Causal Effects (ANOCE), with a newly

introduced definition of the mediator effect, under the linear structure equation model. We further propose a constrained causal structure learning method by incorporating a novel identification constraint that specifies the temporal causal relationship of variables. The proposed algorithm is applied to investigate the causal effects of 2020 Hubei lockdowns on reducing the spread of the coronavirus in Chinese major cities out of Hubei.
****

Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate
Jingfeng Wu,Difan Zou,Vladimir Braverman,Quanquan Gu
Understanding the algorithmic bias of stochastic gradient descent (SGD) is one of the key challenges in modern machine learning and deep learning theory. Most of the existing works, however, focus on very small or even infinitesimal learning rate regime, and fail to cover practical scenarios where the learning rate is moderate and annealing. In this paper, we make an initial attempt to characterize the particular regularization effect of SGD in the moderate learning rate regime by studying its behavior for optimizing an overparameterized linear regression problem. In this case, SGD and GD are known to converge to the unique minimum-norm solution; however, with the moderate and annealing learning rate, we show that they exhibit different directional bias: SGD converges along the large eigenvalue directions of the data matrix, while GD goes after the small eigenvalue directions. Furthermore, we show that such directional bias does matter when early stopping is adopted, where the SGD output is nearly optimal but the GD output is suboptimal. Finally, our theory explains several folk arts in practice used for SGD hyperparameter tuning, such as (1) linearly scaling the initial learning rate with batch size; and (2) overrunning SGD with high learning rate even when the loss stops decreasing.
****

D2p-fed:Differentially Private Federated Learning with Efficient Communication
Lun Wang,Ruoxi Jia,Dawn Song
In this paper, we propose the discrete Gaussian based differentially private federated learning (D2p-fed), a unified scheme to achieve both differential privacy (DP) and communication efficiency in federated learning (FL). In particular, compared with the only prior work taking care of both aspects, D2p-fed provides stronger privacy guarantee, better composability and smaller communication cost. The key idea is to apply the discrete Gaussian noise to the private data transmission. We provide complete analysis of the privacy guarantee, communication cost and convergence rate of D2p-fed. We evaluated D2p-fed on INFIMNIST and CIFAR10. The results show that D2p-fed outperforms the-state-of-the-art by 4.7% to 13.0% in terms of model accuracy while saving one third of the communication cost. The code for evaluation is available in the supplementary material.
****

Visual Imitation with Reinforcement Learning using Recurrent Siamese Networks
Glen Berseth,Florian Golemo,Christopher Pal
It would be desirable for a reinforcement learning (RL) based agent to learn behaviour by merely watching a demonstration. However, defining rewards that facilitate this goal within the RL paradigm remains a challenge. Here we address this problem with Siamese networks, trained to compute distances between observed behaviours and an agent's behaviours. We use an RNN-based comparator model to learn such distances in space and time between motion clips while training an RL policy to minimize this distance. Through experimentation, we have also found that the inclusion of multi-task data and an additional image encoding loss helps enforce temporal consistency and improve policy learning. These two components appear to balance reward for matching a specific instance of a behaviour versus that behaviour in general. Furthermore, we focus here on a particularly challenging form of this problem where only a single demonstration is provided for a given task -- the one-shot learning setting. We demonstrate our approach on humanoid, dog and raptor agents in 2D and a 3D quadruped and humanoid. In these environments, we show that our method outperforms the state-of-the-art, GAIfO (i.e. GAIL without access to actions) and TCNs.

****
NOSE Augment: Fast and Effective Data Augmentation Without Searching
Qingrui Li,Song Xie,An■l Oymagil,Mustafa Furkan Eseoglu,Ziyin Zhang,CM Lee
Data augmentation has been widely used for enhancing the diversity of training data and model generalization. Different from traditional handcrafted methods, recent research introduced automated search for optimal data augmentation policies and achieved state-of-the-art results on image classification tasks. However, these search-based implementations typically incur high computation cost and long search time because of large search spaces and complex searching algorithms. We revisited automated augmentation from alternate perspectives, such as increasing diversity and manipulating the overall usage of augmented data. In this paper, we present an augmentation method without policy searching called NOSE Augment (NO SEarch Augment). Our method completely skips policy searching; instead, it jointly applies multi-stage augmentation strategy and introduces more augmentation operations on top of a simple stochastic augmentation mechanism. With more augmentation operations, we boost the data diversity of stochastic augmentation; and with the phased complexity driven strategy, we ensure the whole training process converged smoothly to a good quality model. We conducted extensive experiments and showed that our method could match or surpass state-of-the-art results provided by search-based methods in terms of accuracies. Without the need for policy search, our method is much more efficient than the existing AutoAugment series of methods. Besides image classification, we also examine the general validity of our proposed method by applying our method to Face Recognition and Text Detection of the Optical Character Recognition (OCR) problems. The results establish our proposed method as a fast and competitive data augmentation strategy that can be used across various CV tasks.
****
Neural Pooling for Graph Neural Networks
Sai Sree Harsha,Deepak Mishra
Tasks such as graph classification, require graph pooling to learn graph-level representations from constituent node representations. In this work, we propose two novel methods using fully connected neural network layers for graph pooling, namely Neural Pooling Method 1 and 2. Our proposed methods have the ability to handle variable number of nodes in different graphs, and are also invariant to the isomorphic structures of graphs. In addition, compared to existing graph pooling methods, our proposed methods are able to capture information from all nodes, collect second-order statistics, and leverage the ability of neural networks to learn relationships among node representations, making them more powerful. We perform experiments on graph classification tasks in the bio-informatics and social network domains to determine the effectiveness of our proposed methods. Experimental results show that our methods lead to an absolute increase of upto 1.2% in classification accuracy over previous works and a general decrease in standard deviation across multiple runs indicating greater reliability. Experimental results also indicate that this improvement in performance is consistent across several datasets.

****
Robust early-learning: Hindering the memorization of noisy labels
Xiaobo Xia,Tongliang Liu,Bo Han,Chen Gong,Nannan Wang,Zongyuan Ge,Yi Chang
The \textit{memorization effects} of deep networks show that they will first memorize training data with clean labels and then those with noisy labels. The \textit{early stopping} method therefore can be exploited for learning with noisy labels. However, the side effect brought by noisy labels will influence the memorization of clean labels before early stopping. In this paper, motivated by the \textit{lottery ticket hypothesis} which shows that only partial parameters are important for generalization, we find that only partial parameters are important for fitting clean labels and generalize well, which we term as \textit{critical parameters}; while the other parameters tend to fit noisy labels and cannot generalize well, which we term as \textit{non-critical parameters}. Based on this, we propose \textit{robust early-learning} to reduce the side effect of noisy label

s before early stopping and thus enhance the memorization of clean labels. Specifically, in each iteration, we divide all parameters into the critical and non-critical ones, and then perform different update rules for different types of parameters. Extensive experiments on benchmark-simulated and real-world label-noise datasets demonstrate the superiority of the proposed method over the state-of-the-art label-noise learning methods.
****

SMiRL: Surprise Minimizing Reinforcement Learning in Unstable Environments
Glen Berseth,Daniel Geng,Coline Manon Devin,Nicholas Rhinehart,Chelsea Finn,Dinesh Jayaraman,Sergey Levine
Every living organism struggles against disruptive environmental forces to carve out and maintain an orderly niche. We propose that such a struggle to achieve and preserve order might offer a principle for the emergence of useful behaviors in artificial agents. We formalize this idea into an unsupervised reinforcement learning method called surprise minimizing reinforcement learning (SMiRL). SMiRL alternates between learning a density model to evaluate the surprise of a stimulus, and improving the policy to seek more predictable stimuli. The policy seeks out stable and repeatable situations that counteract the environment's prevailing sources of entropy. This might include avoiding other hostile agents, or finding a stable, balanced pose for a bipedal robot in the face of disturbance forces. We demonstrate that our surprise minimizing agents can successfully play Tetris, Doom, control a humanoid to avoid falls, and navigate to escape enemies in a maze without any task-specific reward supervision. We further show that SMiRL can be used together with standard task rewards to accelerate reward-driven learning.
****

A Design Space Study for LISTA and Beyond
Tianjian Meng,Xiaohan Chen,Yifan Jiang,Zhangyang Wang
In recent years, great success has been witnessed in building problem-specific deep networks from unrolling iterative algorithms, for solving inverse problems and beyond. Unrolling is believed to incorporate the model-based prior with the learning capacity of deep learning. This paper revisits \textit{the role of unrolling as a design approach for deep networks}: to what extent its resulting special architecture is superior, and can we find better? Using LISTA for sparse recovery as a representative example, we conduct the first thorough \textit{design space study} for the unrolled models. Among all possible variations, we focus on extensively varying the connectivity patterns and neuron types, leading to a gigantic design space arising from LISTA. To efficiently explore this space and identify top performers, we leverage the emerging tool of neural architecture search (NAS). We carefully examine the searched top architectures in a number of settings, and are able to discover networks that consistently better than LISTA. We further present more visualization and analysis to ``open the black box", and find that the searched top architectures demonstrate highly consistent and potentially transferable patterns. We hope our study to spark more reflections and explorations on how to better mingle model-based optimization prior and data-driven learning.
****

PAC Confidence Predictions for Deep Neural Network Classifiers
Sangdon Park,Shuo Li,Insup Lee,Osbert Bastani
A key challenge for deploying deep neural networks (DNNs) in safety critical settings is the need to provide rigorous ways to quantify their uncertainty. In this paper, we propose a novel algorithm for constructing predicted classification confidences for DNNs that comes with provable correctness guarantees. Our approach uses Clopper-Pearson confidence intervals for the Binomial distribution in conjunction with the histogram binning approach to calibrated prediction. In addition, we demonstrate how our predicted confidences can be used to enable downstream guarantees in two settings: (i) fast DNN inference, where we demonstrate how to compose a fast but inaccurate DNN with an accurate but slow DNN in a rigorous way to improve performance without sacrificing accuracy, and (ii) safe planning, where we guarantee safety when using a DNN to predict whether a given action i

s safe based on visual observations. In our experiments, we demonstrate that our approach can be used to provide guarantees for state-of-the-art DNNs.
****
On Trade-offs of Image Prediction in Visual Model-Based Reinforcement Learning
Mohammad Babaeizadeh,Mohammad Taghi Saffar,Danijar Hafner,Dumitru Erhan,Harini Kannan,Chelsea Finn,Sergey Levine
Model-based reinforcement learning (MBRL) methods have shown strong sample efficiency and performance across a variety of tasks, including when faced with high-dimensional visual observations. These methods learn to predict the environment dynamics and expected reward from interaction and use this predictive model to plan and perform the task. However, MBRL methods vary in their fundamental design choices, and it there is no strong consensus in the literature on how these design decisions affect performance. In this paper, we study a number of design decisions for the predictive model in visual MBRL algorithms, focusing specifically on methods that use a predictive model for planning. We find that a range of design decisions that are often considered crucial, such as the use of latent spaces, have little effect on task performance. A big exception to this finding is that predicting future observations (i.e., images) leads to significant task performance improvement compared to only predicting rewards. We also empirically find that image prediction accuracy, somewhat surprisingly, correlates more strongly with downstream task performance than reward prediction accuracy. We show how this phenomenon is related to exploration and how some of the lower-scoring models on standard benchmarks (that require exploration) will perform the same as the best-performing models when trained on the same training data. Simultaneously, in the absence of exploration, models that fit the data better usually perform better on the down-stream task as well, but surprisingly, these are often not the same models that perform the best when learning and exploring from scratch. These findings suggest that performance and exploration place important and potentially contradictory requirements on the model.
****
Accelerating Safe Reinforcement Learning with Constraint-mismatched Policies
Tsung-Yen Yang,Justinian Rosca,Karthik R Narasimhan,Peter Ramadge
We consider the problem of reinforcement learning when provided with (1) a baseline control policy and (2) a set of constraints that the controlled system must satisfy. The baseline policy can arise from a teacher agent, demonstration data or even a heuristic while the constraints might encode safety, fairness or other application-specific requirements. Importantly, the baseline policy may be sub-optimal for the task at hand, and is not guaranteed to satisfy the specified constraints. The key challenge therefore lies in effectively leveraging the baseline policy for faster learning, while still ensuring that the constraints are minimally violated. To reconcile these potentially competing aspects, we propose an iterative policy optimization algorithm that alternates between maximizing expected return on the task, minimizing distance to the baseline policy, and projecting the policy onto the constraint-satisfying set. We analyze the convergence of our algorithm theoretically and provide a finite-sample guarantee. In our empirical experiments on five different control tasks, our algorithm consistently outperforms several state-of-the-art methods, achieving 10 times fewer constraint violations and 40% higher reward on average.
****
Invertible Manifold Learning for Dimension Reduction
Siyuan Li,Haitao Lin,Zelin Zang,Lirong Wu,Jun Xia,Stan Z. Li
It is widely believed that a dimension reduction (DR) process drops information inevitably in most practical scenarios. Thus, most methods try to preserve some essential information of data after DR, as well as manifold based DR methods. However, they usually fail to yield satisfying results, especially in high-dimensional cases. In the context of manifold learning, we think that a good low-dimensional representation should preserve the topological and geometric properties of data manifolds, which involve exactly the entire information of the data manifolds. In this paper, we define the problem of information-lossless NLDR with the manifold assumption and propose a novel two-stage NLDR method, called invertible

manifold learning ($\textit{inv-ML}$), to tackle this problem. A $\textit{local isometry}$ constraint of preserving local geometry is applied under this assumption in $\textit{inv-ML}$. Firstly, a homeomorphic $\textit{sparse coordinate transformation}$ is learned to find the low-dimensional representation without losing topological information. Secondly, a $\textit{linear compression}$ is performed on the learned sparse coding, with the trade-off between the target dimension and the incurred information loss. Experiments are conducted on seven datasets with a neural network implementation of $\textit{inv-ML}$, called $\textit{i-ML-Enc}$, which demonstrate that the proposed $\textit{inv-ML}$ not only achieves invertible NLDR in comparison with typical existing methods but also reveals the characteristics of the learned manifolds through linear interpolation in latent space. Moreover, we find that the reliability of tangent space approximated by the local neighborhood on real-world datasets is key to the success of manifold based DR algorithms. The code will be made available soon.

****

Entropic Risk-Sensitive Reinforcement Learning: A Meta Regret Framework with Function Approximation

Yingjie Fei,Zhuoran Yang,Zhaoran Wang

We study risk-sensitive reinforcement learning with the entropic risk measure and function approximation. We consider the finite-horizon episodic MDP setting, and propose a meta algorithm based on value iteration. We then derive two algorithms for linear and general function approximation, namely RSVI.L and RSVI.G, respectively, as special instances of the meta algorithm. We illustrate that the success of RSVI.L depends crucially on carefully designed feature mapping and regularization that adapt to risk sensitivity. In addition, both RSVI.L and RSVI.G maintain risk-sensitive optimism that facilitates efficient exploration. On the analytic side, we provide regret analysis for the algorithms by developing a meta analytic framework, at the core of which is a risk-sensitive optimism condition. We show that any instance of the meta algorithm that satisfies the condition yields a meta regret bound. We further verify the condition for RSVI.L and RSVI.G under respective function approximation settings to obtain concrete regret bounds that scale sublinearly in the number of episodes.

****

Learning to Actively Learn: A Robust Approach

Jifan Zhang,Kevin Jamieson

This work proposes a procedure for designing algorithms for specific adaptive data collection tasks like active learning and pure-exploration multi-armed bandits. Unlike the design of traditional adaptive algorithms that rely on concentration of measure and careful analysis to justify the correctness and sample complexity of the procedure, our adaptive algorithm is learned via adversarial training over equivalence classes of problems derived from information theoretic lower bounds. In particular, a single adaptive learning algorithm is learned that competes with the best adaptive algorithm learned for each equivalence class. Our procedure takes as input just the available queries, set of hypotheses, loss function, and total query budget. This is in contrast to existing meta-learning work that learns an adaptive algorithm relative to an explicit, user-defined subset or prior distribution over problems which can be challenging to define and be mismatched to the instance encountered at test time. This work is particularly focused on the regime when the total query budget is very small, such as a few dozen, which is much smaller than those budgets typically considered by theoretically derived algorithms. We perform synthetic experiments to justify the stability and effectiveness of the training procedure, and then evaluate the method on tasks derived from real data including a noisy 20 Questions game and a joke recommendation task.

****

Beyond Trivial Counterfactual Generations with Diverse Valuable Explanations

Pau Rodriguez,Massimo Caccia,Alexandre Lacoste,Lee Zamparo,Issam H. Laradji,Laurent Charlin,David Vazquez

Explainability of machine learning models has gained considerable attention with

in our research community given the importance of deploying more reliable machin
e-learning systems. Explanability can also be helpful for model debugging. In co
mputer vision applications, most methods explain models by displaying the region
s in the input image that they focus on for their prediction, but it is difficul
t to improve models based on these explanations since they do not indicate why t
he model fail. Counterfactual methods, on the other hand, indicate how to pertur
b the input to change the model prediction, providing details about the model's
decision-making. Unfortunately, current counterfactual methods make ambiguous in
terpretations as they combine multiple biases of the model and the data in a sin
gle counterfactual interpretation of the model's decision. Moreover, these metho
ds tend to generate trivial counterfactuals about the model's decision, as they
often suggest to exaggerate or remove the presence of the attribute being classi
fied. Trivial counterfactuals are usually not valuable, since the information th
ey provide is often already known to the system's designer. In this work, we pro
pose a counterfactual method that learns a perturbation in a disentangled latent
 space that is constrained using a diversity-enforcing loss to uncover multiple
valuable explanations about the model's prediction. Further, we introduce a mech
anism to prevent the model from producing trivial explanations. Experiments on C
elebA and Synbols demonstrate that our model improves the success rate of produc
ing high-quality valuable explanations when compared to previous state-of-the-ar
t methods. We will make the code public.
****

X2T: Training an X-to-Text Typing Interface with Online Learning from User Feedb
ack

Jensen Gao,Siddharth Reddy,Glen Berseth,Nicholas Hardy,Nikhilesh Natraj,Karunesh
 Ganguly,Anca Dragan,Sergey Levine

We aim to help users communicate their intent to machines using flexible, adapti
ve interfaces that translate arbitrary user input into desired actions. In this
work, we focus on assistive typing applications in which a user cannot operate a
 keyboard, but can instead supply other inputs, such as webcam images that captu
re eye gaze or neural activity measured by a brain implant. Standard methods tra
in a model on a fixed dataset of user inputs, then deploy a static interface tha
t does not learn from its mistakes; in part, because extracting an error signal
from user behavior can be challenging. We investigate a simple idea that would e
nable such interfaces to improve over time, with minimal additional effort from
the user: online learning from user feedback on the accuracy of the interface's
actions. In the typing domain, we leverage backspaces as feedback that the inter
face did not perform the desired action. We propose an algorithm called x-to-tex
t (X2T) that trains a predictive model of this feedback signal, and uses this mo
del to fine-tune any existing, default interface for translating user input into
 actions that select words or characters. We evaluate X2T through a small-scale
online user study with 12 participants who type sentences by gazing at their des
ired words, a large-scale observational study on handwriting samples from 60 use
rs, and a pilot study with one participant using an electrocorticography-based b
rain-computer interface. The results show that X2T learns to outperform a non-ad
aptive default interface, stimulates user co-adaptation to the interface, person
alizes the interface to individual users, and can leverage offline data collecte
d from the default interface to improve its initial performance and accelerate o
nline learning.
****

Discrete Graph Structure Learning for Forecasting Multiple Time Series

Chao Shang,Jie Chen,Jinbo Bi

Time series forecasting is an extensively studied subject in statistics, economi
cs, and computer science. Exploration of the correlation and causation among the
 variables in a multivariate time series shows promise in enhancing the performa
nce of a time series model. When using deep neural networks as forecasting model
s, we hypothesize that exploiting the pairwise information among multiple (multi
variate) time series also improves their forecast. If an explicit graph structur
e is known, graph neural networks (GNNs) have been demonstrated as powerful tool
s to exploit the structure. In this work, we propose learning the structure simu

ltaneously with the GNN if the graph is unknown. We cast the problem as learning a probabilistic graph model through optimizing the mean performance over the graph distribution. The distribution is parameterized by a neural network so that discrete graphs can be sampled differentiably through reparameterization. Empirical evaluations show that our method is simpler, more efficient, and better performing than a recently proposed bilevel learning approach for graph structure learning, as well as a broad array of forecasting models, either deep or non-deep learning based, and graph or non-graph based.
****

SEMANTIC APPROACH TO AGENT ROUTING USING A HYBRID ATTRIBUTE-BASED RECOMMENDER SYSTEM

Anwitha Paruchuri

Traditionally contact centers route an issue to an agent based on ticket load or skill of the agent. When a ticket comes into the system, it is either manually analyzed and pushed to an agent or automatically routed to an agent based on some business rules. A Customer Relationship Management (CRM) system often has predefined categories that an issue could belong to. The agents are generally proficient in handling multiple categories, the categories in the CRM system are often related to each other, and a ticket typically contains content across multiple categories. This makes the traditional approach sub-optimal. We propose a Hybrid Recommendation based approach that recommends top N agents for a ticket by jointly modelling on the interactions between the agents and categories as well as on the semantic features of the categories and the agents.
****

Optimistic Exploration with Backward Bootstrapped Bonus for Deep Reinforcement Learning

Chenjia Bai,Lingxiao Wang,Peng Liu,Zhaoran Wang,Jianye HAO,Yingnan Zhao

Optimism in the face of uncertainty is a principled approach for provably efficient exploration for reinforcement learning in tabular and linear settings. However, such an approach is challenging in developing practical exploration algorithms for Deep Reinforcement Learning (DRL). To address this problem, we propose an Optimistic Exploration algorithm with Backward Bootstrapped Bonus (OEB3) for DRL by following these two principles. OEB3 is built on bootstrapped deep Q-learning, a non-parametric posterior sampling method for temporally-extended exploration. Based on such a temporally-extended exploration, we construct an UCB-bonus indicating the uncertainty of Q-functions. The UCB-bonus is further utilized to estimate an optimistic Q-value, which encourages the agent to explore the scarcely visited states and actions to reduce uncertainty. In the estimation of Q-function, we adopt an episodic backward update strategy to propagate the future uncertainty to the estimated Q-function consistently. Extensive evaluations show that OEB3 outperforms several state-of-the-art exploration approaches in Mnist maze and 49 Atari games.
****

Task-Agnostic Morphology Evolution

Donald Joseph Hejna III,Pieter Abbeel,Lerrel Pinto

Deep reinforcement learning primarily focuses on learning behavior, usually overlooking the fact that an agent's function is largely determined by form. So, how should one go about finding a morphology fit for solving tasks in a given environment? Current approaches that co-adapt morphology and behavior use a specific task's reward as a signal for morphology optimization. However, this often requires expensive policy optimization and results in task-dependent morphologies that are not built to generalize. In this work, we propose a new approach, Task-Agnostic Morphology Evolution (TAME), to alleviate both of these issues. Without any task or reward specification, TAME evolves morphologies by only applying randomly sampled action primitives on a population of agents. This is accomplished using an information-theoretic objective that efficiently ranks agents by their ability to reach diverse states in the environment and the causality of their actions. Finally, we empirically demonstrate that across 2D, 3D, and manipulation environments TAME can evolve morphologies that match the multi-task performance of those learned with task supervised algorithms. Our code and videos can be found

at https://sites.google.com/view/task-agnostic-evolution .

****
Policy Learning Using Weak Supervision
Jingkang Wang,Hongyi Guo,Zhaowei Zhu,Yang Liu
Most existing policy learning solutions require the learning agents to receive high-quality supervision signals, e.g., rewards in reinforcement learning (RL) or high-quality expert demonstrations in behavior cloning (BC). These quality supervisions are either infeasible or prohibitively expensive to obtain in practice. We aim for a unified framework that leverages the weak supervisions to perform policy learning efficiently. To handle this problem, we treat the ``weak supervisions'' as imperfect information coming from a \emph{peer agent}, and evaluate the learning agent's policy based on a ``correlated agreement'' with the peer agent's policy (instead of simple agreements). Our way of leveraging peer agent's information offers us a family of solutions that learn effectively from weak supervisions with theoretical guarantees. Extensive evaluations on tasks including RL with noisy reward, BC with weak demonstrations, and standard policy co-training (RL + BC) show that the proposed approach leads to substantial improvements, especially when the complexity or the noise of the learning environments grows.
****
Deep Equals Shallow for ReLU Networks in Kernel Regimes
Alberto Bietti,Francis Bach
Deep networks are often considered to be more expressive than shallow ones in terms of approximation. Indeed, certain functions can be approximated by deep networks provably more efficiently than by shallow ones, however, no tractable algorithms are known for learning such deep models. Separately, a recent line of work has shown that deep networks trained with gradient descent may behave like (tractable) kernel methods in a certain over-parameterized regime, where the kernel is determined by the architecture and initialization, and this paper focuses on approximation for such kernels. We show that for ReLU activations, the kernels derived from deep fully-connected networks have essentially the same approximation properties as their shallow two-layer counterpart, namely the same eigenvalue decay for the corresponding integral operator. This highlights the limitations of the kernel framework for understanding the benefits of such deep architectures. Our main theoretical result relies on characterizing such eigenvalue decays through differentiability properties of the kernel function, which also easily applies to the study of other kernels defined on the sphere.
****
Loss Function Discovery for Object Detection via Convergence-Simulation Driven Search
Peidong Liu,Gengwei Zhang,Bochao Wang,Hang Xu,Xiaodan Liang,Yong Jiang,Zhenguo Li
Designing proper loss functions for vision tasks has been a long-standing research direction to advance the capability of existing models. For object detection, the well-established classification and regression loss functions have been carefully designed by considering diverse learning challenges (e.g. class imbalance, hard negative samples, and scale variances). Inspired by the recent progress in network architecture search, it is interesting to explore the possibility of discovering new loss function formulations via directly searching the primitive operation combinations. So that the learned losses not only fit for diverse object detection challenges to alleviate huge human efforts, but also have better alignment with evaluation metric and good mathematical convergence property. Beyond the previous auto-loss works on face recognition and image classification, our work makes the first attempt to discover new loss functions for the challenging object detection from primitive operation levels and finds the searched losses are insightful. We propose an effective convergence-simulation driven evolutionary search algorithm, called CSE-Autoloss, for speeding up the search progress by regularizing the mathematical rationality of loss candidates via two progressive convergence simulation modules: convergence property verification and model optimization simulation. CSE-Autoloss involves the search space (i.e. 21 mathematic

al operators, 3 constant-type inputs, and 3 variable-type inputs) that cover a wide range of the possible variants of existing losses and discovers best-searched loss function combination within a short time (around 1.5 wall-clock days with 20x speedup in comparison to the vanilla evolutionary algorithm). We conduct extensive evaluations of loss function search on popular detectors and validate the good generalization capability of searched losses across diverse architectures and various datasets. Our experiments show that the best-discovered loss function combinations outperform default combinations (Cross-entropy/Focal loss for classification and L1 loss for regression) by 1.1% and 0.8% in terms of mAP for two-stage and one-stage detectors on COCO respectively. Our searched losses are available at https://github.com/PerdonLiu/CSE-Autoloss.
****

A Simple Sparse Denoising Layer for Robust Deep Learning
Yueming Lyu,Xingrui Yu,Ivor Tsang
Deep models have achieved great success in many applications. However, vanilla deep models are not well-designed against the input perturbation. In this work, we take an initial step to designing a simple robust layer as a lightweight plug-in for vanilla deep models. To achieve this goal, we first propose a fast sparse coding and dictionary learning algorithm for sparse coding problem with an exact k-sparse constraint or l_0 norm regularization. Our method comes with a closed-form approximation for the sparse coding phase by taking advantage of a novel structured dictionary. With this handy approximation, we propose a simple sparse denoising layer (SDL) as a lightweight robust plug-in. Extensive experiments on both classification and reinforcement learning tasks manifest the effectiveness of our methods.
****

Effective Distributed Learning with Random Features: Improved Bounds and Algorithms
Yong Liu,Jiankun Liu,Shuqiang Wang
In this paper, we study the statistical properties of distributed kernel ridge regression together with random features (DKRR-RF), and obtain optimal generalization bounds under the basic setting, which can substantially relax the restriction on the number of local machines in the existing state-of-art bounds. Specifically, we first show that the simple combination of divide-and-conquer technique and random features can achieve the same statistical accuracy as the exact KRR in expectation requiring only $\mathcal{O}(|\mathcal{D}|)$ memory and $\mathcal{O}(|\mathcal{D}|^{1.5})$ time. Then, beyond the generalization bounds in expectation that demonstrate the average information for multiple trails, we derive generalization bounds in probability to capture the learning performance for a single trail. Finally, we propose an effective communication strategy to further improve the performance of DKRR-RF, and validate the theoretical bounds via numerical experiments.
****

Deep Learning is Singular, and That's Good
Daniel Murfet,Susan Wei,Mingming Gong,Hui Li,Jesse Gell-Redman,Thomas Quella
In singular models, the optimal set of parameters forms an analytic set with singularities and classical statistical inference cannot be applied to such models. This is significant for deep learning as neural networks are singular and thus ``dividing" by the determinant of the Hessian or employing the Laplace approximation are not appropriate. Despite its potential for addressing fundamental issues in deep learning, singular learning theory appears to have made little inroads into the developing canon of deep learning theory. Via a mix of theory and experiment, we present an invitation to singular learning theory as a vehicle for understanding deep learning and suggest important future work to make singular learning theory directly applicable to how deep learning is performed in practice.
****

Better Together: Resnet-50 accuracy with $13 \times $ fewer parameters and at $3 \times $ speed
Utkarsh Nath,Shrinu Kushagra
Recent research on compressing deep neural networks has focused on reducing the

number of parameters. Smaller networks are easier to export and deploy on edge-d
evices. We introduce Adjoined networks as a training approach that can regulariz
e and compress any CNN-based neural architecture. Our one-shot learning paradigm
 trains both the original and the smaller networks together. The parameters of t
he smaller network are shared across both the architectures. We prove strong the
oretical guarantees on the regularization behavior of the adjoint training parad
igm. We complement our theoretical analysis by an extensive empirical evaluation
 of both the compression and regularization behavior of adjoint networks. For re
snet-50 trained adjointly on Imagenet, we are able to achieve a $13.7 \times$ re
duction in the number of parameters and a $3 \times$ improvement in inference ti
me without any significant drop in accuracy. For the same architecture on CIFAR-
100, we are able to achieve a $99.7 \times$ reduction in the number of parameter
s and a $5 \times$ improvement in inference time. On both these datasets, the or
iginal network trained in the adjoint fashion gains about 3% in top-1 accurac
y as compared to the same network trained in the standard fashion.
****

On InstaHide, Phase Retrieval, and Sparse Matrix Factorization
Sitan Chen,Xiaoxiao Li,Zhao Song,Danyang Zhuo
In this work, we examine the security of InstaHide, a scheme recently proposed b
y \cite{hsla20} for preserving the security of private datasets in the context o
f distributed learning. To generate a synthetic training example to be shared am
ong the distributed learners, InstaHide takes a convex combination of private fe
ature vectors and randomly flips the sign of each entry of the resulting vector
with probability 1/2. A salient question is whether this scheme is secure in any
 provable sense, perhaps under a plausible complexity-theoretic assumption.

The answer to this turns out to be quite subtle and closely related to the avera
ge-case complexity of a multi-task, missing-data version of the classic problem
of phase retrieval that is interesting in its own right. Motivated by this conne
ction, under the standard distributional assumption that the public/private feat
ure vectors are isotropic Gaussian, we design an algorithm that can actually rec
over a private vector using only the public vectors and a sequence of synthetic
vectors generated by InstaHide.
****

Optimization Planning for 3D ConvNets
Zhaofan Qiu,Ting Yao,Chong-wah Ngo,Tao Mei
3D Convolutional Neural Networks (3D ConvNets) have been regarded as a powerful
class of models for video recognition. Nevertheless, it is not trivial to optima
lly learn a 3D ConvNets due to high complexity and various options of the traini
ng scheme. The most common hand-tuning process starts from learning 3D ConvNets
using short video clips and then is followed by learning long-term temporal depe
ndency using lengthy clips, while gradually decaying the learning rate from high
 to low as training progresses. The fact that such process comes along with seve
ral heuristic settings motivates the study to seek an optimal ``path'' to automa
te the entire training. In this paper, we decompose the path into a series of tr
aining ``states'' and specify the hyper-parameters, e.g., learning rate and the
length of input clips, in each state. The estimation of the knee point on the pe
rformance-epoch curve triggers the transition from one state to another. We perf
orm dynamic programming over all the candidate states to plan the optimal permut
ation of states, i.e., optimization path. Furthermore, we devise a new 3D ConvNe
ts with a unique design of dual-head classifier to improve the spatial and tempo
ral discrimination. Extensive experiments conducted on seven public video recogn
ition benchmarks demonstrate the advantages of our proposal. With the optimizati
on planning, our 3D ConvNets achieves superior results when comparing to the sta
te-of-the-art video recognition approaches. More remarkably, we obtain the top-1
 accuracy of 82.5% and 84.3% on the large-scale Kinetics-400 and Kinetics-600 da
tasets, respectively.
****

Graph Neural Network Acceleration via Matrix Dimension Reduction
Shunhua Jiang,Yunze Man,Zhao Song,Danyang Zhuo

Graph Neural Networks (GNNs) have become the de facto method for machine learnin g on graph data (e.g., social networks, protein structures, code ASTs), but they require significant time and resource to train. One alternative method is Graph Neural Tangent Kernel (GNTK), a kernel method that corresponds to infinitely wi de multi-layer GNNs. GNTK's parameters can be solved directly in a single step, avoiding time-consuming gradient descent. Today, GNTK is the state-of-the-art me thod to achieve high training speed without compromising accuracy. Unfortunately , solving for the kernel and searching for parameters can still take hours to da ys on real-world graphs. The current computation of GNTK has running time $O(N^4)$, where N is the number of nodes in the graph. This prevents GNTK from scali ng to datasets that contain large graphs. Theoretically, we present two techniqu es to speed up GNTK training while preserving the generalization error: (1) We u se a novel matrix decoupling method to reduce matrix dimensions during the kerne l solving. This allows us to reduce the dominated computation bottleneck term fr om $O(N^4)$ to $O(N^3)$. (2) We apply sketching to further reduce the bottlenec k term to $o(N^{\omega})$, where $\omega \approx 2.373$ is the exponent of curre nt matrix multiplication. Experimentally, we demonstrate that our approaches spe ed up kernel learning by up to $19\times$ on real-world benchmark datasets.
****

Ricci-GNN: Defending Against Structural Attacks Through a Geometric Approach
Ze Ye,Tengfei Ma,Chien-Chun Ni,Kin Sum Liu,Jie Gao,Chao Chen
Graph neural networks (GNNs) rely heavily on the underlying graph topology and t hus can be vulnerable to malicious attacks targeting at graph structures. We pro pose a novel GNN defense algorithm against structural attacks that maliciously m odify graph topology. In particular, we discover a robust representation of the input graph based on the advanced theory of graph Ricci flow, which captures the intrinsic geometry of graphs and is robust to structural perturbation. We propo se an algorithm to train GNNs using re-sampled graphs based on such geometric re presentation. We show that this method can be effective to protect against adver sarial structural attacks. Our method achieves state-of-the-art performance on s ynthetic and real datasets against different types of graph poisoning attacks.
****

Anchor & Transform: Learning Sparse Embeddings for Large Vocabularies
Paul Pu Liang,Manzil Zaheer,Yuan Wang,Amr Ahmed
Learning continuous representations of discrete objects such as text, users, mov ies, and URLs lies at the heart of many applications including language and user modeling. When using discrete objects as input to neural networks, we often ign ore the underlying structures (e.g., natural groupings and similarities) and emb ed the objects independently into individual vectors. As a result, existing meth ods do not scale to large vocabulary sizes. In this paper, we design a simple an d efficient embedding algorithm that learns a small set of anchor embeddings and a sparse transformation matrix. We call our method Anchor & Transform (ANT) as the embeddings of discrete objects are a sparse linear combination of the anchor s, weighted according to the transformation matrix. ANT is scalable, flexible, a nd end-to-end trainable. We further provide a statistical interpretation of our algorithm as a Bayesian nonparametric prior for embeddings that encourages spars ity and leverages natural groupings among objects. By deriving an approximate in ference algorithm based on Small Variance Asymptotics, we obtain a natural exten sion that automatically learns the optimal number of anchors instead of having t o tune it as a hyperparameter. On text classification, language modeling, and mo vie recommendation benchmarks, we show that ANT is particularly suitable for lar ge vocabulary sizes and demonstrates stronger performance with fewer parameters (up to 40x compression) as compared to existing compression baselines.
****

Semi-Supervised Audio Representation Learning for Modeling Beehive Strengths
Tony Zhang,Szymon Zmyslony,Sergei Nozdrenkov,Matthew Smith,Brandon Kingsley Hopk ins
Honey bees are critical to our ecosystem and food security as a pollinator, cont ributing 35% of our global agriculture yield. In spite of their importance, beek eeping is exclusively dependent on human labor and experience-derived heuristics

, while requiring frequent human checkups to ensure the colony is healthy, which can disrupt the colony. Increasingly, pollinator populations are declining due to threats from climate change, pests, environmental toxicity, making their management even more critical than ever before in order to ensure sustained global food security. To start addressing this pressing challenge, we developed an integrated hardware sensing system for beehive monitoring through audio and environment measurements, and a hierarchical semi-supervised deep learning model, composed of an audio modeling module and a predictor, to model the strength of beehives. The model is trained jointly on audio reconstruction and prediction losses based on human inspections, in order to model both low-level audio features and circadian temporal dynamics. We show that this model performs well despite limited labels, and can learn an audio embedding that is useful for characterizing different sound profiles of beehives. This is the first instance to our knowledge of applying audio-based deep learning to model beehives and population size in an observational setting across a large number of hives.
****
Decoupled Greedy Learning of Graph Neural Networks
YEWEN WANG,Jian Tang,Yizhou Sun,Guy Wolf
Graph Neural Networks (GNNs) become very popular for graph-related applications due to their superior performance. However, they have been shown to be computationally expensive in large scale settings, because their produced node embeddings have to be computed recursively, which scales exponentially with the number of layers. To address this issue, several sampling-based methods have recently been proposed to perform training on a subset of nodes while maintaining the fidelity of the trained model. In this work, we introduce a decoupled greedy learning method for GNNs (DGL-GNN) that, instead of sampling the input graph, decouples the GNN into smaller modules and associates each module with greedy auxiliary objectives. Our approach allows GNN layers to be updated during the training process without waiting for feedback from successor layers, thus making parallel GNN training possible. Our method achieves improved efficiency without significantly compromising model performances, which would be important for time or memory limited applications. Further, we propose a lazy-update scheme during training to further improve its efficiency. We empirically analyse our proposed DGL-GNN model, and demonstrate its effectiveness and superior efficiency through a range of experiments. Compared to the sampling-based acceleration, our model is more stable, and we do not have to trade-off between efficiency and accuracy. Finally, we note that while here we focus on comparing the decoupled approach as an alternative to other methods, it can also be regarded as complementary, for example, to sampling and other scalability-enhancing improvements of GNN training.
****
Achieving Linear Speedup with Partial Worker Participation in Non-IID Federated Learning
Haibo Yang,Minghong Fang,Jia Liu
Federated learning (FL) is a distributed machine learning architecture that leverages a large number of workers to jointly learn a model with decentralized data. FL has received increasing attention in recent years thanks to its data privacy protection, communication efficiency and a linear speedup for convergence in training (i.e., convergence performance increases linearly with respect to the number of workers). However, existing studies on linear speedup for convergence are only limited to the assumptions of i.i.d. datasets across workers and/or full worker participation, both of which rarely hold in practice. So far, it remains an open question whether or not the linear speedup for convergence is achievable under non-i.i.d. datasets with partial worker participation in FL. In this paper, we show that the answer is affirmative. Specifically, we show that the federated averaging (FedAvg) algorithm (with two-sided learning rates) on non-i.i.d. datasets in non-convex settings achieves a convergence rate $\mathcal{O}(\frac{1}{\sqrt{mKT}} + \frac{1}{T})$ for full worker participation and a convergence rate $\mathcal{O}(\frac{\sqrt{K}}{\sqrt{nT}} + \frac{1}{T})$ for partial worker participation, where K is the number of local steps, T is the number of total communication rounds, m is the total worker number and n is the worker num

ber in one communication round if for partial worker participation. Our results also reveal that the local steps in FL could help the convergence and show that the maximum number of local steps can be improved to T/m in full worker participation. We conduct extensive experiments on MNIST and CIFAR-10 to verify our theoretical results.
****
Exploiting Playbacks in Unsupervised Domain Adaptation for 3D Object Detection
Yurong You,Carlos Andres Diaz-Ruiz,Yan Wang,Wei-Lun Chao,Bharath Hariharan,Mark Campbell,Kilian Q Weinberger
Self-driving cars must detect other vehicles and pedestrians in 3D to plan safe routes and avoid collisions.
State-of-the-art 3D object detectors, based on deep learning, have shown promising accuracy but are prone to over-fit to domain idiosyncrasies, causing them to fail in new environments---a serious problem if autonomous vehicles are meant to operate freely. In this paper, we propose a novel learning approach that drastically reduces this gap by fine-tuning the detector on pseudo-labels in the target domain, which our method generates while the vehicle is parked, based on replays of previously recorded driving sequences. In these replays, objects are tracked over time, and detections are interpolated and extrapolated---crucially, leveraging future information to catch hard cases. We show, on five autonomous driving datasets, that fine-tuning the detector on these pseudo-labels substantially reduces the domain-gap to new driving environments, yielding drastic improvements in accuracy and detection reliability.
****
Interactive Visualization for Debugging RL
Shuby Deshpande,Benjamin Eysenbach,Jeff Schneider
Visualization tools for supervised learning (SL) allow users to interpret, introspect, and gain an intuition for the successes and failures of their models. While reinforcement learning (RL) practitioners ask many of the same questions while debugging agent policies, existing tools aren't a great fit for the RL setting as these tools address challenges typically found in the SL regime. Whereas SL involves a static dataset, RL often entails collecting new data in challenging environments with partial observability, stochasticity, and non-stationary data distributions. This necessitates the creation of alternate visual interfaces to help us better understand agent policies trained using RL. In this work, we design and implement an interactive visualization tool for debugging and interpreting RL. Our system identifies and addresses important aspects missing from existing tools such as (1) visualizing alternate state representations (different from those seen by the agent) that researchers could use while debugging RL policies; (2) interactive interfaces tailored to metadata stored while training RL agents (3) a conducive workflow designed around RL policy debugging. We provide an example workflow of how this system could be used, along with ideas for future extensions.
****
PABI: A Unified PAC-Bayesian Informativeness Measure for Incidental Supervision Signals
Hangfeng He,Mingyuan Zhang,Qiang Ning,Dan Roth
Real-world applications often require making use of {\em a range of incidental supervision signals}. However, we currently lack a principled way to measure the benefit an incidental training dataset can bring, and the common practice of using indirect, weaker signals is through exhaustive experiments with various models and hyper-parameters. This paper studies whether we can, {\em in a single framework, quantify the benefit of various types of incidental signals for one's target task without going through combinatorial experiments}. We propose PABI, a unified informativeness measure motivated by PAC-Bayesian theory, characterizing the reduction in uncertainty that indirect, weak signals provide. We demonstrate PABI's use in quantifying various types of incidental signals including partial labels, noisy labels, constraints, cross-domain signals, and combinations of these. Experiments with various setups on two natural language processing (NLP) tasks, named entity recognition (NER) and question answering (QA), show that PABI c

orrelates well with learning performance, providing a promising way to determine , ahead of learning, which supervision signals would be beneficial.
****
Real-Time AutoML
Iddo Drori,Brandon Kates,Anant Kharkar,Lu Liu,Qiang Ma,Jonah Deykin,Nihar Sidhu, Madeleine Udell
We present a new zero-shot approach to automated machine learning (AutoML) that predicts a high-quality model for a supervised learning task and dataset in real -time without fitting a single model. In contrast, most AutoML systems require t ens or hundreds of model evaluations. Hence our approach accelerates AutoML by o rders of magnitude. Our method uses a transformer-based language embedding to re present datasets and algorithms using their free-text descriptions and a meta-fe ature extractor to represent the data. We train a graph neural network in which each node represents a dataset to predict the best machine learning pipeline for a new test dataset. The graph neural network generalizes to new datasets and ne w sets of datasets. Our approach leverages the progress of unsupervised represen tation learning in natural language processing to provide a significant boost to AutoML. Performance is competitive with state-of-the-art AutoML systems while r educing running time from minutes to seconds and prediction time from minutes to milliseconds, providing AutoML in real-time.
****
Data-driven Learning of Geometric Scattering Networks
Alexander Tong,Frederik Wenkel,Kincaid Macdonald,Smita Krishnaswamy,Guy Wolf
Many popular graph neural network (GNN) architectures, which are often considere d as the current state of the art, rely on encoding graph structure via smoothne ss or similarity between neighbors. While this approach performs well on a surpr ising number of standard benchmarks, the efficacy of such models does not transl ate consistently to more complex domains, such as graph data in the biochemistry domain. We argue that these more complex domains require priors that encourage learning of longer range features rather than oversmoothed signals of standard G NN architectures. Here, we propose an alternative GNN architecture, based on a r elaxation of recently proposed geometric scattering transforms, which consists o f a cascade of graph wavelet filters. Our learned geometric scattering (LEGS) ar chitecture adaptively tunes these wavelets and their scales to encourage band-pa ss features to emerge in learned representations. This results in a simplified G NN with significantly fewer learned parameters compared to competing methods. We demonstrate the predictive performance of our method on several biochemistry gr aph classification benchmarks, as well as the descriptive quality of its learned features in biochemical graph data exploration tasks. Our results show that the proposed LEGS network matches or outperforms popular GNNs, as well as the origi nal geometric scattering construction, while retaining certain mathematical prop erties of its handcrafted (nonlearned) design.
****
Goal-Auxiliary Actor-Critic for 6D Robotic Grasping with Point Clouds
Lirui Wang,Yu Xiang,Dieter Fox
6D robotic grasping beyond top-down bin-picking scenarios is a challenging task. Previous solutions based on 6D grasp synthesis with robot motion planning usual ly operate in an open-loop setting without considering perception feedback and d ynamics and contacts of objects, which makes them sensitive to grasp synthesis e rrors. In this work, we propose a novel method for learning closed-loop control policies for 6D robotic grasping using point clouds from an egocentric camera. W e combine imitation learning and reinforcement learning in order to grasp unseen objects and handle the continuous 6D action space, where expert demonstrations are obtained from a joint motion and grasp planner. We introduce a goal-auxiliar y actor-critic algorithm, which uses grasping goal prediction as an auxiliary ta sk to facilitate policy learning. The supervision on grasping goals can be obtai ned from the expert planner for known objects or from hindsight goals for unknow n objects. Overall, our learned closed-loop policy achieves over 90% success rates on grasping various ShapeNet objects and YCB objects in simulation. The po licy also transfers well to the real world with only one failure among grasping

of ten different unseen objects in the presence of perception noises.
****
Multiple Descent: Design Your Own Generalization Curve
Lin Chen,Yifei Min,Mikhail Belkin,amin karbasi
This paper explores the generalization loss of linear regression in variably parameterized families of models, both under-parameterized and over-parameterized. We show that the generalization curve can have an arbitrary number of peaks, and moreover, locations of those peaks can be explicitly controlled. Our results highlight the fact that both classical U-shaped generalization curve and the recently observed double descent curve are not intrinsic properties of the model family. Instead, their emergence is due to the interaction between the properties of the data and the inductive biases of learning algorithms.
****
Deep Neural Tangent Kernel and Laplace Kernel Have the Same RKHS
Lin Chen,Sheng Xu
We prove that the reproducing kernel Hilbert spaces (RKHS) of a deep neural tangent kernel and the Laplace kernel include the same set of functions, when both kernels are restricted to the sphere $\mathbb{S}^{d-1}$. Additionally, we prove that the exponential power kernel with a smaller power (making the kernel less smooth) leads to a larger RKHS, when it is restricted to the sphere $\mathbb{S}^{d-1}$ and when it is defined on the entire $\mathbb{R}^d$.
****
Fast convergence of stochastic subgradient method under interpolation
Huang Fang,Zhenan Fan,Michael Friedlander
This paper studies the behaviour of the stochastic subgradient descent (SSGD) method applied to over-parameterized nonsmooth optimization problems that satisfy an interpolation condition. By leveraging the composite structure of the empirical risk minimization problems, we prove that SSGD converges, respectively, with rates $O(1/\epsilon)$ and $O(\log(1/\epsilon))$ for convex and strongly-convex objectives when interpolation holds. These rates coincide with established rates for the stochastic gradient descent (SGD) method applied to smooth problems that also satisfy an interpolation condition. Our analysis provides a partial explanation for the empirical observation that sometimes SGD and SSGD behave similarly for training smooth and nonsmooth machine learning models. We also prove that the rate $O(1/\epsilon)$ is optimal for the subgradient method in the convex and interpolation setting.
****
Generating Adversarial Computer Programs using Optimized Obfuscations
Shashank Srikant,Sijia Liu,Tamara Mitrovska,Shiyu Chang,Quanfu Fan,Gaoyuan Zhang,Una-May O'Reilly
Machine learning (ML) models that learn and predict properties of computer programs are increasingly being adopted and deployed.
These models have demonstrated success in applications such as auto-completing code, summarizing large programs, and detecting bugs and malware in programs.
In this work, we investigate principled ways to adversarially perturb a computer program to fool such learned models, and thus determine their adversarial robustness. We use program obfuscations, which have conventionally been used to avoid attempts at reverse engineering programs, as adversarial perturbations. These perturbations modify programs in ways that do not alter their functionality but can be crafted to deceive an ML model when making a decision. We provide a general formulation for an adversarial program that allows applying multiple obfuscation transformations to a program in any language. We develop first-order optimization algorithms to efficiently determine two key aspects -- which parts of the program to transform, and what transformations to use. We show that it is important to optimize both these aspects to generate the best adversarially perturbed program. Due to the discrete nature of this problem, we also propose using randomized smoothing to improve the attack loss landscape to ease optimization.
We evaluate our work on Python and Java programs on the problem of program summarization.
We show that our best attack proposal achieves a 52% improvement over a state

-of-the-art attack generation approach for programs trained on a \textsc{seq2seq} model.
We further show that our formulation is better at training models that are robust to adversarial attacks.
****
Lie Algebra Convolutional Neural Networks with Automatic Symmetry Extraction
Nima Dehmamy,Yanchen Liu,Robin Walters,Rose Yu
Existing methods for incorporating symmetries into neural network architectures require prior knowledge of the symmetry group. We propose to learn the symmetries during the training of the group equivariant architectures. Our model, the Lie algebra convolutional network (L-conv), is based on infinitesimal generators of continuous groups and does not require discretization or integration over the group. We show that L-conv can approximate any group convolutional layer by composition of layers. We demonstrate how CNNs, Graph Convolutional Networks and fully-connected networks can all be expressed as an L-conv with appropriate groups. By allowing the infinitesimal generators to be learnable, L-conv can learn potential symmetries. We also show how the symmetries are related to the statistics of the dataset in linear settings. We find an analytical relationship between the symmetry group and a subgroup of an orthogonal group preserving the covariance of the input. Our experiments show that L-conv with trainable generators performs well on problems with hidden symmetries. Due to parameter sharing, L-conv also uses far fewer parameters than fully-connected layers.
****
C-Learning: Learning to Achieve Goals via Recursive Classification
Benjamin Eysenbach,Ruslan Salakhutdinov,Sergey Levine
We study the problem of predicting and controlling the future state distribution of an autonomous agent. This problem, which can be viewed as a reframing of goal-conditioned reinforcement learning (RL), is centered around learning a conditional probability density function over future states. Instead of directly estimating this density function, we indirectly estimate this density function by training a classifier to predict whether an observation comes from the future. Via Bayes' rule, predictions from our classifier can be transformed into predictions over future states. Importantly, an off-policy variant of our algorithm allows us to predict the future state distribution of a new policy, without collecting new experience. This variant allows us to optimize functionals of a policy's future state distribution, such as the density of reaching a particular goal state. While conceptually similar to Q-learning, our work lays a principled foundation for goal-conditioned RL as density estimation, providing justification for goal-conditioned methods used in prior work. This foundation makes hypotheses about Q-learning, including the optimal goal-sampling ratio, which we confirm experimentally. Moreover, our proposed method is competitive with prior goal-conditioned RL methods.
****
Off-Dynamics Reinforcement Learning: Training for Transfer with Domain Classifiers
Benjamin Eysenbach,Shreyas Chaudhari,Swapnil Asawa,Sergey Levine,Ruslan Salakhutdinov
We propose a simple, practical, and intuitive approach for domain adaptation in reinforcement learning. Our approach stems from the idea that the agent's experience in the source domain should look similar to its experience in the target domain. Building off of a probabilistic view of RL, we achieve this goal by compensating for the difference in dynamics by modifying the reward function. This modified reward function is simple to estimate by learning auxiliary classifiers that distinguish source-domain transitions from target-domain transitions. Intuitively, the agent is penalized for transitions that would indicate that the agent is interacting with the source domain, rather than the target domain. Formally, we prove that applying our method in the source domain is guaranteed to obtain a near-optimal policy for the target domain, provided that the source and target domains satisfy a lightweight assumption. Our approach is applicable to domains with continuous states and actions and does not require learning an explicit mod

el of the dynamics. On discrete and continuous control tasks, we illustrate the mechanics of our approach and demonstrate its scalability to high-dimensional~tasks.
****
Voting-based Approaches For Differentially Private Federated Learning
Yuqing Zhu,Xiang Yu,Yi-Hsuan Tsai,Francesco Pittaluga,Masoud Faraki,Manmohan Chandraker,Yu-Xiang Wang
While federated learning (FL) enables distributed agents to collaboratively train a centralized model without sharing data with each other, it fails to protect users against inference attacks that mine private information from the centralized model. Thus, facilitating federated learning methods with differential privacy (DPFL) becomes attractive. Existing algorithms based on privately aggregating clipped gradients require many rounds of communication, which may not converge, and cannot scale up to large-capacity models due to explicit dimension-dependence in its added noise. In this paper, we adopt the knowledge transfer model of private learning pioneered by Papernot et al. (2017; 2018) and extend their algorithm PATE, as well as the recent alternative PrivateKNN (Zhu et al., 2020) to the federated learning setting. The key difference is that our method privately aggregates the labels from the agents in a voting scheme, instead of aggregating the gradients, hence avoiding the dimension dependence and achieving signi■cant savings in communication cost. Theoretically, we show that when the margins of the voting scores are large, the agents enjoy exponentially higher accuracy and stronger (data-dependent) differential privacy guarantees on both agent-level and instance-level. Extensive experiments show that our approach signi■cantly improves the privacy-utility trade-off over the current state-of-the-art in DPFL.
****
Learning to Reach Goals via Iterated Supervised Learning
Dibya Ghosh,Abhishek Gupta,Ashwin Reddy,Justin Fu,Coline Manon Devin,Benjamin Eysenbach,Sergey Levine
Current reinforcement learning (RL) algorithms can be brittle and difficult to use, especially when learning goal-reaching behaviors from sparse rewards. Although supervised imitation learning provides a simple and stable alternative, it requires access to demonstrations from a human supervisor. In this paper, we study RL algorithms that use imitation learning to acquire goal reaching policies from scratch, without the need for expert demonstrations or a value function. In lieu of demonstrations, we leverage the property that any trajectory is a successful demonstration for reaching the final state in that same trajectory. We propose a simple algorithm in which an agent continually relabels and imitates the trajectories it generates to progressively learn goal-reaching behaviors from scratch. Each iteration, the agent collects new trajectories using the latest policy, and maximizes the likelihood of the actions along these trajectories under the goal that was actually reached, so as to improve the policy. We formally show that this iterated supervised learning procedure optimizes a bound on the RL objective, derive performance bounds of the learned policy, and empirically demonstrate improved goal-reaching performance and robustness over current RL algorithms in several benchmark tasks.
****
Model-Based Visual Planning with Self-Supervised Functional Distances
Stephen Tian,Suraj Nair,Frederik Ebert,Sudeep Dasari,Benjamin Eysenbach,Chelsea Finn,Sergey Levine
A generalist robot must be able to complete a variety of tasks in its environment. One appealing way to specify each task is in terms of a goal observation. However, learning goal-reaching policies with reinforcement learning remains a challenging problem, particularly when hand-engineered reward functions are not available. Learned dynamics models are a promising approach for learning about the environment without rewards or task-directed data, but planning to reach goals with such a model requires a notion of functional similarity between observations and goal states. We present a self-supervised method for model-based visual goal reaching, which uses both a visual dynamics model as well as a dynamical distance function learned using model-free reinforcement learning. Our approach learns

entirely using offline, unlabeled data, making it practical to scale to large and diverse datasets. In our experiments, we find that our method can successfully learn models that perform a variety of tasks at test-time, moving objects amid distractors with a simulated robotic arm and even learning to open and close a drawer using a real-world robot. In comparisons, we find that this approach substantially outperforms both model-free and model-based prior methods.
****

Iterative Amortized Policy Optimization

Joseph Marino,Alexandre Piché,Alessandro Davide Ialongo,Yisong Yue

Policy networks are a central feature of deep reinforcement learning (RL) algorithms for continuous control, enabling the estimation and sampling of high-value actions. From the variational inference perspective on RL, policy networks, when employed with entropy or KL regularization, are a form of amortized optimization, optimizing network parameters rather than the policy distributions directly. However, this direct amortized mapping can empirically yield suboptimal policy estimates and limited exploration. Given this perspective, we consider the more flexible class of iterative amortized optimizers. We demonstrate that the resulting technique, iterative amortized policy optimization, yields performance improvements over direct amortization methods on benchmark continuous control tasks.
****

Mathematical Reasoning via Self-supervised Skip-tree Training

Markus Norman Rabe,Dennis Lee,Kshitij Bansal,Christian Szegedy

We demonstrate that self-supervised language modeling applied to mathematical formulas enables logical reasoning. To measure the logical reasoning abilities of language models, we formulate several evaluation (downstream) tasks, such as inferring types, suggesting missing assumptions and completing equalities. For training language models for formal mathematics, we propose a novel skip-tree task. We find that models trained on the skip-tree task show surprisingly strong mathematical reasoning abilities, and outperform models trained on standard skip-sequence tasks. We also analyze the models' ability to formulate new conjectures by measuring how often the predictions are provable and useful in other proofs.
****

Fast Training of Contrastive Learning with Intermediate Contrastive Loss

Chengyue Gong,Xingchao Liu,qiang liu

Recently, representations learned by self-supervised approaches have significantly reduced the gap with their supervised counterparts in many different computer vision tasks. However, these self-supervised methods are computationally challenging. In this work, we focus on accelerating contrastive learning algorithms with little or even no loss of accuracy. Our insight is that, contrastive learning concentrates on optimizing similarity (dissimilarity) between pairs of inputs, and the similarity on the intermediate layers is a good surrogate of the final similarity. We exploit our observation by introducing additional intermediate contrastive losses. In this way, we can truncate the back-propagation and updates only a part of the parameters for each gradient descent update. Additionally, we do selection based on the intermediate losses to filter easy regions for each image, which further reduces the computational cost. We apply our method to recently-proposed MOCO, SimCLR, SwAV and notice that we can reduce the computational cost with little loss on the performance of ImageNet linear classification and other downstream tasks.
****

DeepAveragers: Offline Reinforcement Learning By Solving Derived Non-Parametric MDPs

Aayam Kumar Shrestha,Stefan Lee,Prasad Tadepalli,Alan Fern

We study an approach to offline reinforcement learning (RL) based on optimally solving finitely-represented MDPs derived from a static dataset of experience. This approach can be applied on top of any learned representation and has the potential to easily support multiple solution objectives as well as zero-shot adjustment to changing environments and goals. Our main contribution is to introduce the Deep Averagers with Costs MDP (DAC-MDP) and to investigate its solutions for offline RL. DAC-MDPs are a non-parametric model that can leverage dee

p representations and account for limited data by introducing costs for exploiti ng under-represented parts of the model. In theory, we show conditions that all ow for lower-bounding the performance of DAC-MDP solutions. We also investigate the empirical behavior in a number of environments, including those with image-b ased observations. Overall, the experiments demonstrate that the framework can w ork in practice and scale to large complex offline RL problems.
****
Multi-resolution modeling of a discrete stochastic process identifies causes of cancer
Adam Uri Yaari,Maxwell Sherman,Oliver Clarke Priebe,Po-Ru Loh,Boris Katz,Andrei Barbu,Bonnie Berger
Detection of cancer-causing mutations within the vast and mostly unexplored huma n genome is a major challenge. Doing so requires modeling the background mutatio n rate, a highly non-stationary stochastic process, across regions of interest v arying in size from one to millions of positions. Here, we present the split-Poi sson-Gamma (SPG) distribution, an extension of the classical Poisson-Gamma formu lation, to model a discrete stochastic process at multiple resolutions. We demon strate that the probability model has a closed-form posterior, enabling efficien t and accurate linear-time prediction over any length scale after the parameters of the model have been inferred a single time. We apply our framework to model mutation rates in tumors and show that model parameters can be accurately inferr ed from high-dimensional epigenetic data using a convolutional neural network, G aussian process, and maximum-likelihood estimation. Our method is both more accu rate and more efficient than existing models over a large range of length scales . We demonstrate the usefulness of multi-resolution modeling by detecting genomi c elements that drive tumor emergence and are of vastly differing sizes.
****
A Siamese Neural Network for Behavioral Biometrics Authentication
Jesús Solano,Esteban Rivera,Alejandra Castelblanco,Lizzy Tengana,Christian Lopez ,Martin Ochoa
The raise in popularity of personalized web and mobile applications brings about a need of robust authentication systems. Although password authentication is th e most popular authentication mechanism, it has also several drawbacks. Behavior al Biometrics Authentication has emerged as a complementary risk-based authentic ation approach which aims at profiling users based on their behavior while inter acting with computers/smartphones. In this work we propose a novel Siamese Neura l Network to perform a few-shot verification of user's behavior. We develop our approach to identify behavior from either human-computer or human-smartphone int eraction. For computer interaction our approach learns from mouse and keyboard d ynamics, while for smartphone interaction it learns from holding patterns and to uch patterns. We show that our approach has a few-shot classification accuracy o f up to 99.8% and 90.8% for mobile and web interactions, respectively. We also t est our approach on a database that contains over 100K different web interaction s collected in the wild.
****
Assisting the Adversary to Improve GAN Training
Andreas Munk,William Harvey,Frank Wood
Some of the most popular methods for improving the stability and performance of GANs involve constraining or regularizing the discriminator. In this paper we co nsider a largely overlooked regularization technique which we refer to as the Ad versary's Assistant (AdvAs). We motivate this using a different perspective to t hat of prior work. Specifically, we consider a common mismatch between theoretic al analysis and practice: analysis often assumes that the discriminator reaches its optimum on each iteration. In practice, this is essentially never true, ofte n leading to poor gradient estimates for the generator. To address this, AdvAs i s a theoretically motivated penalty imposed on the generator based on the norm o f the gradients used to train the discriminator. This encourages the generator t o move towards points where the discriminator is optimal. We demonstrate the eff ect of applying AdvAs to several GAN objectives, datasets and network architectu res. The results indicate a reduction in the mismatch between theory and practic

e and that AdvAs can lead to improvement of GAN training, as measured by FID sco
res.

****
Estimating Example Difficulty using Variance of Gradients
Chirag Agarwal,Sara Hooker
In machine learning, a question of great interest is understanding what examples
 are challenging for a model to classify. Identifying atypical examples helps in
form safe deployment of models, isolates examples that require further human ins
pection, and provides interpretability into model behavior. In this work, we pro
pose the Variance of Gradients (VOG) as a valuable and efficient proxy metric fo
r detecting outliers in the data distribution. We provide quantitative and quali
tative support that VOG is a meaningful way to rank data by difficulty and to su
rface a tractable subset of the most challenging examples for human-in-the-loop
auditing. Data points with high VOG scores are more difficult for the model to l
earn and over-index on examples that require memorization.
****
Parametric Copula-GP model for analyzing multidimensional neuronal and behaviora
l relationships
Nina Kudryashova,Theoklitos Amvrosiadis,Nathalie Dupuy,Nathalie Rochefort,Arno O
nken
One of the main challenges in current systems neuroscience is the analysis of hi
gh-dimensional neuronal and behavioral data that are characterized by different
statistics and timescales of the recorded variables. We propose a parametric cop
ula model which separates the statistics of the individual variables from their
dependence structure, and escapes the curse of dimensionality by using vine copu
la constructions. We use a Bayesian framework with Gaussian Process (GP) priors
over copula parameters, conditioned on a continuous task-related variable. We im
prove the flexibility of this method by 1) using non-parametric conditional (rat
her than unconditional) marginals; 2) linearly mixing copula elements with quali
tatively different tail dependencies. We validate the model on synthetic data an
d compare its performance in estimating mutual information against the commonly
used non-parametric algorithms. Our model provides accurate information estimate
s when the dependencies in the data match the parametric copulas used in our fra
mework. Moreover, even when the exact density estimation with a parametric model
 is not possible, our Copula-GP model is still able to provide reasonable inform
ation estimates, close to the ground truth and comparable to those obtained with
 a neural network estimator. Finally, we apply our framework to real neuronal an
d behavioral recordings obtained in awake mice. We demonstrate the ability of ou
r framework to 1) produce accurate and interpretable bivariate models for the an
alysis of inter-neuronal noise correlations or behavioral modulations; 2) expand
 to more than 100 dimensions and measure information content in the whole-popula
tion statistics. These results demonstrate that the Copula-GP framework is parti
cularly useful for the analysis of complex multidimensional relationships betwee
n neuronal, sensory and behavioral data.
****
Channel-Directed Gradients for Optimization of Convolutional Neural Networks
Dong Lao,Peihao Zhu,Peter Wonka,Ganesh Sundaramoorthi
We introduce optimization methods for convolutional neural networks that can be
used to improve existing gradient-based optimization in terms of generalization
error. The method requires only simple processing of existing stochastic gradien
ts, can be used in conjunction with any optimizer, and has only a linear overhea
d (in the number of parameters) compared to computation of the stochastic gradie
nt. The method works by computing the gradient of the loss function with respect
 to output-channel directed re-weighted L2 or Sobolev metrics, which has the eff
ect of smoothing components of the gradient across a certain direction of the pa
rameter tensor. We show that defining the gradients along the output channel dir
ection leads to a performance boost, while other directions can be detrimental.
We present the continuum theory of such gradients, its discretization, and appli
cation to deep networks. Experiments on benchmark datasets, several networks, an

d baseline optimizers show that optimizers can be improved in generalization err
or by simply computing the stochastic gradient with respect to output-channel di
rected metrics.
****
On the Theory of Implicit Deep Learning: Global Convergence with Implicit Layers
Kenji Kawaguchi
A deep equilibrium model uses implicit layers, which are implicitly defined thro
ugh an equilibrium point of an infinite sequence of computation. It avoids any e
xplicit computation of the infinite sequence by finding an equilibrium point dir
ectly via root-finding and by computing gradients via implicit differentiation.
In this paper, we analyze the gradient dynamics of deep equilibrium models with
nonlinearity only on weight matrices and non-convex objective functions of weigh
ts for regression and classification. Despite non-convexity, convergence to glob
al optimum at a linear rate is guaranteed without any assumption on the width of
 the models, allowing the width to be smaller than the output dimension and the
number of data points. Moreover, we prove a relation between the gradient dynami
cs of the deep implicit layer and the dynamics of trust region Newton method of
a shallow explicit layer. This mathematically proven relation along with our num
erical observation suggests the importance of understanding implicit bias of imp
licit layers and an open problem on the topic. Our proofs deal with implicit lay
ers, weight tying and nonlinearity on weights, and differ from those in the rela
ted literature.

****
Uncertainty Weighted Offline Reinforcement Learning
Yue Wu,Shuangfei Zhai,Nitish Srivastava,Joshua M. Susskind,Jian Zhang,Ruslan Sal
akhutdinov,Hanlin Goh
Offline Reinforcement Learning promises to learn effective policies from previou
sly-collected, static datasets without the need for exploration. However, existi
ng Q-learning and actor-critic based off-policy RL algorithms fail when bootstra
pping from out-of-distribution (OOD) actions or states. We hypothesize that a ke
y missing ingredient from the existing methods is a proper treatment of uncertai
nty in the offline setting. We propose Uncertainty Weighted Actor-Critic (UWAC),
 an algorithm that models the epistemic uncertainty to detect OOD state-action p
airs and down-weights their contribution in the training objectives accordingly.
 Implementation-wise, we adopt a practical and effective dropout-based uncertain
ty estimation method that introduces very little overhead over existing RL algor
ithms. Empirically, we observe that UWAC substantially improves model stability
during training. In addition, UWAC out-performs existing offline RL methods on a
 variety of competitive tasks, and achieves significant performance gains over t
he state-of-the-art baseline on datasets with sparse demonstrations collected fr
om human experts.
****
On the Critical Role of Conventions in Adaptive Human-AI Collaboration
Andy Shih,Arjun Sawhney,Jovana Kondic,Stefano Ermon,Dorsa Sadigh
Humans can quickly adapt to new partners in collaborative tasks (e.g. playing ba
sketball), because they understand which fundamental skills of the task (e.g. ho
w to dribble, how to shoot) carry over across new partners. Humans can also quic
kly adapt to similar tasks with the same partners by carrying over conventions t
hat they have developed (e.g. raising hand signals pass the ball), without learn
ing to coordinate from scratch. To collaborate seamlessly with humans, AI agents
 should adapt quickly to new partners and new tasks as well. However, current ap
proaches have not attempted to distinguish between the complexities intrinsic to
 a task and the conventions used by a partner, and more generally there has been
 little focus on leveraging conventions for adapting to new settings. In this wo
rk, we propose a learning framework that teases apart rule-dependent representat
ion from convention-dependent representation in a principled way. We show that,
under some assumptions, our rule-dependent representation is a sufficient statis
tic of the distribution over best-response strategies across partners. Using thi

s separation of representations, our agents are able to adapt quickly to new par
tners, and to coordinate with old partners on new tasks in a zero-shot manner. W
e experimentally validate our approach on three collaborative tasks varying in c
omplexity: a contextual multi-armed bandit, a block placing task, and the card g
ame Hanabi.
****

On the Predictability of Pruning Across Scales
Jonathan S Rosenfeld,Jonathan Frankle,Michael Carbin,Nir Shavit
We show that the error of iteratively-pruned networks empirically follows a scal
ing law with interpretable coefficients that depend on the architecture and task
. We functionally approximate the error of the pruned networks, showing that it
is predictable in terms of an invariant tying width, depth, and pruning level, s
uch that networks of vastly different sparsities are freely interchangeable. We
demonstrate the accuracy of this functional approximation over scales spanning o
rders of magnitude in depth, width, dataset size, and sparsity. We show that the
 scaling law functional form holds (generalizes) for large scale data (CIFAR-10,
 ImageNet), architectures (ResNets, VGGs) and iterative pruning algorithms (IMP,
 SynFlow). As neural networks become ever larger and more expensive to train, ou
r findings suggest a framework for reasoning conceptually and analytically about
 pruning.
****

A Lazy Approach to Long-Horizon Gradient-Based Meta-Learning
Muhammad Abdullah Jamal,Liqiang Wang,Boqing Gong
Gradient-based meta-learning relates task-specific models to a meta-model by gra
dients. By this design, an algorithm first optimizes the task-specific models by
 an inner loop and then backpropagates meta-gradients through the loop to update
 the meta-model. The number of inner-loop optimization steps has to be small (e.
g., one step) to avoid high-order derivatives, big memory footprints, and the ri
sk of vanishing or exploding meta-gradients. We propose an intuitive teacher-stu
dent scheme to enable the gradient-based meta-learning algorithms to explore lon
g horizons by the inner loop. The key idea is to employ a student network to exp
lore the search space of task-specific models adequately (e.g., by more than ten
 steps), and a teacher then takes a ``leap'' toward the regions probed by the st
udent. The teacher not only arrives at a high-quality model but also defines a l
ightweight computation graph for meta-gradients. Our approach is generic, as we
verify its effectiveness with four meta-learning algorithms over three tasks: fe
w-shot learning, long-tailed classification, and meta-attack.
****

WaveGrad: Estimating Gradients for Waveform Generation
Nanxin Chen,Yu Zhang,Heiga Zen,Ron J Weiss,Mohammad Norouzi,William Chan
This paper introduces WaveGrad, a conditional model for waveform generation whic
h estimates gradients of the data density. The model is built on prior work on s
core matching and diffusion probabilistic models. It starts from a Gaussian whit
e noise signal and iteratively refines the signal via a gradient-based sampler c
onditioned on the mel-spectrogram.
WaveGrad offers a natural way to trade inference speed for sample quality by adj
usting the number of refinement steps, and bridges the gap between non-autoregre
ssive and autoregressive models in terms of audio quality.
We find that it can generate high fidelity audio samples using as few as six ite
rations.
Experiments reveal WaveGrad to generate high fidelity audio, outperforming adver
sarial non-autoregressive baselines and matching a strong likelihood-based autor
egressive baseline using fewer sequential operations. Audio samples are availab
le at https://wavegrad.github.io/.
****

BUSTLE: Bottom-Up Program Synthesis Through Learning-Guided Exploration
Augustus Odena,Kensen Shi,David Bieber,Rishabh Singh,Charles Sutton,Hanjun Dai
Program synthesis is challenging largely because of the difficulty of search in
a large space of programs. Human programmers routinely tackle the task of writin
g complex programs by writing sub-programs and then analyzing their intermediate

results to compose them in appropriate ways. Motivated by this intuition, we present a new synthesis approach that leverages learning to guide a bottom-up search over programs. In particular, we train a model to prioritize compositions of intermediate values during search conditioned on a given set of input-output examples. This is a powerful combination because of several emergent properties. First, in bottom-up search, intermediate programs can be executed, providing semantic information to the neural network. Second, given the concrete values from those executions, we can exploit rich features based on recent work on property signatures. Finally, bottom-up search allows the system substantial flexibility in what order to generate the solution, allowing the synthesizer to build up a program from multiple smaller sub-programs. Overall, our empirical evaluation finds that the combination of learning and bottom-up search is remarkably effective, even with simple supervised learning approaches. We demonstrate the effectiveness of our technique on two datasets, one from the SyGuS competition and one of our own creation.

****

Increasing-Margin Adversarial (IMA) training to Improve Adversarial Robustness of Neural Networks

Linhai Ma,Liang Liang

Deep neural networks (DNNs), including convolutional neural networks, are known to be vulnerable to adversarial attacks, which may lead to disastrous consequences in life-critical applications. Adversarial samples are usually generated by attack algorithms and can also be induced by white noises, and therefore the threats are real. In this study, we propose a novel training method, named Increasing Margin Adversarial (IMA) Training, to improve DNN robustness against adversarial noises. During training, the IMA method increases the margins of training samples by moving the decision boundaries of the DNN model far away from the training samples to improve robustness. The IMA method is evaluated on six publicly available datasets (including a COVID-19 CT image dataset) under strong 100-PGD white-box adversarial attacks, and the results show that the proposed method significantly improved classification accuracy on noisy data while keeping a relatively high accuracy on clean data. We hope our approach may facilitate the development of robust DNN applications, especially for COVID-19 diagnosis using CT images.

****

A Critique of Self-Expressive Deep Subspace Clustering

Benjamin David Haeffele,Chong You,Rene Vidal

Subspace clustering is an unsupervised clustering technique designed to cluster data that is supported on a union of linear subspaces, with each subspace defining a cluster with dimension lower than the ambient space. Many existing formulations for this problem are based on exploiting the self-expressive property of linear subspaces, where any point within a subspace can be represented as linear combination of other points within the subspace. To extend this approach to data supported on a union of non-linear manifolds, numerous studies have proposed learning an embedding of the original data using a neural network which is regularized by a self-expressive loss function on the data in the embedded space to encourage a union of linear subspaces prior on the data in the embedded space. Here we show that there are a number of potential flaws with this approach which have not been adequately addressed in prior work. In particular, we show the model formulation is often ill-posed in that it can lead to a degenerate embedding of the data, which need not correspond to a union of subspaces at all and is poorly suited for clustering. We validate our theoretical results experimentally and also repeat prior experiments reported in the literature, where we conclude that a significant portion of the previously claimed performance benefits can be attributed to an ad-hoc post processing step rather than the deep subspace clustering model.

****

An Algorithm for Out-Of-Distribution Attack to Neural Network Encoder

Liang Liang,Linhai Ma,Linchen Qian,Jiasong Chen

Deep neural networks (DNNs), especially convolutional neural networks, have achi

eved superior performance on image classification tasks. However, such performance is only guaranteed if the input to a trained model is similar to the training samples, i.e., the input follows the probability distribution of the training set. Out-Of-Distribution (OOD) samples do not follow the distribution of training set, and therefore the predicted class labels on OOD samples become meaningless. Classi■cation-based methods have been proposed for OOD detection; however, in this study we show that this type of method has no theoretical guarantee and is practically breakable by our OOD Attack algorithm because of dimensionality reduction in the DNN models. We also show that Glow likelihood-based OOD detection is breakable as well.
****

Explaining by Imitating: Understanding Decisions by Interpretable Policy Learning
Alihan Hüyük,Daniel Jarrett,Cem Tekin,Mihaela van der Schaar
Understanding human behavior from observed data is critical for transparency and accountability in decision-making. Consider real-world settings such as healthcare, in which modeling a decision-maker’s policy is challenging—with no access to underlying states, no knowledge of environment dynamics, and no allowance for live experimentation. We desire learning a data-driven representation of decision- making behavior that (1) inheres transparency by design, (2) accommodates partial observability, and (3) operates completely offline. To satisfy these key criteria, we propose a novel model-based Bayesian method for interpretable policy learning (“Interpole”) that jointly estimates an agent’s (possibly biased) belief-update process together with their (possibly suboptimal) belief-action mapping. Through experiments on both simulated and real-world data for the problem of Alzheimer’s disease diagnosis, we illustrate the potential of our approach as an investigative device for auditing, quantifying, and understanding human decision-making behavior.
****

Predicting Video with VQVAE
Jacob C Walker,Ali Razavi,Aaron van den Oord
In recent years, the task of video prediction---forecasting future video given past video frames---has attracted attention in the research community. In this paper we propose a novel approach to this problem with Vector Quantized Variational AutoEncoders (VQ-VAE). With VQ-VAE we compress high-resolution videos into a hierarchical set of multi-scale discrete latent variables. Compared to pixels, this compressed latent space has dramatically reduced dimensionality, allowing us to apply scalable autoregressive generative models to predict video. In contrast to previous work that has largely emphasized highly constrained datasets, we focus
on very diverse, large-scale datasets such as Kinetics-600. We predict video at a higher resolution, 256$\times$256, than any other previous method to our knowledge. We further validate our approach against prior work via a crowdsourced human evaluation.
****

For self-supervised learning, Rationality implies generalization, provably
Yamini Bansal,Gal Kaplun,Boaz Barak
We prove a new upper bound on the generalization gap of classifiers that are obtained by first using self-supervision to learn a representation r of the training~data, and then fitting a simple (e.g., linear) classifier g to the labels. Specifically, we show that (under the assumptions described below) the generalization gap of such classifiers tends to zero if $\mathsf{C}(g) \ll n$, where $\mathsf{C}(g)$ is an appropriately-defined measure of the simple classifier g's complexity, and n is the number of training samples. We stress that our bound is independent of the complexity of the representation r.
We do not make any structural or conditional-independence assumptions on the representation-learning task, which can use the same training dataset that is later used for classification. Rather, we assume that the training procedure satisfies certain natural noise-robustness (adding small amount of label noise causes small degradation in performance) and rationality (getting the wrong label is no

t better than getting no label at all) conditions that widely hold across many s
tandard architectures.
We also conduct an extensive empirical study of the generalization gap and the q
uantities used in our assumptions for a variety of self-supervision based algori
thms, including SimCLR, AMDIM and BigBiGAN, on the CIFAR-10 and ImageNet datase
ts. We show that, unlike standard supervised classifiers, these algorithms displ
ay small generalization gap, and the bounds we prove on this gap are often non v
acuous.
****

Shape-Tailored Deep Neural Networks Using PDEs for Segmentation
Naeemullah Khan,Angira Sharma,Philip Torr,Ganesh Sundaramoorthi
We present Shape-Tailored Deep Neural Networks (ST-DNN). ST-DNN extend convoluti
onal networks, which aggregate data from fixed shape (square) neighbor-hoods to
compute descriptors, to be defined on arbitrarily shaped regions. This is useful
 for segmentation applications, where it is desired to have descriptors that agg
regate data only within regions of segmentation to avoid mixing data from differ
ent regions, otherwise, the descriptors are difficult to group to a unique regio
n. We formulate these descriptors through partial differential equations (PDE)
that naturally generalize convolution to arbitrary regions, and derive the metho
dology to jointly estimate the segmentation and ST-DNN descriptor. We also show
that ST-DNN inherit covariance to translations and rotations from the PDE, a nat
ural property of a segmentation method, which existing CNN based methods lack. S
T-DNN are 3-4 order of magnitude smaller than typical CNN. We empirically show t
hat they exceed segmentation performance compared to state-of-the-art CNN-based
descriptors using 2-3 orders smaller training sets on the texture segmentation p
roblem.
****

Fine-grained Synthesis of Unrestricted Adversarial Examples
Omid Poursaeed,Tianxing Jiang,Yordanos Abraham Goshu,Harry Yang,Serge Belongie,S
er-Nam Lim
We propose a novel approach for generating unrestricted adversarial examples by
manipulating fine-grained aspects of image generation. Unlike existing unrestric
ted attacks that typically hand-craft geometric transformations, we learn stylis
tic and stochastic modifications leveraging state-of-the-art generative models.
This allows us to manipulate an image in a controlled, fine-grained manner witho
ut being bounded by a norm threshold. Our approach can be used for targeted and
non-targeted unrestricted attacks on classification, semantic segmentation and o
bject detection models. Our attacks can bypass certified defenses, yet our adver
sarial images look indistinguishable from natural images as verified by human ev
aluation. Moreover, we demonstrate that adversarial training with our examples i
mproves performance of the model on clean images without requiring any modificat
ions to the architecture. We perform experiments on LSUN, CelebA-HQ and COCO-Stu
ff as high resolution datasets to validate efficacy of our proposed approach.
****

Directed Acyclic Graph Neural Networks
Veronika Thost,Jie Chen
Graph-structured data ubiquitously appears in science and engineering. Graph neu
ral networks (GNNs) are designed to exploit the relational inductive bias exhibi
ted in graphs; they have been shown to outperform other forms of neural networks
 in scenarios where structure information supplements node features. The most co
mmon GNN architecture aggregates information from neighborhoods based on message
 passing. Its generality has made it broadly applicable. In this paper, we focus
 on a special, yet widely used, type of graphs---DAGs---and inject a stronger in
ductive bias---partial ordering---into the neural network design. We propose the
 directed acyclic graph neural network, DAGNN, an architecture that processes in
formation according to the flow defined by the partial order. DAGNN can be consi
dered a framework that entails earlier works as special cases (e.g., models for
trees and models updating node representations recurrently), but we identify sev
eral crucial components that prior architectures lack. We perform comprehensive
experiments, including ablation studies, on representative DAG datasets (i.e., s

ource code, neural architectures, and probabilistic graphical models) and demons trate the superiority of DAGNN over simpler DAG architectures as well as general graph architectures.

****

The Traveling Observer Model: Multi-task Learning Through Spatial Variable Embed dings

Elliot Meyerson,Risto Miikkulainen

This paper frames a general prediction system as an observer traveling around a continuous space, measuring values at some locations, and predicting them at oth ers. The observer is completely agnostic about any particular task being solved; it cares only about measurement locations and their values. This perspective le ads to a machine learning framework in which seemingly unrelated tasks can be so lved by a single model, by embedding their input and output variables into a sha red space. An implementation of the framework is developed in which these variab le embeddings are learned jointly with internal model parameters. In experiments , the approach is shown to (1) recover intuitive locations of variables in space and time, (2) exploit regularities across related datasets with completely disj oint input and output spaces, and (3) exploit regularities across seemingly unre lated tasks, outperforming task-specific single-task models and multi-task learn ing alternatives. The results suggest that even seemingly unrelated tasks may or iginate from similar underlying processes, a fact that the traveling observer mo del can use to make better predictions.

****

Learning Hyperbolic Representations for Unsupervised 3D Segmentation

Joy Hsu,Jeffrey Gu,Gong Her Wu,Wah Chiu,Serena Yeung

There exists a need for unsupervised 3D segmentation on complex volumetric data, particularly when annotation ability is limited or discovery of new categories is desired. Using the observation that much of 3D volumetric data is innately hi erarchical, we propose learning effective representations of 3D patches for unsu pervised segmentation through a variational autoencoder (VAE) with a hyperbolic latent space and a proposed gyroplane convolutional layer, which better models t he underlying hierarchical structure within a 3D image. We also introduce a hier archical triplet loss and multi-scale patch sampling scheme to embed relationshi ps across varying levels of granularity. We demonstrate the effectiveness of our hyperbolic representations for unsupervised 3D segmentation on a hierarchical t oy dataset, BraTS whole tumor dataset, and cryogenic electron microscopy data.

****

Enhancing Visual Representations for Efficient Object Recognition during Online Distillation

Shashanka Venkataramanan,Bruce W McIntosh,Abhijit Mahalanobis

We propose ENVISE, an online distillation framework that ENhances VISual repres entations for Efficient object recognition. We are motivated by the observation that in many real-world scenarios, the probability of occurrence of all classes is not the same and only a subset of classes occur frequently. Exploiting this f act, we aim to reduce the computations of our framework by employing a binary st udent network (BSN) to learn the frequently occurring classes using the pseudo-l abels generated by the teacher network (TN) on an unlabeled image stream. To mai ntain overall accuracy, the BSN must also accurately determine when a rare (or u nknown) class is present in the image stream so that the TN can be used in such cases. To achieve this, we propose an attention triplet loss which ensures that the BSN emphasizes the same semantically meaningful regions of the image as the TN. When the prior class probabilities in the image stream vary, we demonstrate that the BSN adapts to the TN faster than the real-valued student network. We al so introduce Gain in Efficiency (GiE), a new metric which estimates the relative reduction in FLOPS based on the number of times the BSN and TN are used to proc ess the image stream. We benchmark CIFAR-100 and tiny-imagenet datasets by creat ing meaningful inlier (frequent) and outlier (rare) class pairs that mimic real- world scenarios. We show that ENVISE outperforms state-of-the-art (SOTA) outlier detection methods in terms of GiE, and also achieves greater separation between

inlier and outlier classes in the feature space.
****
How Important is the Train-Validation Split in Meta-Learning?
Yu Bai,Minshuo Chen,Pan Zhou,Tuo Zhao,Jason D. Lee,Sham M. Kakade,Huan Wang,Caiming Xiong
Meta-learning aims to perform fast adaptation on a new task through learning a "prior" from multiple existing tasks. A common practice in meta-learning is to perform a train-validation split where the prior adapts to the task on one split of the data, and the resulting predictor is evaluated on another split. Despite its prevalence, the importance of the train-validation split is not well understood either in theory or in practice, particularly in comparison to the more direct non-splitting method, which uses all the per-task data for both training and evaluation.
We provide a detailed theoretical study on the whether and when the train-validation split is helpful on the linear centroid meta-learning problem, in the asymptotic setting where the number of tasks goes to infinity. We show that the splitting method converges to the optimal prior as expected, whereas the non-splitting method does not in general without structural assumptions on the data. In contrast, if the data are generated from linear models (the realizable regime), we show that both the splitting and non-splitting methods converge to the optimal prior. Further, perhaps surprisingly, our main result shows that the non-splitting method achieves a strictly better asymptotic excess risk under this data distribution, even when the regularization parameter and split ratio are optimally tuned for both methods. Our results highlight that data splitting may not always be preferable, especially when the data is realizable by the model. We validate our theories by experimentally showing that the non-splitting method can indeed outperform the splitting method, on both simulations and real meta-learning tasks.
****
My Body is a Cage: the Role of Morphology in Graph-Based Incompatible Control
Vitaly Kurin,Maximilian Igl,Tim Rocktäschel,Wendelin Boehmer,Shimon Whiteson
Multitask Reinforcement Learning is a promising way to obtain models with better performance, generalisation, data efficiency, and robustness. Most existing work is limited to compatible settings, where the state and action space dimensions are the same across tasks. Graph Neural Networks (GNN) are one way to address incompatible environments, because they can process graphs of arbitrary size. They also allow practitioners to inject biases encoded in the structure of the input graph. Existing work in graph-based continuous control uses the physical morphology of the agent to construct the input graph, i.e., encoding limb features as node labels and using edges to connect the nodes if their corresponded limbs are physically connected.
In this work, we present a series of ablations on existing methods that show that morphological information encoded in the graph does not improve their performance. Motivated by the hypothesis that any benefits GNNs extract from the graph structure are outweighed by difficulties they create for message passing, we also propose Amorpheus, a transformer-based approach. Further results show that, while Amorpheus ignores the morphological information that GNNs encode, it nonetheless substantially outperforms GNN-based methods.
****
Video Prediction with Variational Temporal Hierarchies
Vaibhav Saxena,Jimmy Ba,Danijar Hafner
Deep learning has shown promise for accurately predicting high-dimensional video sequences. Existing video prediction models succeeded in generating sharp but often short video sequences. Toward improving long-term video prediction, we study hierarchical latent variable models with levels that process at different time scales. To gain insights into the representations of such models, we study the information stored at each level of the hierarchy via the KL divergence, predictive entropy, datasets of varying speed, and generative distributions. Our analysis confirms that faster changing details are generally captured by lower levels, while slower changing facts are remembered by higher levels. On synthetic datasets where common methods fail after 25 frames, we show that temporally abstract

latent variable models can make accurate predictions for up to 200 frames.
****
Incremental few-shot learning via vector quantization in deep embedded space
Kuilin Chen,Chi-Guhn Lee
The capability of incrementally learning new tasks without forgetting old ones is a challenging problem due to catastrophic forgetting. This challenge becomes greater when novel tasks contain very few labelled training samples. Currently, most methods are dedicated to class-incremental learning and rely on sufficient training data to learn additional weights for newly added classes. Those methods cannot be easily extended to incremental regression tasks and could suffer from severe overfitting when learning few-shot novel tasks. In this study, we propose a nonparametric method in deep embedded space to tackle incremental few-shot learning problems. The knowledge about the learned tasks are compressed into a small number of quantized reference vectors. The proposed method learns new tasks sequentially by adding more reference vectors to the model using few-shot samples in each novel task. For classification problems, we employ the nearest neighbor scheme to make classification on sparsely available data and incorporate intra-class variation, less forgetting regularization and calibration of reference vectors to mitigate catastrophic forgetting. In addition, the proposed learning vector quantization (LVQ) in deep embedded space can be customized as a kernel smoother to handle incremental few-shot regression tasks. Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods in incremental learning.
****
Variational Auto-Encoder Architectures that Excel at Causal Inference
Negar Hassanpour,Russell Greiner
This paper provides a generative approach for causal inference using data from observational studies. Inspired by the work of Kingma et al. (2014), we propose a sequence of three architectures (namely Series, Parallel, and Hybrid) that each incorporate their M1 and M2 models as building blocks. Each architecture is an improvement over the previous one in terms of estimating causal effect, culminating in the Hybrid model. The Hybrid model is designed to encourage decomposing the underlying factors of any observational dataset; this in turn, helps to accurately estimate all treatment outcomes. Our empirical results demonstrate the superiority of all three proposed architectures compared to both state-of-the-art discriminative as well as other generative approaches in the literature.
****
Delay-Tolerant Local SGD for Efficient Distributed Training
An Xu,Xiao Yan,Hongchang Gao,Heng Huang
The heavy communication for model synchronization is a major bottleneck for scaling up the distributed deep neural network training to many workers. Moreover, model synchronization can suffer from long delays in scenarios such as federated learning and geo-distributed training. Thus, it is crucial that the distributed training methods are both \textit{delay-tolerant} AND \textit{communication-efficient}. However, existing works cannot simultaneously address the communication delay and bandwidth constraint. To address this important and challenging problem, we propose a novel training framework OLCO\textsubscript{3} to achieve delay tolerance with a low communication budget by using stale information. OLCO\textsubscript{3} introduces novel staleness compensation and compression compensation to combat the influence of staleness and compression error. Theoretical analysis shows that OLCO\textsubscript{3} achieves the same sub-linear convergence rate as the vanilla synchronous stochastic gradient descent (SGD) method. Extensive experiments on deep learning tasks verify the effectiveness of OLCO\textsubscript{3} and its advantages over existing works.
****
Revisiting Few-sample BERT Fine-tuning
Tianyi Zhang,Felix Wu,Arzoo Katiyar,Kilian Q Weinberger,Yoav Artzi
This paper is a study of fine-tuning of BERT contextual representations, with focus on commonly observed instabilities in few-sample scenarios. We identify several factors that cause this instability: the common use of a non-standard optimi

zation method with biased gradient estimation; the limited applicability of significant parts of the BERT network for down-stream tasks; and the prevalent practice of using a pre-determined, and small number of training iterations. We empirically test the impact of these factors, and identify alternative practices that resolve the commonly observed instability of the process. In light of these observations, we re-visit recently proposed methods to improve few-sample fine-tuning with BERT and re-evaluate their effectiveness. Generally, we observe the impact of these methods diminishes significantly with our modified process.
****
WAFFLe: Weight Anonymized Factorization for Federated Learning
Weituo Hao,Nikhil Mehta,Kevin J Liang,Pengyu Cheng,Mostafa El-Khamy,Lawrence Carin
In domains where data are sensitive or private, there is great value in methods that can learn in a distributed manner without the data ever leaving the local devices. In light of this need, federated learning has emerged as a popular training paradigm. However, many federated learning approaches trade transmitting data for communicating updated weight parameters for each local device. Therefore, a successful breach that would have otherwise directly compromised the data instead grants whitebox access to the local model, which opens the door to a number of attacks, including exposing the very data federated learning seeks to protect. Additionally, in distributed scenarios, individual client devices commonly exhibit high statistical heterogeneity. Many common federated approaches learn a single global model; while this may do well on average, performance degrades when the i.i.d. assumption is violated, underfitting individuals further from the mean and raising questions of fairness. To address these issues, we propose Weight Anonymized Factorization for Federated Learning (WAFFLe), an approach that combines the Indian Buffet Process with a shared dictionary of weight factors for neural networks. Experiments on MNIST, FashionMNIST, and CIFAR-10 demonstrate WAFFLe's significant improvement to local test performance and fairness while simultaneously providing an extra layer of security.
****
Alpha Net: Adaptation with Composition in Classifier Space
Nadine Chang,Jayanth Koushik,Michael Tarr,Martial Hebert,Yu-Xiong Wang
Deep learning classification models typically train poorly on classes with small numbers of examples. Motivated by the human ability to solve this task, models have been developed that transfer knowledge from classes with many examples to learn classes with few examples. Critically, the majority of these models transfer knowledge within model feature space. In this work, we demonstrate that transferring knowledge within classifier space is more effective and efficient. Specifically, by linearly combining strong nearest neighbor classifiers along with a weak classifier, we are able to compose a stronger classifier. Uniquely, our model can be implemented on top of any existing classification model that includes a classifier layer. We showcase the success of our approach in the task of long-tailed recognition, whereby the classes with few examples, otherwise known as the tail classes, suffer the most in performance and are the most challenging classes to learn. Using classifier-level knowledge transfer, we are able to drastically improve - by a margin as high as 10.5% - the state-of-the-art performance on the tail categories.
****
MDP Playground: Controlling Orthogonal Dimensions of Hardness in Toy Environments
Raghu Rajan,Jessica Lizeth Borja Diaz,Suresh Guttikonda,Fabio Ferreira,André Biedenkapp,Frank Hutter
We present MDP Playground, an efficient benchmark for Reinforcement Learning (RL) algorithms with various dimensions of hardness that can be controlled independently to challenge algorithms in different ways and to obtain varying degrees of hardness in generated environments. We consider and allow control over a wide variety of key hardness dimensions, including delayed rewards, rewardable sequences, sparsity of rewards, stochasticity, image representations, irrelevant features, time unit, and action max. While it is very time consuming to run RL algorit

hms on standard benchmarks, we define a parameterised collection of fast-to-run toy benchmarks in OpenAI Gym by varying these dimensions. Despite their toy nature and low compute requirements, we show that these benchmarks present substantial challenges to current RL algorithms. Furthermore, since we can generate environments with a desired value for each of the dimensions, in addition to having fine-grained control over the environments' hardness, we also have the ground truth available for evaluating algorithms. Finally, we evaluate the kinds of transfer for these dimensions that may be expected from our benchmarks to more complex benchmarks. We believe that MDP Playground is a valuable testbed for researchers designing new, adaptive and intelligent RL algorithms and those wanting to unit test their algorithms.

****

Linear Convergent Decentralized Optimization with Compression
Xiaorui Liu,Yao Li,Rongrong Wang,Jiliang Tang,Ming Yan
Communication compression has become a key strategy to speed up distributed optimization. However, existing decentralized algorithms with compression mainly focus on compressing DGD-type algorithms. They are unsatisfactory in terms of convergence rate, stability, and the capability to handle heterogeneous data. Motivated by primal-dual algorithms, this paper proposes the first \underline{L}in\underline{EA}r convergent \underline{D}ecentralized algorithm with compression, LEAD. Our theory describes the coupled dynamics of the inexact primal and dual update as well as compression error, and we provide the first consensus error bound in such settings without assuming bounded gradients. Experiments on convex problems validate our theoretical analysis, and empirical study on deep neural nets shows that LEAD is applicable to non-convex problems.
****

Learning from Demonstration with Weakly Supervised Disentanglement
Yordan Hristov,Subramanian Ramamoorthy
Robotic manipulation tasks, such as wiping with a soft sponge, require control from multiple rich sensory modalities. Human-robot interaction, aimed at teach- ing robots, is difficult in this setting as there is potential for mismatch between human and machine comprehension of the rich data streams. We treat the task of interpretable learning from demonstration as an optimisation problem over a probabilistic generative model. To account for the high-dimensionality of the data, a high-capacity neural network is chosen to represent the model. The latent variables in this model are explicitly aligned with high-level notions and concepts that are manifested in a set of demonstrations. We show that such alignment is best achieved through the use of labels from the end user, in an appropriately restricted vocabulary, in contrast to the conventional approach of the designer picking a prior over the latent variables. Our approach is evaluated in the context of two table-top robot manipulation tasks performed by a PR2 robot – that of dabbing liquids with a sponge (forcefully pressing a sponge and moving it along a surface) and pouring between different containers. The robot provides visual information, arm joint positions and arm joint efforts. We have made videos of the tasks and data available - see supplementary materials at: https://sites.google.com/view/weak-label-lfd.
****

Incorporating Symmetry into Deep Dynamics Models for Improved Generalization
Rui Wang,Robin Walters,Rose Yu
Recent work has shown deep learning can accelerate the prediction of physical dynamics relative to numerical solvers. However, limited physical accuracy and an inability to generalize under distributional shift limit its applicability to the real world. We propose to improve accuracy and generalization by incorporating symmetries into convolutional neural networks. Specifically, we employ a variety of methods each tailored to enforce a different symmetry. Our models are both theoretically and experimentally robust to distributional shift by symmetry group transformations and enjoy favorable sample complexity. We demonstrate the advantage of our approach on a variety of physical dynamics including Rayleigh–Bénard convection and real-world ocean currents and temperatures. Compare with image

or text applications, our work is a significant step towards applying equivarian t neural networks to high-dimensional systems with complex dynamics.
****
Ensemble-based Adversarial Defense Using Diversified Distance Mapping
Ehsan Kazemi,Mohamed E. Hussein,Wael AbdAlmgaeed
We propose an ensemble-based defense against adversarial examples using distance map layers (DMLs). Similar to fully connected layers, DMLs can be used to outpu t logits for a multi-class classification model. We show in this paper how DMLs can be deployed to prevent transferability of attacks across ensemble members b y adapting pairwise (almost) orthogonal covariance matrices. We also illustrate how DMLs provide an efficient way to regularize the Lipschitz constant of the en semble's member models, which further boosts the resulting robustness. Through e mpirical evaluations across multiple datasets and attack models, we demonstrate that the ensembles based on DMLs can achieve high benign accuracy while exhibiti ng robustness against adversarial attacks using multiple white-box techniques al ong with AutoAttack.
****
Contrastive Learning with Stronger Augmentations
Xiao Wang,Guo-Jun Qi
Representation learning has been greatly improved with the advance of contrastiv e learning methods with the performance being closer to their supervised learnin g counterparts. Those methods have greatly benefited from various data augmentat ions that are carefully designated to maintain their identities so that the imag es transformed from the same instance can still be retrieved. Although stronger augmentations could expose novel patterns of representations to improve their ge neralizability, directly using stronger augmentations in instance discrimination -based contrastive learning may even deteriorate the performance, because the di stortions induced from the stronger augmentations could ridiculously change the image structures and thus the transformed images cannot be viewed as the same as the original ones any more. Additional efforts are needed for us to explore the role of the stronger augmentations in further pushing the performance of unsupe rvised learning to the fully supervised upper bound. Instead of applying the str onger augmentations directly to minimize the contrastive loss, we propose to min imize the distribution divergence between the weakly and strongly augmented imag es over the representation bank to supervise the retrieval of strongly augmented queries from a pool of candidates. This avoids an overoptimistic assumption tha t could overfit the strongly augmented queries containing distorted visual struc tures into the positive targets in the representation bank, while still being ab le to distinguish them from the negative samples by leveraging the distributions of weakly augmented counterparts. The proposed method achieves top-1 accuracy o f 76.2% on ImageNet with a standard ResNet-50 architecture with a single-layer c lassifier fine-tuned. This is almost the same as 76.5% of top-1 accuracy with a fully supervised ResNet-50. Moreover, it outperforms the previous self-supervise d and supervised methods on both the transfer learning and object detection task s.

****
The Risks of Invariant Risk Minimization
Elan Rosenfeld,Pradeep Kumar Ravikumar,Andrej Risteski
Invariant Causal Prediction (Peters et al., 2016) is a technique for out-of-dist ribution generalization which assumes that some aspects of the data distribution vary across the training set but that the underlying causal mechanisms remain c onstant. Recently, Arjovsky et al. (2019) proposed Invariant Risk Minimization (IRM), an objective based on this idea for learning deep, invariant features of d ata which are a complex function of latent variables; many alternatives have sub sequently been suggested. However, formal guarantees for all of these works are severely lacking. In this paper, we present the first analysis of classificat ion under the IRM objective—as well as these recently proposed alternatives—unde r a fairly natural and general model. In the linear case, we show simple conditi ons under which the optimal solution succeeds or, more often, fails to recover t

he optimal invariant predictor. We furthermore present the very first results in the non-linear regime: we demonstrate that IRM can fail catastrophically unless the test data is sufficiently similar to the training distribution—this is precisely the issue that it was intended to solve. Thus, in this setting we find that IRM and its alternatives fundamentally do not improve over standard Empirical Risk Minimization.
****
Scaling Symbolic Methods using Gradients for Neural Model Explanation
Subham Sekhar Sahoo,Subhashini Venugopalan,Li Li,Rishabh Singh,Patrick Riley
Symbolic techniques based on Satisfiability Modulo Theory (SMT) solvers have been proposed for analyzing and verifying neural network properties, but their usage has been fairly limited owing to their poor scalability with larger networks. In this work, we propose a technique for combining gradient-based methods with symbolic techniques to scale such analyses and demonstrate its application for model explanation. In particular, we apply this technique to identify minimal regions in an input that are most relevant for a neural network's prediction. Our approach uses gradient information (based on Integrated Gradients) to focus on a subset of neurons in the first layer, which allows our technique to scale to large networks. The corresponding SMT constraints encode the minimal input mask discovery problem such that after masking the input, the activations of the selected neurons are still above a threshold. After solving for the minimal masks, our approach scores the mask regions to generate a relative ordering of the features within the mask. This produces a saliency map which explains" where a model is looking" when making a prediction. We evaluate our technique on three datasets-MNIST, ImageNet, and Beer Reviews, and demonstrate both quantitatively and qualitatively that the regions generated by our approach are sparser and achieve higher saliency scores compared to the gradient-based methods alone. Code and examples are at - https://github.com/google-research/google-research/tree/master/smug_saliency
****
Contextual Dropout: An Efficient Sample-Dependent Dropout Module
XINJIE FAN,Shujian Zhang,Korawat Tanwisuth,Xiaoning Qian,Mingyuan Zhou
Dropout has been demonstrated as a simple and effective module to not only regularize the training process of deep neural networks, but also provide the uncertainty estimation for prediction. However, the quality of uncertainty estimation is highly dependent on the dropout probabilities. Most current models use the same dropout distributions across all data samples due to its simplicity. Despite the potential gains in the flexibility of modeling uncertainty, sample-dependent dropout, on the other hand, is less explored as it often encounters scalability issues or involves non-trivial model changes. In this paper, we propose contextual dropout with an efficient structural design as a simple and scalable sample-dependent dropout module, which can be applied to a wide range of models at the expense of only slightly increased memory and computational cost. We learn the dropout probabilities with a variational objective, compatible with both Bernoulli dropout and Gaussian dropout. We apply the contextual dropout module to various models with applications to image classification and visual question answering and demonstrate the scalability of the method with large-scale datasets, such as ImageNet and VQA 2.0. Our experimental results show that the proposed method outperforms baseline methods in terms of both accuracy and quality of uncertainty estimation.
****
Empirically Verifying Hypotheses Using Reinforcement Learning
Kenneth Marino,Rob Fergus,Arthur Szlam,Abhinav Gupta
This paper formulates hypothesis verification as an RL problem. Specifically, we aim to build an agent that, given a hypothesis about the dynamics of the world, can take actions to generate observations which can help predict whether the hypothesis is true or false. Existing RL algorithms fail to solve this task, even for simple environments.
In order to train the agents, we exploit the underlying structure of many hypotheses, factorizing them as {pre-condition, action sequence, post-condition} tripl

ets. By leveraging this structure we show that RL agents are able to succeed at the task. Furthermore, subsequent fine-tuning of the policies allows the agent to correctly verify hypotheses not amenable to the above factorization.
****

TraDE: A Simple Self-Attention-Based Density Estimator
Rasool Fakoor,Pratik Anil Chaudhari,Jonas Mueller,Alex Smola
We present TraDE, a self-attention-based architecture for auto-regressive density estimation with continuous and discrete valued data. Our model is trained using a penalized maximum likelihood objective, which ensures that samples from the density estimate resemble the training data distribution. The use of self-attention means that the model need not retain conditional sufficient statistics during the auto-regressive process beyond what is needed for each covariate. On standard tabular and image data benchmarks, TraDE produces significantly better density estimates than existing approaches such as normalizing flow estimators and recurrent auto-regressive models. However log-likelihood on held-out data only partially reflects how useful these estimates are in real-world applications. In order to systematically evaluate density estimators, we present a suite of tasks such as regression using generated samples, out-of-distribution detection, and robustness to noise in the training data and demonstrate that TraDE works well in these scenarios.
****

Disentangling style and content for low resource video domain adaptation: a case study on keystroke inference attacks
John Lim,Fabian Monrose,Jan-Michael Frahm
Keystroke inference attacks are a form of side-channels attacks in which an attacker leverages various techniques to recover a user’s keystrokes as she inputs information into some display (for example, while sending a text message or entering her pin). Typically, these attacks leverage machine learning approaches, but assessing the realism of the threat space has lagged behind the pace of machine learning advancements, due in-part, to the challenges in curating large real-life datasets. This paper aims to overcome the challenge of having limited number of real data by introducing a video domain adaptation technique that is able to leverage synthetic data through supervised disentangled learning. Specifically, for a given domain, we decompose the observed data into two factors of variation: Style and Content. Doing so provides four learned representations: real-life style, synthetic style, real-life content and synthetic content. Then, we combine them into feature representations from all combinations of style-content pairings across domains, and train a model on these combined representations to classify the content (i.e., labels) of a given datapoint in the style of another domain. We evaluate our method on real-life data using a variety of metrics to quantify the amount of information an attacker is able to recover. We show that our method prevents our model from overfitting to a small real-life training set, indicating that our method is an effective form of data augmentation. Code and data will be released after reviewal.
****

RG-Flow: A hierarchical and explainable flow model based on renormalization group and sparse prior
Hong-Ye Hu,Dian Wu,Yi-Zhuang You,Bruno Olshausen,Yubei Chen
Flow-based generative models have become an important class of unsupervised learning approaches. In this work, we incorporate the key idea of renormalization group (RG) and sparse prior distribution to design a hierarchical flow-based generative model, called RG-Flow, which can separate different scale information of images with disentangle representations at each scale. We demonstrate our method mainly on the CelebA dataset and show that the disentangled representation at different scales enables semantic manipulation and style mixing of the images. To visualize the latent representation, we introduce the receptive fields for flow-based models and find receptive fields learned by RG-Flow are similar to convolutional neural networks. In addition, we replace the widely adopted Gaussian prior distribution by sparse prior distributions to further enhance the disentanglement of representations. From a theoretical perspective, the proposed method has

$O(\log L)$ complexity for image inpainting compared to previous flow-based models with $O(L^2)$ complexity.
****
Contrastive Learning with Hard Negative Samples
Joshua David Robinson,Ching-Yao Chuang,Suvrit Sra,Stefanie Jegelka
We consider the question: how can you sample good negative examples for contrastive learning? We argue that, as with metric learning, learning contrastive representations benefits from hard negative samples (i.e., points that are difficult to distinguish from an anchor point). The key challenge toward using hard negatives is that contrastive methods must remain unsupervised, making it infeasible to adopt existing negative sampling strategies that use label information. In response, we develop a new class of unsupervised methods for selecting hard negative samples where the user can control the amount of hardness. A limiting case of this sampling results in a representation that tightly clusters each class, and pushes different classes as far apart as possible. The proposed method improves downstream performance across multiple modalities, requires only few additional lines of code to implement, and introduces no computational overhead.

****
Debiasing Concept-based Explanations with Causal Analysis
Mohammad Taha Bahadori,David Heckerman
Concept-based explanation approach is a popular model interpertability tool because it expresses the reasons for a model's predictions in terms of concepts that are meaningful for the domain experts. In this work, we study the problem of the concepts being correlated with confounding information in the features. We propose a new causal prior graph for modeling the impacts of unobserved variables and a method to remove the impact of confounding information and noise using a two-stage regression technique borrowed from the instrumental variable literature. We also model the completeness of the concepts set and show that our debiasing method works when the concepts are not complete. Our synthetic and real-world experiments demonstrate the success of our method in removing biases and improving the ranking of the concepts in terms of their contribution to the explanation of the predictions.
****
Reintroducing Straight-Through Estimators as Principled Methods for Stochastic Binary Networks
Alexander Shekhovtsov,Viktor Yanush
Training neural networks with binary weights and activations is a challenging problem due to the lack of gradients and difficulty of optimization over discrete weights.
Many successful experimental results have been achieved with empirical straight-through (ST) approaches, proposing a variety of ad-hoc rules for propagating gradients through non-differentiable activations and updating discrete weights. At the same time, ST methods can be truly derived as estimators in the stochastic binary network (SBN) model with Bernoulli weights. We advance these derivations to a more complete and systematic study. We analyze properties, estimation accuracy, obtain different forms of correct ST estimators for activations and weights, explain existing empirical approaches and their shortcomings, explain how latent weights arise from the mirror descent method when optimizing over probabilities. This allows to reintroduce, once empirical, ST methods as sound approximations, apply them with clarity and develop further improvements.
****
Selecting Treatment Effects Models for Domain Adaptation Using Causal Knowledge
Trent Kyono,Ioana Bica,Zhaozhi Qian,Mihaela van der Schaar
Selecting causal inference models for estimating individualized treatment effects (ITE) from observational data presents a unique challenge since the counterfactual outcomes are never observed. The problem is challenged further in the unsupervised domain adaptation (UDA) setting where we only have access to labeled samples in the source domain, but desire selecting a model that achieves good performance on a target domain for which only unlabeled samples are available. Existi

ng techniques for UDA model selection are designed for the predictive setting. T
hese methods examine discriminative density ratios between the input covariates
in the source and target domain and do not factor in the model's predictions in
the target domain. Because of this, two models with identical performance on the
 source domain would receive the same risk score by existing methods, but in rea
lity, have significantly different performance in the test domain. We leverage t
he invariance of causal structures across domains to propose a novel model selec
tion metric specifically designed for ITE methods under the UDA setting. In part
icular, we propose selecting models whose predictions of interventions' effects
satisfy known causal structures in the target domain. Experimentally, our method
 selects ITE models that are more robust to covariate shifts on several healthca
re datasets, including estimating the effect of ventilation in COVID-19 patients
 from different geographic locations.
****

Self-training For Few-shot Transfer Across Extreme Task Differences

Cheng Perng Phoo,Bharath Hariharan

Most few-shot learning techniques are pre-trained on a large, labeled “base data
set”. In problem domains where such large labeled datasets are not available for
 pre-training (e.g., X-ray, satellite images), one must resort to pre-training i
n a different “source” problem domain (e.g., ImageNet), which can be very differ
ent from the desired target task. Traditional few-shot and transfer learning tec
hniques fail in the presence of such extreme differences between the source and
target tasks. In this paper, we present a simple and effective solution to tackl
e this extreme domain gap: self-training a source domain representation on unlab
eled data from the target domain. We show that this improves one-shot performanc
e on the target domain by 2.9 points on average on the challenging BSCD-FSL benc
hmark consisting of datasets from multiple domains.
****

An Examination of Preference-based Reinforcement Learning for Treatment Recommen
dation

Nan Xu,Nitin Kamra,Yan Liu

Treatment recommendation is a complex multi-faceted problem with many conflictin
g objectives, e.g., optimizing the survival rate (or expected lifetime), mitigat
ing negative impacts, reducing financial expenses and time costs, avoiding over-
treatment, etc. While this complicates the hand-engineering of a reward function
 for learning treatment policies, fortunately, qualitative feedback from human e
xperts is readily available and can be easily exploited. Since direct estimatio
n of rewards via inverse reinforcement learning is a challenging task and requir
es the existence of an optimal human policy, the field of treatment recommendati
on has recently witnessed the development of the preference-based ReinforcementL
earning (PRL) framework, which infers a reward function from only qualitative an
d imperfect human feedback to ensure that a human expert’s preferred policy has
a higher expected return over a less preferred policy. In this paper, we first p
resent an open simulation platform to model the progression of two diseases, nam
ely Cancer and Sepsis, and the reactions of the affected individuals to the rece
ived treatment. Secondly, we investigate important problems in adopting preferen
ce-basedRL approaches for treatment recommendation, such as advantages of learni
ng from preference over hand-engineered reward, addressing incomparable policies
, reward interpretability, and agent design via simulated experiments. The desig
ned simulation platform and insights obtained for preference-based RL approaches
 are beneficial for achieving the right trade-off between various human objectiv
es during treatment recommendation.
****

Risk-Averse Offline Reinforcement Learning

Núria Armengol Urpí,Sebastian Curi,Andreas Krause

Training Reinforcement Learning (RL) agents in high-stakes applications might be
 too prohibitive due to the risk associated to exploration. Thus, the agent can
only use data previously collected by safe policies. While previous work conside
rs optimizing the average performance using offline data, we focus on optimizing
 a risk-averse criteria, namely the CVaR. In particular, we present the Offline

Risk-Averse Actor-Critic (O-RAAC), a model-free RL algorithm that is able to lea rn risk-averse policies in a fully offline setting. We show that O-RAAC learns p olicies with higher CVaR than risk-neutral approaches in different robot control tasks. Furthermore, considering risk-averse criteria guarantees distributional robustness of the average performance with respect to particular distribution sh ifts. We demonstrate empirically that in the presence of natural distribution-sh ifts, O-RAAC learns policies with good average performance.

****
The shape and simplicity biases of adversarially robust ImageNet-trained CNNs
Peijie Chen,Chirag Agarwal,Anh Nguyen
Adversarial training has been the topic of dozens of studies and a leading metho d for defending against adversarial attacks.
Yet, it remains largely unknown (a) how adversarially-robust ImageNet classifier s (R classifiers) generalize to out-of-distribution examples; and (b) how their generalization capability relates to their hidden representations. In this paper , we perform a thorough, systematic study to answer these two questions across A lexNet, GoogLeNet, and ResNet-50 architectures. We found that while standard Ima geNet classifiers have a strong texture bias, their R counterparts rely heavily on shapes. Remarkably, adversarial training induces three simplicity biases into hidden neurons in the process of ``"robustifying" networks. That is, each convo lutional neuron in R networks often changes to detecting (1) pixel-wise smoother patterns i.e. a mechanism that blocks high-frequency noise from passing through the network; (2) more lower-level features i.e. textures and colors (instead of objects); and (3) fewer types of inputs. Our findings reveal the interesting me chanisms that made networks more adversarially robust and also explain some rece nt findings e.g. why R networks benefit from much larger capacity and can act as a strong image prior in image synthesis.
****
Trojans and Adversarial Examples: A Lethal Combination
Guanxiong Liu,Issa Khalil,Abdallah Khreishah,Hai Phan
In this work, we naturally unify adversarial examples and Trojan backdoors into a new stealthy attack, that is activated only when 1) adversarial perturbation i s injected into the input examples and 2) a Trojan backdoor is used to poison th e training process simultaneously. Different from traditional attacks, we levera ge adversarial noise in the input space to move Trojan-infected examples across the model decision boundary, thus making it difficult to be detected. Our attack can fool the user into accidentally trusting the infected model as a robust cla ssifier against adversarial examples. We perform a thorough analysis and conduct an extensive set of experiments on several benchmark datasets to show that our attack can bypass existing defenses with a success rate close to 100%.
****
Parameter Efficient Multimodal Transformers for Video Representation Learning
Sangho Lee,Youngjae Yu,Gunhee Kim,Thomas Breuel,Jan Kautz,Yale Song
The recent success of Transformers in the language domain has motivated adapting it to a multimodal setting, where a new visual model is trained in tandem with an already pretrained language model. However, due to the excessive memory requi rements from Transformers, existing work typically fixes the language model and train only the vision module, which limits its ability to learn cross-modal info rmation in an end-to-end manner. In this work, we focus on reducing the paramete rs of multimodal Transformers in the context of audio-visual video representatio n learning. We alleviate the high memory requirement by sharing the parameters o f Transformers across layers and modalities; we decompose the Transformer into m odality-specific and modality-shared parts so that the model learns the dynamics of each modality both individually and together, and propose a novel parameter sharing scheme based on low-rank approximation. We show that our approach reduce s parameters of the Transformers up to 97%, allowing us to train our model end-t o-end from scratch. We also propose a negative sampling approach based on an ins tance similarity measured on the CNN embedding space that our model learns toget her with the Transformers. To demonstrate our approach, we pretrain our model on

30-second clips (480 frames) from Kinetics-700 and transfer it to audio-visual classification tasks.
****
Characterizing Structural Regularities of Labeled Data in Overparameterized Models
Ziheng Jiang,Chiyuan Zhang,Kunal Talwar,Michael Curtis Mozer
Humans are accustomed to environments that contain both regularities and exceptions. For example, at most gas stations, one pays prior to pumping, but the occasional rural station does not accept payment in advance.
Likewise, deep neural networks can generalize across instances that share common patterns or structures yet have the capacity to memorize rare or irregular forms. We analyze how individual instances are treated by a model via a consistency score. The score characterizes the expected accuracy for a held-out instance given training sets of varying size sampled from the data distribution. We obtain empirical estimates of this score for individual instances in multiple data-sets, and we show that the score identifies out-of-distribution and mislabeled examples at one end of the continuum and strongly regular examples at the other end. We identify computationally inexpensive proxies to the consistency score using statistics collected during training. We apply the score toward understanding the dynamics of representation learning and to filter outliers during training, and we discuss other potential applications including curriculum learning and active data collection.
****
Universal Approximation Theorem for Equivariant Maps by Group CNNs
Wataru Kumagai,Akiyoshi Sannai
Group symmetry is inherent in a wide variety of data distributions. Data processing that preserves symmetry is described as an equivariant map and often effective in achieving high performance. Convolutional neural networks (CNNs) have been known as models with equivariance and shown to approximate equivariant maps for some specific groups. However, universal approximation theorems for CNNs have been separately derived with individual techniques according to each group and setting. This paper provides a unified method to obtain universal approximation theorems for equivariant maps by CNNs in various settings. As its significant advantage, we can handle non-linear equivariant maps between infinite-dimensional spaces for non-compact groups.
****
Defending against black-box adversarial attacks with gradient-free trained sign activation neural networks
Yunzhe Xue,Meiyan Xie,Zhibo Yang,Usman Roshan
While machine learning models today can achieve high accuracies on classification tasks, they can be deceived by minor imperceptible distortions to the data. These are known as adversarial attacks and can be lethal in the black-box setting which does not require knowledge of the target model type or its parameters. Binary neural networks that have sign activation and are trained with gradient descent have been shown to be harder to attack than conventional sigmoid activation networks but their improvements are marginal. We instead train sign activation networks with a novel gradient-free stochastic coordinate descent algorithm and propose an ensemble of such networks as a defense model. We evaluate the robustness of our model (a hard problem in itself) on image, text, and medical ECG data and find it to be more robust than ensembles of binary, full precision, and convolutional neural networks, and than random forests while attaining comparable clean test accuracy. In order to explain our model's robustness we show that an adversary targeting a single network in our ensemble fails to attack (and thus non-transferable to) other networks in the ensemble. Thus a datapoint requires a large distortion to fool the majority of networks in our ensemble and is likely to be detected in advance. This property of non-transferability arises naturally from the non-convexity of sign activation networks and randomization in our gradient-free training algorithm without any adversarial defense effort.
****
Tradeoffs in Data Augmentation: An Empirical Study

Raphael Gontijo-Lopes,Sylvia Smullin,Ekin Dogus Cubuk,Ethan Dyer
Though data augmentation has become a standard component of deep neural network training, the underlying mechanism behind the effectiveness of these techniques remains poorly understood. In practice, augmentation policies are often chosen using heuristics of distribution shift or augmentation diversity. Inspired by these, we conduct an empirical study to quantify how data augmentation improves model generalization. We introduce two interpretable and easy-to-compute measures: Affinity and Diversity. We find that augmentation performance is predicted not by either of these alone but by jointly optimizing the two.
****

Concept Learners for Few-Shot Learning
Kaidi Cao,Maria Brbic,Jure Leskovec
Developing algorithms that are able to generalize to a novel task given only a few labeled examples represents a fundamental challenge in closing the gap between machine- and human-level performance. The core of human cognition lies in the structured, reusable concepts that help us to rapidly adapt to new tasks and provide reasoning behind our decisions. However, existing meta-learning methods learn complex representations across prior labeled tasks without imposing any structure on the learned representations. Here we propose COMET, a meta-learning method that improves generalization ability by learning to learn along human-interpretable concept dimensions. Instead of learning a joint unstructured metric space, COMET learns mappings of high-level concepts into semi-structured metric spaces, and effectively combines the outputs of independent concept learners. We evaluate our model on few-shot tasks from diverse domains, including fine-grained image classification, document categorization and cell type annotation on a novel dataset from a biological domain developed in our work. COMET significantly outperforms strong meta-learning baselines, achieving 6-15% relative improvement on the most challenging 1-shot learning tasks, while unlike existing methods providing interpretations behind the model's predictions.
****

Analogical Reasoning for Visually Grounded Compositional Generalization
Bo Wu,Haoyu Qin,Alireza Zareian,Carl Vondrick,Shih-Fu Chang
Children acquire language subconsciously by observing the surrounding world and listening to descriptions. They can discover the meaning of words even without explicit language knowledge, and generalize to novel compositions effortlessly. In this paper, we bring this ability to AI, by studying the task of multimodal compositional generalization within the context of visually grounded language acquisition. We propose a multimodal transformer model augmented with a novel mechanism for analogical reasoning, which approximates novel compositions by learning semantic mapping and reasoning operations from previously seen compositions. Our proposed method, Analogical Reasoning Transformer Networks (ARTNet), is trained on raw multimedia data (video frames and transcripts), and after observing a set of compositions such as "washing apple" or "cutting carrot", it can generalize and recognize new compositions in new video frames, such as "washing carrot" or "cutting apple". To this end, ARTNet refers to relevant instances in the training data and uses their visual features and captions to establish analogies with the query image. Then it chooses a suitable verb and noun to create a new composition that describes the new image best. Extensive experiments on an instructional video dataset demonstrate that the proposed method achieves significantly better generalization capability and recognition accuracy compared to state-of-the-art transformer models.
****

SALR: Sharpness-aware Learning Rates for Improved Generalization
Xubo Yue,Maher Nouiehed,Raed Al Kontar
In an effort to improve generalization in deep learning, we propose SALR: a sharpness-aware learning rate update technique designed to recover flat minimizers. Our method dynamically updates the learning rate of gradient-based optimizers based on the local sharpness of the loss function. This allows optimizers to automatically increase learning rates at sharp valleys to increase the chance of escaping them. We demonstrate the effectiveness of SALR when adopted by various algo

rithms over a broad range of networks. Our experiments indicate that SALR improv es generalization, converges faster, and drives solutions to significantly flatt er regions.
****
Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamic s
Daniel Kunin,Javier Sagastuy-Brena,Surya Ganguli,Daniel LK Yamins,Hidenori Tanak a
Understanding the dynamics of neural network parameters during training is one o f the key challenges in building a theoretical foundation for deep learning. A c entral obstacle is that the motion of a network in high-dimensional parameter sp ace undergoes discrete finite steps along complex stochastic gradients derived f rom real-world datasets. We circumvent this obstacle through a unifying theoreti cal framework based on intrinsic symmetries embedded in a network's architecture that are present for any dataset. We show that any such symmetry imposes string ent geometric constraints on gradients and Hessians, leading to an associated co nservation law in the continuous-time limit of stochastic gradient descent (SGD) , akin to Noether's theorem in physics. We further show that finite learning rat es used in practice can actually break these symmetry induced conservation laws. We apply tools from finite difference methods to derive modified gradient flow, a differential equation that better approximates the numerical trajectory taken by SGD at finite learning rates. We combine modified gradient flow with our fra mework of symmetries to derive exact integral expressions for the dynamics of ce rtain parameter combinations. We empirically validate our analytic expressions f or learning dynamics on VGG-16 trained on Tiny ImageNet. Overall, by exploiting symmetry, our work demonstrates that we can analytically describe the learning d ynamics of various parameter combinations at finite learning rates and batch siz es for state of the art architectures trained on any dataset.
****
Federated Learning Based on Dynamic Regularization
Durmus Alp Emre Acar,Yue Zhao,Ramon Matas,Matthew Mattina,Paul Whatmough,Venkate sh Saligrama
We propose a novel federated learning method for distributively training neural network models, where the server orchestrates cooperation between a subset of ra ndomly chosen devices in each round. We view Federated Learning problem primaril y from a communication perspective and allow more device level computations to s ave transmission costs. We point out a fundamental dilemma, in that the minima o f the local-device level empirical loss are inconsistent with those of the globa l empirical loss. Different from recent prior works, that either attempt inexact minimization or utilize devices for parallelizing gradient computation, we prop ose a dynamic regularizer for each device at each round, so that in the limit th e global and device solutions are aligned. We demonstrate both through empirical results on real and synthetic data as well as analytical results that our schem e leads to efficient training, in both convex and non-convex settings, while bei ng fully agnostic to device heterogeneity and robust to large number of devices, partial participation and unbalanced data.
****
Fair Mixup: Fairness via Interpolation
Ching-Yao Chuang,Youssef Mroueh
Training classifiers under fairness constraints such as group fairness, regulari zes the disparities of predictions between the groups. Nevertheless, even though the constraints are satisfied during training, they might not generalize at eva luation time. To improve the generalizability of fair classifiers, we propose fa ir mixup, a new data augmentation strategy for imposing the fairness constraint. In particular, we show that fairness can be achieved by regularizing the models on paths of interpolated samples between the groups. We use mixup, a powerful data augmentation strategy to generate these interpolates. We analyze fair mixu p and empirically show that it ensures a better generalization for both accuracy and fairness measurement in tabular, vision, and language benchmarks.
****

Fidelity-based Deep Adiabatic Scheduling
Eli Ovits,Lior Wolf
Adiabatic quantum computation is a form of computation that acts by slowly interpolating a quantum system between an easy to prepare initial state and a final state that represents a solution to a given computational problem. The choice of the interpolation schedule is critical to the performance: if at a certain time point, the evolution is too rapid, the system has a high probability to transfer to a higher energy state, which does not represent a solution to the problem. On the other hand, an evolution that is too slow leads to a loss of computation time and increases the probability of failure due to decoherence. In this work, we train deep neural models to produce optimal schedules that are conditioned on the problem at hand. We consider two types of problem representation: the Hamiltonian form, and the Quadratic Unconstrained Binary Optimization (QUBO) form. A novel loss function that scores schedules according to their approximated success probability is introduced. We benchmark our approach on random QUBO problems, Grover search, 3-SAT, and MAX-CUT problems and show that our approach outperforms, by a sizable margin, the linear schedules as well as alternative approaches that were very recently proposed.
****
Guiding Representation Learning in Deep Generative Models with Policy Gradients
Luca Lach,Timo Korthals,Malte Schilling,Helge Ritter
Variational Auto Encoder (VAE) provide an efficient latent space representation of complex data distributions which is learned in an unsupervised fashion.
Using such a representation as input to Reinforcement Learning (RL) approaches may reduce learning time, enable domain transfer or improve interpretability of the model.
However, current state-of-the-art approaches that combine VAE with RL fail at learning good performing policies on certain RL domains.
Typically, the VAE is pre-trained in isolation and may omit the embedding of task-relevant features due to insufficiencies of its loss.
As a result, the RL approach can not successfully maximize the reward on these domains.
Therefore, this paper investigates the issues of joint training approaches and explores incorporation of policy gradients from RL into the VAE's latent space to find a task-specific latent space representation.
We show that using pre-trained representations can lead to policies being unable to learn any rewarding behaviour in these environments.
Subsequently, we introduce two types of models which overcome this deficiency by using policy gradients to learn the representation.
Thereby the models are able to embed features into its representation that are crucial for performance on the RL task but would not have been learned with previous methods.
****
Heating up decision boundaries: isocapacitory saturation, adversarial scenarios and generalization bounds
Bogdan Georgiev,Lukas Franken,Mayukh Mukherjee
In the present work we study classifiers' decision boundaries via Brownian motion processes in ambient data space and associated probabilistic techniques. Intuitively, our ideas correspond to placing a heat source at the decision boundary and observing how effectively the sample points warm up. We are largely motivated by the search for a soft measure that sheds further light on the decision boundary's geometry. En route, we bridge aspects of potential theory and geometric analysis (Maz'ya 2011, Grigor'Yan and Saloff-Coste 2002) with active fields of ML research such as adversarial examples and generalization bounds. First, we focus on the geometric behavior of decision boundaries in the light of adversarial attack/defense mechanisms. Experimentally, we observe a certain capacitory trend over different adversarial defense strategies: decision boundaries locally become flatter as measured by isoperimetric inequalities (Ford et al 2019); however, our more sensitive heat-diffusion metrics extend this analysis and further reveal that some non-trivial geometry invisible to plain distance-based methods is s

till preserved. Intuitively, we provide evidence that the decision boundaries ne vertheless retain many persistent "wiggly and fuzzy" regions on a finer scale. Second, we show how Brownian hitting probabilities translate to soft generalizat ion bounds which are in turn connected to compression and noise stability (Arora et al 2018), and these bounds are significantly stronger if the decision bounda ry has controlled geometric features.
****
Information Lattice Learning
Haizi Yu,James Evans,Lav R. Varshney
Information Lattice Learning (ILL) is a general framework to learn decomposed re presentations, called rules, of a signal such as an image or a probability distr ibution. Each rule is a coarsened signal used to gain some human-interpretable i nsight into what might govern the nature of the original signal. To summarize th e signal, we need several disentangled rules arranged in a hierarchy, formalized by a lattice structure. ILL focuses on explainability and generalizability from "small data", and aims for rules akin to those humans distill from experience (rather than a representation optimized for a specific task like classification). This paper focuses on a mathematical and algorithmic presentation of ILL, then demonstrates how ILL addresses the core question "what makes X an X" or "what ma kes X different from Y" to create effective, rule-based explanations designed to help human learners understand. The key part here is \emph{what} rather than ta sks like generating X or predicting labels X,Y. Typical applications of ILL are presented for artistic and scientific knowledge discovery. These use ILL to lear n music theory from scores and chemical laws from molecule data, revealing relat ionships between domains. We include initial benchmarks and assessments for ILL to demonstrate efficacy.
****
Watching the World Go By: Representation Learning from Unlabeled Videos
Daniel Gordon,Kiana Ehsani,Dieter Fox,Ali Farhadi
Recent unsupervised representation learning techniques show remarkable success o n many single image tasks by using instance discrimination: learning to differe ntiate between two augmented versions of the same image and a large batch of unr elated images. Prior work uses artificial data augmentation techniques such as cropping, and color jitter which can only affect the image in superficial waysan d are not aligned with how objects actually change e.g. occlusion, deformation,v iewpoint change. We argue that videos offer this natural augmentation for free. Videos can provide entirely new views of objects, show deformation, and even c onnect semantically similar but visually distinct concepts.We propose Video Nois e Contrastive Estimation, a method for using unlabeled video to learnstrong, tr ansferable, single image representations.We demonstrate improvements over recent unsupervised single image techniques,as well as over fully supervised ImageNet pretraining, across temporal and non-temporal tasks.
****
Deciphering and Optimizing Multi-Task Learning: a Random Matrix Approach
Malik Tiomoko,Hafiz Tiomoko Ali,Romain Couillet
This article provides theoretical insights into the inner workings of multi-task and transfer learning methods, by studying the tractable least-square support v ector machine multi-task learning (LS-SVM MTL) method, in the limit of large ($p $) and numerous ($n$) data. By a random matrix analysis applied to a Gaussian mi xture data model, the performance of MTL LS-SVM is shown to converge, as $n,p\to \infty$, to a deterministic limit involving simple (small-dimensional) statistic s of the data.

We prove (i) that the standard MTL LS-SVM algorithm is in general strongly biase d and may dramatically fail (to the point that individual single-task LS-SVMs ma y outperform the MTL approach, even for quite resembling tasks): our analysis pr ovides a simple method to correct these biases, and that we reveal (ii) the suff icient statistics at play in the method, which can be efficiently estimated, eve n for quite small datasets. The latter result is exploited to automatically opti mize the hyperparameters without resorting to any cross-validation procedure.

Experiments on popular datasets demonstrate that our improved MTL LS-SVM method is computationally-efficient and outperforms sometimes much more elaborate state-of-the-art multi-task and transfer learning techniques.
****
Predicting What You Already Know Helps: Provable Self-Supervised Learning
Jason D. Lee,Qi Lei,Nikunj Saunshi,Jiacheng Zhuo
Self-supervised representation learning solves auxiliary prediction tasks (known as pretext tasks), that do not require labeled data, to learn semantic representations. These pretext tasks are created solely using the input features, such as predicting a missing image patch, recovering the color channels of an image from context, or predicting missing words, yet predicting this \textit{known} information helps in learning representations effective for downstream prediction tasks. This paper posits a mechanism based on approximate conditional independence to formalize how solving certain pretext tasks can learn representations that provably decrease the sample complexity of downstream supervised tasks. Formally, we quantify how the approximate independence between the components of the pretext task (conditional on the label and latent variables) allows us to learn representations that can solve the downstream task with drastically reduced sample complexity by just training a linear layer on top of the learned representation.
****
On Flat Minima, Large Margins and Generalizability
Daniel Lengyel,Nicholas Jennings,Panos Parpas,Nicholas Kantas
The intuitive connection to robustness and convincing empirical evidence have made the flatness of the loss surface an attractive measure of generalizability for neural networks.

Yet it suffers from various problems such as computational difficulties, reparametrization issues, and a growing concern that it may only be an epiphenomenon of optimization methods.

We provide empirical evidence that under the cross-entropy loss once a neural network reaches a non-trivial training error, the flatness correlates (via Pearson Correlation Coefficient) well to the classification margins, which allows us to better reason about the concerns surrounding flatness.

Our results lead to the practical recommendation that when assessing generalizability one should consider a margin-based measure instead, as it is computationally more efficient, provides further insight, and is highly correlated to flatness.

We also use our insight to replace the misleading folklore that small-batch methods generalize better because they are able to escape sharp minima. Instead, we argue that large-batch methods did not have enough time to maximize margins and hence generalize worse.
****
Overinterpretation reveals image classification model pathologies
Brandon Carter,Siddhartha Jain,Jonas Mueller,David Gifford
Image classifiers are typically scored on their test set accuracy, but high accuracy can mask a subtle type of model failure. We find that high scoring convolutional neural networks (CNNs) on popular benchmarks exhibit troubling pathologies that allow them to display high accuracy even in the absence of semantically salient features. When a model provides a high-confidence decision without salient supporting input features, we say the classifier has overinterpreted its input, finding too much class-evidence in patterns that appear nonsensical to humans. Here, we demonstrate that neural networks trained on CIFAR-10 and ImageNet suffer from overinterpretation, and we find models on CIFAR-10 make confident predictions even when 95% of input images are masked and humans cannot discern salient features in the remaining pixel-subsets. Although these patterns portend potential model fragility in real-world deployment, they are in fact valid statistical patterns of the benchmark that alone suffice to attain high test accuracy. Unlike adversarial examples, overinterpretation relies upon unmodified image pixels. We find ensembling and input dropout can each help mitigate overinterpretation.
****

Improving the Unsupervised Disentangled Representation Learning with VAE Ensembl
e
Nanxiang Li,Shabnam Ghaffarzadegan,Liu Ren
Variational Autoencoder (VAE) based frameworks have achieved the state-of-the-ar
t performance on the unsupervised disentangled representation learning. A recent
 theoretical analysis shows that such success is mainly due to the VAE implement
ation choices that encourage a PCA-like behavior locally on data samples. Despit
e this implied model identifiability, the VAE based disentanglement frameworks s
till face the trade-off between the local orthogonality and data reconstruction.
 As a result, models with the same architecture and hyperparameter setting can s
ometime learn entangled representations. To address this challenge, we propose a
 simple yet effective VAE ensemble framework consisting of multiple VAEs. It is
based on the assumption that entangled representations are unique in their own w
ays, and the disentangled representations are "alike" (similar up to a signed pe
rmutation transformation). In the proposed VAE ensemble, each model not only mai
ntains its original objective, but also encodes to and decodes from other models
 through pair-wise linear transformations between the latent representations. We
 show both theoretically and experimentally, the VAE ensemble objective encourag
es the linear transformations connecting the VAEs to be trivial transformations,
 aligning the latent representations of different models to be "alike". We compa
re our approach with the state-of-the-art unsupervised disentangled representati
on learning approaches and show the improved performance.
****
Influence Functions in Deep Learning Are Fragile
Samyadeep Basu,Phil Pope,Soheil Feizi
Influence functions approximate the effect of training samples in test-time pred
ictions and have a wide variety of applications in machine learning interpretabi
lity and uncertainty estimation. A commonly-used (first-order) influence functio
n can be implemented efficiently as a post-hoc method requiring access only to t
he gradients and Hessian of the model. For linear models, influence functions ar
e well-defined due to the convexity of the underlying loss function and are gene
rally accurate even across difficult settings where model changes are fairly lar
ge such as estimating group influences. Influence functions, however, are not we
ll-understood in the context of deep learning with non-convex loss functions. I
n this paper, we provide a comprehensive and large-scale empirical study of succ
esses and failures of influence functions in neural network models trained on da
tasets such as Iris, MNIST, CIFAR-10 and ImageNet. Through our extensive experim
ents, we show that the network architecture, its depth and width, as well as the
 extent of model parameterization and regularization techniques have strong effe
cts in the accuracy of influence functions. In particular, we find that (i) infl
uence estimates are fairly accurate for shallow networks, while for deeper netwo
rks the estimates are often erroneous; (ii) for certain network architectures an
d datasets, training with weight-decay regularization is important to get high-q
uality influence estimates; and (iii) the accuracy of influence estimates can va
ry significantly depending on the examined test points. These results suggest th
at in general influence functions in deep learning are fragile and call for deve
loping improved influence estimation methods to mitigate these issues in non-con
vex setups.
****
Few-Shot Learning via Learning the Representation, Provably
Simon Shaolei Du,Wei Hu,Sham M. Kakade,Jason D. Lee,Qi Lei
This paper studies few-shot learning via representation learning, where one uses
 T source tasks with n_1 data per task to learn a representation in order to
 reduce the sample complexity of a target task for which there is only $n_2 (\ll
 n_1)$ data. Specifically, we focus on the setting where there exists a good com
mon representation between source and target, and our goal is to understand how
much a sample size reduction is possible. First, we study the setting where this
 common representation is low-dimensional and provide a risk bound of $\tilde{O}
(\frac{dk}{n_1T} + \frac{k}{n_2})$ on the target task for the linear representat
ion class; here d is the ambient input dimension and $k (\ll d)$ is the dimens

ion of the representation. This result bypasses the $\Omega(\frac{1}{T})$ barrie r under the i.i.d. task assumption, and can capture the desired property that al l n_1T samples from source tasks can be \emph{pooled} together for representat ion learning. We further extend this result to handle a general representation f unction class and obtain a similar result. Next, we consider the setting where t he common representation may be high-dimensional but is capacity-constrained (sa y in norm); here, we again demonstrate the advantage of representation learning in both high-dimensional linear regression and neural networks, and show that re presentation learning can fully utilize all n_1T samples from source tasks.
****
Self-Supervised Learning of Compressed Video Representations
Youngjae Yu,Sangho Lee,Gunhee Kim,Yale Song
Self-supervised learning of video representations has received great attention. Existing methods typically require frames to be decoded before being processed, which increases compute and storage requirements and ultimately hinders large-sc ale training. In this work, we propose an efficient self-supervised approach to learn video representations by eliminating the expensive decoding step. We use a three-stream video architecture that encodes I-frames and P-frames of a compres sed video. Unlike existing approaches that encode I-frames and P-frames individu ally, we propose to jointly encode them by establishing bidirectional dynamic co nnections across streams. To enable self-supervised learning, we propose two pre text tasks that leverage the multimodal nature (RGB, motion vector, residuals) a nd the internal GOP structure of compressed videos. The first task asks our netw ork to predict zeroth-order motion statistics in a spatio-temporal pyramid; the second task asks correspondence types between I-frames and P-frames after applyi ng temporal transformations. We show that our approach achieves competitive perf ormance on compressed video recognition both in supervised and self-supervised r egimes.

****
Uncertainty Prediction for Deep Sequential Regression Using Meta Models
Jiri Navratil,Matthew Arnold,Benjamin Elder
Generating high quality uncertainty estimates for sequential regression, particu larly deep recurrent networks, remains a challenging and open problem.
Existing approaches often make restrictive assumptions (such as stationarity) ye t still perform poorly in practice, particularly in presence of real world non-s tationary signals and drift.
This paper describes a flexible method that can generate symmetric and asymmetri c uncertainty estimates, makes no assumptions about stationarity, and outperfor ms competitive baselines on both drift and non drift scenarios.
This work helps make sequential regression more effective and practical for use in real-world applications, and is a powerful new addition to the modeling toolb ox for sequential uncertainty quantification in general.
****
FLAGNet : Feature Label based Automatic Generation Network for symbolic music
SeongHyeon Go
The technology for automatic music generation has been very actively studied in recent years. However, almost in these studies, handling domain knowledge of mus ic was omitted or considered a difficult task. In particular, research that anal yzes and utilizes the characteristics of each bar of music is very rare, even th ough it is essential in the human composition. We propose a model that generate music with musical characteristics of bars by conditional generative adversarial network, and analyze the good combination of the sequence of which characterize d bars for symbolic-domain music generation by Recurrent Neural Network with Lo ng short term memory layer. Also, by analyzing symbolic music data as image-like based on relational pitch approach, it increases the utilization of the data se t with arbitrary chord scales and enables the use of generational results extens ively. The resulting model FLAGNet generates music with the understanding of mus ical domain knowledge while handling inputs like minimum unit of note, length of music, chart scales, and chord condition.

****
Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis
Rafael Valle,Kevin J. Shih,Ryan Prenger,Bryan Catanzaro
In this paper we propose Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis with style transfer and speech variation. Flowtron borrows insights from Autoregressive Flows and revamps Tacotron 2 in order to provide high-quality and expressive mel-spectrogram synthesis. Flowtron is optimized by maximizing the likelihood of the training data, which makes training simple and stable. Flowtron learns an invertible mapping of data to a latent space that can be used to modulate many aspects of speech synthesis (timbre, expressivity, accent). Our mean opinion scores (MOS) show that Flowtron matches state-of-the-art TTS models in terms of speech quality. We provide results on speech variation, interpolation over time between samples and style transfer between seen and unseen speakers. Code and pre-trained models are publicly available at \href{https://github.com/NVIDIA/flowtron}{https://github.com/NVIDIA/flowtron}.
****
Adaptive Personalized Federated Learning
Yuyang Deng,Mohammad Mahdi Kamani,Mehrdad Mahdavi
Investigation of the degree of personalization in federated learning algorithms has shown that only maximizing the performance of the global model will confine the capacity of the local models to personalize. In this paper, we advocate an adaptive personalized federated learning (APFL) algorithm, where each client will train their local models while contributing to the global model. We derive the generalization bound of mixture of local and global models, and find the optimal mixing parameter. We also propose a communication-efficient optimization method to collaboratively learn the personalized models and analyze its convergence in both smooth strongly convex and nonconvex settings. The extensive experiments demonstrate the effectiveness of our personalization schema, as well as the correctness of established generalization theories.

****
Novel Policy Seeking with Constrained Optimization
Hao Sun,Zhenghao Peng,Bo Dai,Jian Guo,Dahua Lin,Bolei Zhou
We address the problem of seeking novel policies in reinforcement learning tasks. Instead of following the multi-objective framework commonly used in existing methods, we propose to rethink the problem under a novel perspective of constrained optimization. We at first introduce a new metric to evaluate the difference between policies, and then design two practical novel policy seeking methods following the new perspective, namely the Constrained Task Novel Bisector (CTNB), and the Interior Policy Differentiation (IPD), corresponding to the feasible direction method and the interior point method commonly known in the constrained optimization literature. Experimental comparisons on the MuJuCo control suite show our methods can achieve substantial improvements over previous novelty-seeking methods in terms of both the novelty of policies and their performances in the primal task.
****
Discrete Predictive Representation for Long-horizon Planning
Thanard Kurutach,Julia Peng,Yang Gao,Stuart Russell,Pieter Abbeel
Discrete representations have been key in enabling robots to plan at more abstract levels and solve temporally-extended tasks more efficiently for decades. However, they typically require expert specifications. On the other hand, deep reinforcement learning aims to learn to solve tasks end-to-end, but struggles with long-horizon tasks. In this work, we propose Discrete Object-factorized Representation Planning (DORP), which learns temporally-abstracted discrete representations from exploratory video data in an unsupervised fashion via a mutual information maximization objective. DORP plans a sequence of abstract states for a low-level model-predictive controller to follow. In our experiments, we show that DORP robustly solves unseen long-horizon tasks. Interestingly, it discovers independent representations per object and binary properties such as a key-and-door.

****
R-GAP: Recursive Gradient Attack on Privacy
Junyi Zhu,Matthew B. Blaschko
Federated learning frameworks have been regarded as a promising approach to break the dilemma between demands on privacy and the promise of learning from large collections of distributed data. Many such frameworks only ask collaborators to share their local update of a common model, i.e. gradients with respect to locally stored data, instead of exposing their raw data to other collaborators. However, recent optimization-based gradient attacks show that raw data can often be accurately recovered from gradients. It has been shown that minimizing the Euclidean distance between true gradients and those calculated from estimated data is often effective in fully recovering private data. However, there is a fundamental lack of theoretical understanding of how and when gradients can lead to unique recovery of original data. Our research fills this gap by providing a closed-form recursive procedure to recover data from gradients in deep neural networks. We name it Recursive Gradient Attack on Privacy (R-GAP). Experimental results demonstrate that R-GAP works as well as or even better than optimization-based approaches at a fraction of the computation under certain conditions. Additionally, we propose a Rank Analysis method, which can be used to estimate the risk of gradient attacks inherent in certain network architectures, regardless of whether an optimization-based or closed-form-recursive attack is used. Experimental results demonstrate the utility of the rank analysis towards improving the network's security. Source code is available for download from https://github.com/JunyiZhu-AI/R-GAP.
****
Towards Understanding Linear Value Decomposition in Cooperative Multi-Agent Q-Learning
Jianhao Wang,Zhizhou Ren,Beining Han,Jianing Ye,Chongjie Zhang
Value decomposition is a popular and promising approach to scaling up multi-agent reinforcement learning in cooperative settings. However, the theoretical understanding of such methods is limited. In this paper, we introduce a variant of the fitted Q-iteration framework for analyzing multi-agent Q-learning with value decomposition. Based on this framework, we derive a closed-form solution to the empirical Bellman error minimization with linear value decomposition. With this novel solution, we further reveal two interesting insights: 1) linear value decomposition implicitly implements a classical multi-agent credit assignment called counterfactual difference rewards; and 2) On-policy data distribution or richer Q function classes can improve the training stability of multi-agent Q-learning. In the empirical study, our experiments demonstrate the realizability of our theoretical closed-form formulation and implications in the didactic examples and a broad set of StarCraft II unit micromanagement tasks, respectively.
****
Quickly Finding a Benign Region via Heavy Ball Momentum in Non-Convex Optimization
Jun-Kun Wang,Jacob Abernethy
The Heavy Ball Method, proposed by Polyak over five decades ago, is a first-order method for optimizing continuous functions. While its stochastic counterpart has proven extremely popular in training deep networks, there are almost no known functions where deterministic Heavy Ball is provably faster than the simple and classical gradient descent algorithm in non-convex optimization. The success of Heavy Ball has thus far eluded theoretical understanding. Our goal is to address this gap, and in the present work we identify two non-convex problems where we provably show that the Heavy Ball momentum helps the iterate to enter a benign region that contains a global optimal point faster. We show that Heavy Ball exhibits simple dynamics that clearly reveal the benefit of using a larger value of momentum parameter for the problems. The first of these optimization problems is the phase retrieval problem, which has useful applications in physical science. The second of these optimization problems is the cubic-regularized minimization, a critical subroutine required by Nesterov-Polyak cubic-regularized method to

find second-order stationary points in general smooth non-convex problems.
****
ME-MOMENTUM: EXTRACTING HARD CONFIDENT EXAMPLES FROM NOISILY LABELED DATA
Yingbin Bai,Tongliang Liu
Examples that are close to the decision boundary—that we term hard examples, are essential to shaping accurate classifiers. Extracting confident examples has been widely studied in the community of learning with noisy labels. However, it remains elusive how to extract hard confident examples from the noisy training data. In this paper, we propose a deep learning paradigm to solve this problem, which is built on the memorization effect of deep neural networks that they would first learn simple patterns, i.e., which are defined by these shared by multiple training examples. To extract hard confident examples that contain non-simple patterns and are entangled with the inaccurately labeled examples, we borrow the idea of momentum from physics. Specifically, we alternately update the confident examples and refine the classifier. Note that the extracted confident examples in the previous round can be exploited to learn a better classifier and that the better classifier will help identify better (and hard) confident examples. We call the approach the “Momentum of Memorization” (Me-Momentum). Empirical results on benchmark-simulated and real-world label-noise data illustrate the effectiveness of Me-Momentum for extracting hard confident examples, leading to better classification performance.
****
Robust Temporal Ensembling
Abel Brown,Benedikt Schifferer,Robert DiPietro
Successful training of deep neural networks with noisy labels is an essential capability as most real-world datasets contain some amount of mislabeled data. Left unmitigated, label noise can sharply degrade typical supervised learning approaches. In this paper, we present robust temporal ensembling (RTE), a simple supervised learning approach which combines robust task loss, temporal pseudo-labeling, and a new ensemble consistency regularization term to achieve noise-robust learning. We demonstrate that RTE achieves state-of-the-art performance across the CIFAR-10, CIFAR-100, and ImageNet datasets, while forgoing the recent trend of label filtering/fixing. In particular, RTE achieves 93.64% accuracy on CIFAR-10 and 66.43% accuracy on CIFAR-100 under 80% label corruption, and achieves 74.79% accuracy on ImageNet under 40% corruption. These are substantial gains over previous state-of-the-art accuracies of 86.6%, 60.2%, and 71.31%, respectively, achieved using three distinct methods. Finally, we show that RTE retains competitive corruption robustness to unforeseen input noise using CIFAR-10-C, obtaining a mean corruption error (mCE) of 13.50% even in the presence of an 80% noise ratio, versus 26.9% mCE with standard methods on clean data.
****
Non-Local Graph Neural Networks
Meng Liu,Zhengyang Wang,Shuiwang Ji
Modern graph neural networks (GNNs) learn node embeddings through multilayer local aggregation and achieve great success in applications on assortative graphs. However, tasks on disassortative graphs usually require non-local aggregation. In addition, we find that local aggregation is even harmful for some disassortative graphs. In this work, we propose a simple yet effective non-local aggregation framework with an efficient attention-guided sorting for GNNs. Based on it, we develop various non-local GNNs. We perform thorough experiments to analyze disassortative graph datasets and evaluate our non-local GNNs. Experimental results demonstrate that our non-local GNNs significantly outperform previous state-of-the-art methods on six benchmark datasets of disassortative graphs, in terms of both model performance and efficiency.
****
Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients
Jing An,Lexing Ying,Yuhua Zhu
A data set sampled from a certain population is biased if the subgroups of the population are sampled at proportions that are significantly different from their

underlying proportions. Training machine learning models on biased data sets re
quires correction techniques to compensate for the bias. We consider two commonl
y-used techniques, resampling and reweighting, that rebalance the proportions of
 the subgroups to maintain the desired objective function. Though statistically
equivalent, it has been observed that resampling outperforms reweighting when co
mbined with stochastic gradient algorithms. By analyzing illustrative examples,
we explain the reason behind this phenomenon using tools from dynamical stabilit
y and stochastic asymptotics. We also present experiments from regression, class
ification, and off-policy prediction to demonstrate that this is a general pheno
menon. We argue that it is imperative to consider the objective function design
and the optimization algorithm together while addressing the sampling bias.

****
Unsupervised Video Decomposition using Spatio-temporal Iterative Inference
Polina Zablotskaia,Edoardo Alberto Dominici,Leonid Sigal,Andreas Lehrmann
Unsupervised multi-object scene decomposition is a fast-emerging problem in repr
esentation learning. Despite significant progress in static scenes, such models
are unable to leverage important dynamic cues present in video. We propose a nov
el spatio-temporal iterative inference framework that is powerful enough to join
tly model complex multi-object representations and explicit temporal dependencie
s between latent variables across frames. This is achieved by leveraging 2D-LSTM
, temporally conditioned inference and generation within the iterative amortized
 inference for posterior refinement. Our method improves the overall quality of
decompositions, encodes information about the objects' dynamics, and can be used
 to predict trajectories of each object separately. Additionally, we show that o
ur model has a high accuracy even without color information. We demonstrate the
decomposition, segmentation, and prediction capabilities of our model and show t
hat it outperforms the state-of-the-art on several benchmark datasets, one of wh
ich was curated for this work and will be made publicly available.
****
Whitening for Self-Supervised Representation Learning
Aleksandr Ermolov,Aliaksandr Siarohin,Enver Sangineto,Nicu Sebe
Most of the self-supervised representation learning methods are based on the con
trastive loss and the instance-discrimination task, where augmented versions of
the same image instance ("positives") are contrasted with instances extracted f
rom other images ("negatives"). For the learning to be effective, a lot of negat
ives should be compared with a positive pair, which is computationally demanding
. In this paper, we propose a different direction and a new loss function for se
lf-supervised representation learning which is based on the whitening of the lat
ent-space features. The whitening operation has a "scattering" effect on the bat
ch samples, which compensates the use of negatives, avoiding degenerate solution
s where all the sample representations collapse to a single point. Our Whitening
 MSE (W-MSE) loss does not require special heuristics (e.g. additional networks)
 and it is conceptually simple. Since negatives are not needed, we can extract m
ultiple positive pairs from the same image instance. We empirically show that W
-MSE is competitive with respect to popular, more complex self-supervised method
s. The source code of the method and all the experiments is included in the Supp
lementary Material.
****
Adaptive Gradient Methods Converge Faster with Over-Parameterization (and you ca
n do a line-search)
Sharan Vaswani,Issam H. Laradji,Frederik Kunstner,Si Yi Meng,Mark Schmidt,Simon
Lacoste-Julien
Adaptive gradient methods are typically used for training over-parameterized mod
els capable of exactly fitting the data; we thus study their convergence in this
 interpolation setting. Under an interpolation assumption, we prove that AMSGrad
 with a constant step-size and momentum can converge to the minimizer at the fas
ter $O(1/T)$ rate for smooth, convex functions. Furthermore, in this setting, we
 show that AdaGrad can achieve an $O(1)$ regret in the online convex optimizatio
n framework. When interpolation is only approximately satisfied, we show that co

nstant step-size AMSGrad converges to a neighbourhood of the solution. On the ot
her hand, we prove that AdaGrad is robust to the violation of interpolation and
converges to the minimizer at the optimal rate. However, we demonstrate that eve
n for simple, convex problems satisfying interpolation, the empirical performanc
e of these methods heavily depends on the step-size and requires tuning. We alle
viate this problem by using stochastic line-search (SLS) and Polyak's step-sizes
 (SPS) to help these methods adapt to the function's local smoothness. By using
these techniques, we prove that AdaGrad and AMSGrad do not require knowledge of
problem-dependent constants and retain the convergence guarantees of their const
ant step-size counterparts. Experimentally, we show that these techniques help i
mprove the convergence and generalization performance across tasks, from binary
classification with kernel mappings to classification with deep neural networks.
****

Learning with Instance-Dependent Label Noise: A Sample Sieve Approach
Hao Cheng,Zhaowei Zhu,Xingyu Li,Yifei Gong,Xing Sun,Yang Liu
Human-annotated labels are often prone to noise, and the presence of such noise
will degrade the performance of the resulting deep neural network (DNN) models.
Much of the literature (with several recent exceptions) of learning with noisy l
abels focuses on the case when the label noise is independent of features. Pract
ically, annotations errors tend to be instance-dependent and often depend on the
 difficulty levels of recognizing a certain task. Applying existing results from
 instance-independent settings would require a significant amount of estimation
of noise rates. Therefore, providing theoretically rigorous solutions for learni
ng with instance-dependent label noise remains a challenge. In this paper, we pr
opose CORES^{2} (COnfidence REgularized Sample Sieve), which progressively sie
ves out corrupted examples. The implementation of CORES^{2} does not require
specifying noise rates and yet we are able to provide theoretical guarantees of
CORES^{2} in filtering out the corrupted examples. This high-quality sample si
eve allows us to treat clean examples and the corrupted ones separately in train
ing a DNN solution, and such a separation is shown to be advantageous in the ins
tance-dependent noise setting. We demonstrate the performance of CORES^{2} on
CIFAR10 and CIFAR100 datasets with synthetic instance-dependent label noise and
Clothing1M with real-world human noise. As of independent interests, our sample
sieve provides a generic machinery for anatomizing noisy datasets and provides a
 flexible interface for various robust training techniques to further improve th
e performance. Code is available at https://github.com/UCSC-REAL/cores.
****

EMTL: A Generative Domain Adaptation Approach
Jianfeng Zhang,Illyyne Saffar,Aladin Virmaux,Balázs Kégl
We propose an unsupervised domain adaptation approach based on generative models
. We show that when the source probability density function can be learned, one-
step Expectation–Maximization iteration plus an additional marginal density func
tion constraint will produce a proper mediator probability density function to b
ridge the gap between the source and target domains. The breakthrough is based o
n modern generative models (autoregressive mixture density nets) that are compet
itive to discriminative models on moderate-dimensional classification problems.
By decoupling the source density estimation from the adaption steps, we can desi
gn a domain adaptation approach where the source data is locked away after being
 processed only once, opening the door to transfer when data security or privacy
 concerns impede the use of traditional domain adaptation. We demonstrate that o
ur approach can achieve state-of-the-art performance on synthetic and real data
sets, without accessing the source data at the adaptation phase.
****

Sim2SG: Sim-to-Real Scene Graph Generation for Transfer Learning
Aayush Prakash,Shoubhik Debnath,Jean Francois Lafleche,Eric Cameracci,Gavriel St
ate,Marc T Law
Scene graph (SG) generation has been gaining a lot of traction recently. Current
 SG generation techniques, however, rely on the availability of expensive and li
mited number of labeled datasets. Synthetic data offers a viable alternative as
labels are essentially free. However, neural network models trained on synthetic

data, do not perform well on real data because of the domain gap.
To overcome this challenge, we propose Sim2SG, a scalable technique for sim-to-real transfer for scene graph generation. Sim2SG addresses the domain gap by decomposing it into appearance, label and prediction discrepancies between the two domains. We handle these discrepancies by introducing pseudo statistic based self-learning and adversarial techniques. Sim2SG does not require costly supervision from the real-world dataset.
Our experiments demonstrate significant improvements over baselines in reducing the domain gap both qualitatively and quantitatively.
We validate our approach on toy simulators, as well as realistic simulators evaluated on real-world data.
****
DynamicVAE: Decoupling Reconstruction Error and Disentangled Representation Learning
Huajie Shao,Haohong Lin,Qinmin Yang,Shuochao Yao,Han Zhao,Tarek Abdelzaher
This paper challenges the common assumption that the weight β, in β-VAE, should be larger than 1 in order to effectively disentangle latent factors. We demonstrate that β-VAE, with $\beta < 1$, can not only attain good disentanglement but also significantly improve reconstruction accuracy via dynamic control. The paper \textit{removes the inherent trade-off} between reconstruction accuracy and disentanglement for β-VAE. Existing methods, such as β-VAE and FactorVAE, assign a large weight to the KL-divergence term in the objective function, leading to high reconstruction errors for the sake of better disentanglement. To mitigate this problem, a ControlVAE has recently been developed that dynamically tunes the KL-divergence weight in an attempt to \textit{control the trade-off} to more a favorable point. However, ControlVAE fails to eliminate the conflict between the need for a large β (for disentanglement) and the need for a small β (for smaller reconstruction error). Instead, we propose DynamicVAE that maintains a different β at different stages of training, thereby \textit{decoupling disentanglement and reconstruction accuracy}. In order to evolve the weight, β, along a trajectory that enables such decoupling, DynamicVAE leverages a modified incremental PI (proportional-integral) controller, a variant of proportional-integral-derivative controller (PID) algorithm, and employs a moving average as well as a hybrid annealing method to evolve the value of KL-divergence smoothly in a tightly controlled fashion. We theoretically prove the stability of the proposed approach. Evaluation results on three benchmark datasets demonstrate that DynamicVAE significantly improves the reconstruction accuracy while achieving disentanglement comparable to the best of existing methods. The results verify that our method can separate disentangled representation learning and reconstruction, removing the inherent tension between the two.
****
Unsupervised Audiovisual Synthesis via Exemplar Autoencoders
Kangle Deng,Aayush Bansal,Deva Ramanan
We present an unsupervised approach that converts the input speech of any individual into audiovisual streams of potentially-infinitely many output speakers. Our approach builds on simple autoencoders that project out-of-sample data onto the distribution of the training set. We use exemplar autoencoders to learn the voice, stylistic prosody, and visual appearance of a specific target exemplar speech. In contrast to existing methods, the proposed approach can be easily extended to an arbitrarily large number of speakers and styles using only 3 minutes of target audio-video data, without requiring any training data for the input speaker. To do so, we learn audiovisual bottleneck representations that capture the structured linguistic content of speech. We outperform prior approaches on both audio and video synthesis.

****
Quantifying Exposure Bias for Open-ended Language Generation
Tianxing He,Jingzhao Zhang,Zhiming Zhou,James R. Glass
The exposure bias problem refers to the incrementally distorted generation induc

ed by the training-generation discrepancy, in teacher-forcing training for auto-regressive neural network language models (LM). It has been regarded as a central problem for LMs trained for open-ended language generation. Although a lot of algorithms have been proposed to avoid teacher forcing and therefore alleviate exposure bias, there is little work showing how serious the exposure bias problem actually is. In this work, we propose novel metrics to quantify the impact of exposure bias in the generation of MLE-trained LMs. Our key intuition is that if we feed ground-truth data prefixes (instead of prefixes generated by the model itself) into the model and ask it to continue the generation, the performance should become much better because the training-generation discrepancy in the prefix is removed. We conduct both automatic and human evaluation in our experiments, and our observations are two-fold: (1) We confirm that the prefix discrepancy indeed induces some level of performance loss. (2) However, the induced distortion seems to be limited, and is not incremental during the generation, which contradicts the claim of exposure bias.

****

Single-Timescale Actor-Critic Provably Finds Globally Optimal Policy

Zuyue Fu,Zhuoran Yang,Zhaoran Wang

We study the global convergence and global optimality of actor-critic, one of the most popular families of reinforcement learning algorithms. While most existing works on actor-critic employ bi-level or two-timescale updates, we focus on the more practical single-timescale setting, where the actor and critic are updated simultaneously. Specifically, in each iteration, the critic update is obtained by applying the Bellman evaluation operator only once while the actor is updated in the policy gradient direction computed using the critic. Moreover, we consider two function approximation settings where both the actor and critic are represented by linear or deep neural networks. For both cases, we prove that the actor sequence converges to a globally optimal policy at a sublinear $O(K^{-1/2})$ rate, where K is the number of iterations. To the best of our knowledge, we establish the rate of convergence and global optimality of single-timescale actor-critic with linear function approximation for the first time. Moreover, under the broader scope of policy optimization with nonlinear function approximation, we prove that actor-critic with deep neural network finds the globally optimal policy at a sublinear rate for the first time.

****

MULTI-SPAN QUESTION ANSWERING USING SPAN-IMAGE NETWORK

Tarik Arici,Hayreddin Ceker,Ismail Baha Tutar

Question-answering (QA) models aim to find an answer given a question and con- text. Language models like BERT are used to associate question and context to find an answer span. Prior art on QA focuses on finding the best answer. There is a need for multi-span QA models to output the top-K likely answers to questions such as "Which companies Elon Musk started?" or "What factors cause global warming?" In this work, we introduce Span-Image architecture that can learn to identify multiple answers in a context for a given question. This architecture can incorporate prior information about the span length distribution or valid span patterns (e.g., end index has to be larger than start index), thus eliminating the need for post-processing. Span-Image architecture outperforms the state-of-the-art in top-K answer accuracy on SQuAD dataset and in multi-span answer accuracy on an Amazon internal dataset.

****

Wandering within a world: Online contextualized few-shot learning

Mengye Ren,Michael Louis Iuzzolino,Michael Curtis Mozer,Richard Zemel

We aim to bridge the gap between typical human and machine-learning environments by extending the standard framework of few-shot learning to an online, continual setting. In this setting, episodes do not have separate training and testing phases, and instead models are evaluated online while learning novel classes. As in the real world, where the presence of spatiotemporal context helps us retrieve learned skills in the past, our online few-shot learning setting also features an underlying context that changes throughout time. Object classes are correlat

ed within a context and inferring the correct context can lead to better perform
ance. Building upon this setting, we propose a new few-shot learning dataset bas
ed on large scale indoor imagery that mimics the visual
experience of an agent wandering within a world. Furthermore, we convert popular
 few-shot learning approaches into online versions and we also propose a new mod
el that can make use of spatiotemporal contextual information from the recent pa
st.
****

Learning Discrete Adaptive Receptive Fields for Graph Convolutional Networks
Xiaojun Ma,Ziyao Li,Lingjun Xu,Guojie Song,Yi Li,Chuan Shi
Different nodes in a graph neighborhood generally yield different importance. In
 previous work of Graph Convolutional Networks (GCNs), such differences are typi
cally modeled with attention mechanisms. However, as we prove in our paper, soft
 attention weights suffer from over-smoothness in large neighborhoods. To addres
s this weakness, we introduce a novel framework of conducting graph convolutions
, where nodes are discretely selected among multi-hop neighborhoods to construct
 adaptive receptive fields (ARFs). ARFs enable GCNs to get rid of the over-smoot
hness of soft attention weights, as well as to efficiently explore long-distance
 dependencies in graphs. We further propose GRARF (GCN with Reinforced Adaptive
Receptive Fields) as an instance, where an optimal policy of constructing ARFs i
s learned with reinforcement learning. GRARF achieves or matches state-of-the-ar
t performances on public datasets from different domains. Our further analysis c
orroborates that GRARF is more robust than attention models against neighborhood
 noises.
****

PanRep: Universal node embeddings for heterogeneous graphs
Vassilis N. Ioannidis,Da Zheng,George Karypis
Learning unsupervised node embeddings facilitates several downstream tasks such
as node classification and link prediction. A node embedding is universal if it
is designed to be used by and benefit various downstream tasks. This work introd
uces PanRep, a graph neural network (GNN) model, for unsupervised learning of un
iversal node representations for heterogenous graphs. PanRep consists of a GNN e
ncoder that obtains node embeddings and four decoders, each capturing different
topological and node feature properties. Abiding to these properties the novel u
nsupervised framework learns universal embeddings applicable to different downst
ream tasks. PanRep can be furthered fine-tuned to account for possible limited l
abels. In this operational setting PanRep is considered as a pretrained model fo
r extracting node embeddings of heterogenous graph data. PanRep outperforms all
unsupervised and certain supervised methods in node classification and link pred
iction, especially when the labeled data for the supervised methods is small. Pa
nRep-FT (with fine-tuning) outperforms all other supervised approaches, which co
rroborates the merits of pretraining models. Finally, we apply PanRep-FT for dis
covering novel drugs for Covid-19. We showcase the advantage of universal embedd
ings in drug repurposing and identify several drugs used in clinical trials as p
ossible drug candidates.
****

Learning-based Support Estimation in Sublinear Time
Talya Eden,Piotr Indyk,Shyam Narayanan,Ronitt Rubinfeld,Sandeep Silwal,Tal Wagne
r
We consider the problem of estimating the number of distinct elements in a larg
e data set (or, equivalently, the support size of the distribution induced by th
e data set) from a random sample of its elements. The problem occurs in many app
lications, including biology, genomics, computer systems and linguistics. A line
 of research spanning the last decade resulted in algorithms that estimate the s
upport up to $ \pm \varepsilon n$ from a sample of size $O(\log^2(1/\varepsilon)
 \cdot n/\log n)$, where n is the data set size. Unfortunately, this bound is
 known to be tight, limiting further improvements to the complexity of this prob
lem. In this paper we consider estimation algorithms augmented with a machine-le
arning-based predictor that, given any element, returns an estimation of its fr
equency. We show that if the predictor is correct up to a constant approximatio

n factor, then the sample complexity can be reduced significantly, to
$$ \ \log (1/\varepsilon) \cdot n^{1-\Theta(1/\log(1/\varepsilon))}. $$
We evaluate the proposed algorithms on a collection of data sets, using the neural-network based estimators from {Hsu et al, ICLR'19} as predictors. Our experiments demonstrate substantial (up to 3x) improvements in the estimation accuracy compared to the state of the art algorithm.
****

Neural Delay Differential Equations
Qunxi Zhu,Yao Guo,Wei Lin
Neural Ordinary Differential Equations (NODEs), a framework of continuous-depth neural networks, have been widely applied, showing exceptional efficacy in coping with some representative datasets. Recently, an augmented framework has been successfully developed for conquering some limitations emergent in application of the original framework. Here we propose a new class of continuous-depth neural networks with delay, named as Neural Delay Differential Equations (NDDEs), and, for computing the corresponding gradients, we use the adjoint sensitivity method to obtain the delayed dynamics of the adjoint. Since the differential equations with delays are usually seen as dynamical systems of infinite dimension possessing more fruitful dynamics, the NDDEs, compared to the NODEs, own a stronger capacity of nonlinear representations. Indeed, we analytically validate that the NDDEs are of universal approximators, and further articulate an extension of the NDDEs, where the initial function of the NDDEs is supposed to satisfy ODEs. More importantly, we use several illustrative examples to demonstrate the outstanding capacities of the NDDEs and the NDDEs with ODEs' initial value. More precisely, (1) we successfully model the delayed dynamics where the trajectories in the lower-dimensional phase space could be mutually intersected, while the traditional NODEs without any argumentation are not directly applicable for such modeling, and (2) we achieve lower loss and higher accuracy not only for the data produced synthetically by complex models but also for the real-world image datasets, i.e., CIFAR10, MNIST and SVHN. Our results on the NDDEs reveal that appropriately articulating the elements of dynamical systems into the network design is truly beneficial to promoting the network performance.
****

Discriminative Cross-Modal Data Augmentation for Medical Imaging Applications
Yue Yang,Pengtao Xie
While deep learning methods have shown great success in medical image analysis, they require a number of medical images to train. Due to data privacy concerns and unavailability of medical annotators, it is oftentimes very difficult to obtain a lot of labeled medical images for model training. In this paper, we study cross-modality data augmentation to mitigate the data deficiency issue in medical imaging domain. We propose a discriminative unpaired image-to-image translation model which translate images in source modality into images in target modality where the translation task is conducted jointly with the downstream prediction task and the translation is guided by the prediction. Experiments on two applications demonstrate the effectiveness of our method.
****

Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning
Ruozi Huang,Huang Hu,Wei Wu,Kei Sawada,Mi Zhang,Daxin Jiang
Dancing to music is one of human's innate abilities since ancient times. In machine learning research, however, synthesizing dance movements from music is a challenging problem. Recently, researchers synthesize human motion sequences through autoregressive models like recurrent neural network (RNN). Such an approach often generates short sequences due to an accumulation of prediction errors that are fed back into the neural network. This problem becomes even more severe in the long motion sequence generation. Besides, the consistency between dance and music in terms of style, rhythm and beat is yet to be taken into account during modeling. In this paper, we formalize the music-driven dance generation as a sequence-to-sequence learning problem and devise a novel seq2seq architecture to efficiently process long sequences of music features and capture the fine-grained correspondence between music and dance. Furthermore, we propose a novel curriculum

learning strategy to alleviate error accumulation of autoregressive models in long motion sequence generation, which gently changes the training process from a fully guided teacher-forcing scheme using the previous ground-truth movements, towards a less guided autoregressive scheme mostly using the generated movements instead. Extensive experiments show that our approach significantly outperforms the existing state-of-the-arts on automatic metrics and human evaluation. We also make a demo video to demonstrate the superior performance of our proposed approach at https://www.youtube.com/watch?v=lmE20MEheZ8.
****
Diversity Actor-Critic: Sample-Aware Entropy Regularization for Sample-Efficient Exploration
Seungyul Han,Youngchul Sung
Policy entropy regularization is commonly used for better exploration in deep reinforcement learning (RL). However, policy entropy regularization is sample-inefficient in off-policy learning since it does not take the distribution of previous samples stored in the replay buffer into account. In order to take advantage of the previous sample distribution from the replay buffer for sample-efficient exploration, we propose sample-aware entropy regularization which maximizes the entropy of weighted sum of the policy action distribution and the sample action distribution from the replay buffer. We formulate the problem of sample-aware entropy regularized policy iteration, prove its convergence, and provide a practical algorithm named diversity actor-critic (DAC) which is a generalization of soft actor-critic (SAC). Numerical results show that DAC significantly outperforms SAC baselines and other state-of-the-art RL algorithms.
****
Adversarial Attacks on Binary Image Recognition Systems
Eric Balkanski,Harrison Chase,Kojin Oshiba,Alexander Rilee,Yaron Singer,Richard Wang
We initiate the study of adversarial attacks on models for binary (i.e. black and white) image classification. Although there has been a great deal of work on attacking models for colored and grayscale images, little is known about attacks on models for binary images. Models trained to classify binary images are used in text recognition applications such as check processing, license plate recognition, invoice processing, and many others. In contrast to colored and grayscale images, the search space of attacks on binary images is extremely restricted and noise cannot be hidden with minor perturbations in each pixel. Thus, the optimization landscape of attacks on binary images introduces new fundamental challenges.

In this paper we introduce a new attack algorithm called Scar, designed to fool classifiers of binary images. We show that Scar significantly outperforms existing L0 attacks applied to the binary setting and use it to demonstrate the vulnerability of real-world text recognition systems. Scar’s strong performance in practice contrasts with hardness results that show the existence of worst-case classifiers for binary images that are robust to large perturbations. In many cases, altering a single pixel is sufficient to trick Tesseract, a popular open-source text recognition system, to misclassify a word as a different word in the English dictionary. We also demonstrate the vulnerability of check recognition by fooling commercial check processing systems used by major US banks for mobile deposits. These systems are substantially harder to fool since they classify both the handwritten amounts in digits and letters, independently. Nevertheless, we generalize Scar to design attacks that fool state-of-the-art check processing systems using unnoticeable perturbations that lead to misclassification of deposit amounts. Consequently, this is a powerful method to perform financial fraud.
****
Dissecting graph measures performance for node clustering in LFR parameter space
Vladimir Ivashkin,Pavel Chebotarev
Graph measures can be used for graph node clustering using metric clustering algorithms. There are multiple measures applicable to this task, and which one performs better is an open question. We study the performance of 25 graph measures o

n generated graphs with different parameters. While usually measure comparisons are limited to general measure ranking on a particular dataset, we aim to explor e the performance of various measures depending on graph features. Using an LFR generator, we create a dataset of ~7500 graphs covering the whole LFR parameter space. For each graph, we assess the quality of clustering with k-means algorith m for every considered measure. We determine the best measure for every area of the parameter space. We find that the parameter space consists of distinct zones where one particular measure is the best. We analyze the geometry of the result ing zones and describe it with simple criteria. Given particular graph parameter s, this allows us to choose the best measure to use for clustering.
****

Learning Parametrised Graph Shift Operators
George Dasoulas,Johannes F. Lutzeyer,Michalis Vazirgiannis
In many domains data is currently represented as graphs and therefore, the graph representation of this data becomes increasingly important in machine learning. Network data is, implicitly or explicitly, always represented using a graph shi ft operator (GSO) with the most common choices being the adjacency, Laplacian ma trices and their normalisations. In this paper, a novel parametrised GSO (PGSO) is proposed, where specific parameter values result in the most commonly used GS Os and message-passing operators in graph neural network (GNN) frameworks. The P GSO is suggested as a replacement of the standard GSOs that are used in state-of -the-art GNN architectures and the optimisation of the PGSO parameters is seamle ssly included in the model training. It is proved that the PGSO has real eigenva lues and a set of real eigenvectors independent of the parameter values and spec tral bounds on the PGSO are derived. PGSO parameters are shown to adapt to the s parsity of the graph structure in a study on stochastic blockmodel networks, whe re they are found to automatically replicate the GSO regularisation found in the literature. On several real-world datasets the accuracy of state-of-the-art GNN architectures is improved by the inclusion of the PGSO in both node- and graph- classification tasks.
****

An empirical study of a pruning mechanism
Minju Jung,Hyounguk Shon,Eojindl Yi,SungHyun Baek,Junmo Kim
Many methods aim to prune neural network to the maximum extent. However, there a re few studies that investigate the pruning mechanism. In this work, we empirica lly investigate a standard framework for network pruning: pretraining large netw ork and then pruning and retraining it. The framework has been commonly used bas ed on heuristics, i.e., finding a good minima with a large network (pretraining phase) and retaining it with careful pruning and retraining (pruning and retrain ing phase). For the pretraining phase, the reason for which the large network is required to achieve good performance is examined. We hypothesize that this migh t come from the network relying on only a portion of its weights when trained fr om scratch. This way of weight utilization is referred to as imbalanced utility. The measures for weight utility and utility imbalance are proposed. We investig ate the cause of the utility imbalance and the characteristics of the weight uti lity. For the pruning and retraining phase, whether the pruned-and-retrained net work benefits from the pretrained network indded is examined. We visualize the a ccuracy surface of the pretrained, pruned and retrained networks and investigate the relation between them. The validation accuracy is also interpreted in assoc iation with the surface.
****

Augmented Sliced Wasserstein Distances
Xiongjie Chen,Yongxin Yang,Yunpeng Li
While theoretically appealing, the application of the Wasserstein distance to la rge-scale machine learning problems has been hampered by its prohibitive computa tional cost. The sliced Wasserstein distance and its variants improve the comput ational efficiency through random projection, yet they suffer from low projectio n efficiency because the majority of projections result in trivially small value s. In this work, we propose a new family of distance metrics, called augmented s liced Wasserstein distances (ASWDs), constructed by first mapping samples to hig

her-dimensional hypersurfaces parameterized by neural networks. It is derived fr
om a key observation that (random) linear projections of samples residing on the
se hypersurfaces would translate to much more flexible nonlinear projections in
the original sample space, so they can capture complex structures of the data di
stribution. We show that the hypersurfaces can be optimized by gradient ascent e
fficiently. 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. Nume
rical results demonstrate that the ASWD significantly outperforms other Wasserst
ein variants for both synthetic and real-world problems.
****

News-Driven Stock Prediction Using Noisy Equity State Representation
Xiao Liu,Heyan Huang,Yue Zhang
News-driven stock prediction investigates the correlation between news events an
d stock price movements.
Previous work has considered effective ways for representing news events and the
ir sequences, but rarely exploited the representation of underlying equity state
s.
We address this issue by making use of a recurrent neural network to represent a
n equity state transition sequence, integrating news representation using contex
tualized embeddings as inputs to the state transition mechanism.
Thanks to the separation of news and equity representations, our model can accom
modate additional input factors.
We design a novel random noise factor for modeling influencing factors beyond ne
ws events, and a future event factor to address the delay of news information (e
.g., insider trading).
Results show that the proposed model outperforms strong baselines in the literat
ure.
****

Equivariant Normalizing Flows for Point Processes and Sets
Marin Biloš,Stephan Günnemann
A point process describes how random sets of exchangeable points are generated.
The points usually influence the positions of each other via attractive and repu
lsive forces. To model this behavior, it is enough to transform the samples from
 the uniform process with a sufficiently complex equivariant function. However,
learning the parameters of the resulting process is challenging since the likeli
hood is hard to estimate and often intractable. This leads us to our proposed mo
del - CONFET. Based on continuous normalizing flows, it allows arbitrary interac
tions between points while having tractable likelihood. Experiments on various r
eal and synthetic datasets show the improved performance of our new scalable app
roach.
****

Inverse Constrained Reinforcement Learning
Shehryar Malik,Usman Anwar,Alireza Aghasi,Ali Ahmed
Standard reinforcement learning (RL) algorithms train agents to maximize given r
eward functions. However, many real-world applications of RL require agents to a
lso satisfy certain constraints which may, for example, be motivated by safety c
oncerns. Constrained RL algorithms approach this problem by training agents to m
aximize given reward functions while respecting \textit{explicitly} defined cons
traints. However, in many cases, manually designing accurate constraints is a ch
allenging task. In this work, given a reward function and a set of demonstration
s from an expert that maximizes this reward function while respecting \textit{un
known} constraints, we propose a framework to learn the most likely constraints
that the expert respects. We then train agents to maximize the given reward func
tion subject to the learned constraints. Previous works in this regard have eith
er mainly been restricted to tabular settings or specific types of constraints o
r assume knowledge of transition dynamics of the environment. In contrast, we em
pirically show that our framework is able to learn arbitrary \textit{Markovian}
constraints in high-dimensions in a model-free setting.
****

Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based

Models Reliable?
Anna-Kathrin Kopetzki,Bertrand Charpentier,Daniel Zügner,Sandhya Giri,Stephan Günnemann
Robustness to adversarial perturbations and accurate uncertainty estimation are crucial for reliable application of deep learning in real world settings. Dirichlet-based uncertainty (DBU) models are a family of models that predict the parameters of a Dirichlet distribution (instead of a categorical one) and promise to signal when not to trust their predictions. Untrustworthy predictions are obtained on unknown or ambiguous samples and marked with a high uncertainty by the models.

In this work, we show that DBU models with standard training are not robust w.r.t. three important tasks in the field of uncertainty estimation. First, we evaluate how useful the uncertainty estimates are to (1) indicate correctly classified samples. Our results show that while they are a good indicator on unperturbed data, performance on perturbed data decreases dramatically. (2) We evaluate if uncertainty estimates are able to detect adversarial examples that try to fool classification. It turns out that uncertainty estimates are able to detect FGSM attacks but not able to detect PGD attacks. We further evaluate the reliability of DBU models on the task of (3) distinguishing between in-distribution (ID) and out-of-distribution (OOD) data. To this end, we present the first study of certifiable robustness for DBU models. Furthermore, we propose novel uncertainty attacks that fool models into assigning high confidence to OOD data and low confidence to ID data, respectively.
Both approaches show that detecting OOD samples and distinguishing between ID-data and OOD-data is not robust.

Based on our results, we explore the first approaches to make DBU models more robust. We use adversarial training procedures based on label attacks, uncertainty attacks, or random noise and demonstrate how they affect robustness of DBU models on ID data and OOD data.
****
CoLES: Contrastive learning for event sequences with self-supervision
Dmitrii Babaev,Nikita Ovsov,Ivan A Kireev,Gleb Gusev,Maria Ivanova,Alexander Tuzhilin
We address the problem of self-supervised learning on discrete event sequences generated by real-world users. Self-supervised learning incorporates complex information from the raw data in low-dimensional fixed-length vector representations that could be easily applied in various downstream machine learning tasks. In this paper, we propose a new method CoLES, which adopts contrastive learning, previously used for audio and computer vision domains, to the discrete event sequences domain in a self-supervised setting. Unlike most previous studies, we theoretically justify under mild conditions that the augmentation method underlying CoLES provides representative samples of discrete event sequences. We evaluated CoLES on several public datasets and showed that CoLES representations consistently outperform other methods on different downstream tasks.

****
Unlearnable Examples: Making Personal Data Unexploitable
Hanxun Huang,Xingjun Ma,Sarah Monazam Erfani,James Bailey,Yisen Wang
The volume of "free" data on the internet has been key to the current success of deep learning. However, it also raises privacy concerns about the unauthorized exploitation of personal data for training commercial models. It is thus crucial to develop methods to prevent unauthorized data exploitation. This paper raises the question: can data be made unlearnable for deep learning models? We present a type of error-minimizing noise that can indeed make training examples unlearnable. Error-minimizing noise is intentionally generated to reduce the error of one or more of the training example(s) close to zero, which can trick the model into believing there is "nothing" to learn from these example(s). The noise is restricted to be imperceptible to human eyes, and thus does not affect normal data

utility. We empirically verify the effectiveness of error-minimizing noise in both sample-wise and class-wise forms. We also demonstrate its flexibility under extensive experimental settings and practicability in a case study of face recognition. Our work establishes an important ■rst step towards making personal data unexploitable to deep learning models.

****

ChePAN: Constrained Black-Box Uncertainty Modelling with Quantile Regression

Axel Brando,Joan Gimeno,Jose Antonio Rodriguez-Serrano,Jordi Vitria

Most predictive systems currently in use do not report any useful information for auditing their associated uncertainty and evaluating the corresponding risk. Taking it for granted that their replacement may not be advisable in the short term, in this paper we propose a novel approach to modelling confidence in such systems while preserving their predictions. The method is based on the Chebyshev Polynomial Approximation Network (the ChePAN), a new way of modelling aleatoric uncertainty in a regression scenario. In the case addressed here, uncertainty is modelled by building conditional quantiles on top of the original pointwise forecasting system considered as a black box, i.e. without making assumptions about its internal structure. Furthermore, the ChePAN allows users to consistently choose how to constrain any predicted quantile with respect to the original forecaster. Experiments show that the proposed method scales to large size data sets and transfers the advantages of quantile regression to estimating black-box uncertainty.

****

Practical Evaluation of Out-of-Distribution Detection Methods for Image Classification

Engkarat Techapanurak,Takayuki Okatani

We reconsider the evaluation of OOD detection methods for image recognition. Although many studies have been conducted so far to build better OOD detection methods, most of them follow Hendrycks and Gimpel's work for the method of experimental evaluation. While the unified evaluation method is necessary for a fair comparison, there is a question of if its choice of tasks and datasets reflect real-world applications and if the evaluation results can generalize to other OOD detection application scenarios. In this paper, we experimentally evaluate the performance of representative OOD detection methods for three scenarios, i.e., irrelevant input detection, novel class detection, and domain shift detection, on various datasets and classification tasks. The results show that differences in scenarios and datasets alter the relative performance among the methods. Our results can also be used as a guide for practitioners for the selection of OOD detection methods.

****

Rao-Blackwellizing the Straight-Through Gumbel-Softmax Gradient Estimator

Max B Paulus,Chris J. Maddison,Andreas Krause

Gradient estimation in models with discrete latent variables is a challenging problem, because the simplest unbiased estimators tend to have high variance. To counteract this, modern estimators either introduce bias, rely on multiple function evaluations, or use learned, input-dependent baselines. Thus, there is a need for estimators that require minimal tuning, are computationally cheap, and have low mean squared error. In this paper, we show that the variance of the straight-through variant of the popular Gumbel-Softmax estimator can be reduced through Rao-Blackwellization without increasing the number of function evaluations. This provably reduces the mean squared error. We empirically demonstrate that this leads to variance reduction, faster convergence, and generally improved performance in two unsupervised latent variable models.

****

Unsupervised Task Clustering for Multi-Task Reinforcement Learning

Johannes Ackermann,Oliver Paul Richter,Roger Wattenhofer

Meta-learning, transfer learning and multi-task learning have recently laid a path towards more generally applicable reinforcement learning agents that are not limited to a single task. However, most existing approaches implicitly assume a uniform similarity between tasks. We argue that this assumption is limiting in s

ettings where the relationship between tasks is unknown a-priori. In his work, w
e propose a general approach to automatically cluster together similar tasks dur
ing training. Our method, inspired by the expectation-maximization algorithm, su
cceeds at finding clusters of related tasks and uses these to improve sample com
plexity. We achieve this by designing an agent with multiple policies. In the ex
pectation step, we evaluate the performance of the policies on all tasks and ass
ign each task to the best performing policy. In the maximization step, each poli
cy trains by sampling tasks from its assigned set. This method is intuitive, sim
ple to implement and orthogonal to other multi-task learning algorithms. We show
 the generality of our approach by evaluating on simple discrete and continuous
control tasks, as well as complex bipedal walker tasks and Atari games. Results
show improvements in sample complexity as well as a more general applicability w
hen compared to other approaches.
****

Symbol-Shift Equivariant Neural Networks
David Salinas,Hady Elsahar
Neural networks have been shown to have poor compositionality abilities: while t
hey can produce sophisticated output given sufficient data, they perform patchy
generalization and fail to generalize to new symbols (e.g. switching a name in a
 sentence by a less frequent one or one not seen yet). In this paper, we define
a class of models whose outputs are equivariant to entity permutations (an analo
g being convolution networks whose outputs are invariant through translation) wi
thout requiring to specify or detect entities in a pre-processing step. We then
show how two question-answering models can be made robust to entity permutation
using a novel differentiable hybrid semantic-symbolic representation. The benefi
ts of this approach are demonstrated on a set of synthetic NLP tasks where sampl
e complexity and generalization are significantly improved even allowing models
to generalize to words that are never seen in the training set. When using only
1K training examples for bAbi, we obtain a test error of 1.8% and fail only one
task while the best results reported so far obtained an error of 9.9% and failed
 7 tasks.

****

Drift Detection in Episodic Data: Detect When Your Agent Starts Faltering
Ido Greenberg,Shie Mannor
Detection of deterioration of agent performance in dynamic environments is chall
enging due to the non-i.i.d nature of the observed performance. We consider an e
pisodic framework, where the objective is to detect when an agent begins to falt
er. We devise a hypothesis testing procedure for non-i.i.d rewards, which is opt
imal under certain conditions. To apply the procedure sequentially in an online
manner, we also suggest a novel Bootstrap mechanism for False Alarm Rate control
 (BFAR). We demonstrate our procedure in problems where the rewards are not inde
pendent, nor identically-distributed, nor normally-distributed. The statistical
power of the new testing procedure is shown to outperform alternative tests - of
ten by orders of magnitude - for a variety of environment modifications (which c
ause deterioration in agent performance). Our detection method is entirely exter
nal to the agent, and in particular does not require model-based learning. Furth
ermore, it can be applied to detect changes or drifts in any episodic signal.
****

Structural Landmarking and Interaction Modelling: on Resolution Dilemmas in Grap
h Classification
Kai Zhang,Yaokang Zhu,Jun Wang,Haibin Ling,Jie Zhang,Hongyuan Zha
Graph neural networks are promising architecture for learning and inference with
 graph-structured data. However, generating informative graph level features has
 long been a challenge. Current practice of graph-pooling typically summarizes
 a graph by squeezing it into a single vector. This may lead to significant loss
 of predictive, iterpretable structural information, because properties of a com
plex system are believed to arise largely from the interaction among its compone
nts. In this paper, we analyze the intrinsic difficulty in graph classification
under the unified concept of ``"resolution dilemmas" and propose `SLIM, an induc

tive neural network model for Structural Landmarking and Interaction Modelling, to remedy the information loss in graph pooling. We show that, by projecting graphs onto end-to-end optimizable, and well-aligned substructure landmarks (representatives), the resolution dilemmas can be resolved effectively, so that explicit interacting relation between component parts of a graph can be leveraged directly in explaining its complexity and predicting its property. Empirical evaluations, in comparison with state-of-the-art, demonstrate promising results of our approach on a number of benchmark datasets for graph classification.

****

Ruminating Word Representations with Random Noise Masking
Hwiyeol Jo,Byoung-Tak Zhang
We introduce a training method for better word representation and performance, which we call \textbf{GraVeR} (\textbf{Gra}dual \textbf{Ve}ctor \textbf{R}umination). The method is to gradually and iteratively add random noises and bias to word embeddings after training a model, and re-train the model from scratch but initialize with the noised word embeddings. Through the re-training process, some noises can be compensated and other noises can be utilized to learn better representations. As a result, we can get word representations further fine-tuned and specialized in the task. On six text classification tasks, our method improves model performances with a large gap. When GraVeR is combined with other regularization techniques, it shows further improvements. Lastly, we investigate the usefulness of GraVeR.
****

DOP: Off-Policy Multi-Agent Decomposed Policy Gradients
Yihan Wang,Beining Han,Tonghan Wang,Heng Dong,Chongjie Zhang
Multi-agent policy gradient (MAPG) methods recently witness vigorous progress. However, there is a significant performance discrepancy between MAPG methods and state-of-the-art multi-agent value-based approaches. In this paper, we investigate causes that hinder the performance of MAPG algorithms and present a multi-agent decomposed policy gradient method (DOP). This method introduces the idea of value function decomposition into the multi-agent actor-critic framework. Based on this idea, DOP supports efficient off-policy learning and addresses the issue of centralized-decentralized mismatch and credit assignment in both discrete and continuous action spaces. We formally show that DOP critics have sufficient representational capability to guarantee convergence. In addition, empirical evaluations on the StarCraft II micromanagement benchmark and multi-agent particle environments demonstrate that DOP outperforms both state-of-the-art value-based and policy-based multi-agent reinforcement learning algorithms. Demonstrative videos are available at https://sites.google.com/view/dop-mapg/.
****

Time-varying Graph Representation Learning via Higher-Order Skip-Gram with Negative Sampling
Simone Piaggesi,André Panisson
Representation learning models for graphs are a successful family of techniques that project nodes into feature spaces that can be exploited by other machine learning algorithms. Since many real-world networks are inherently dynamic, with interactions among nodes changing over time, these techniques can be defined both for static and for time-varying graphs. Here, we show how the skip-gram embedding approach can be used to perform implicit tensor factorization on different tensor representations of time-varying graphs. We show that higher-order skip-gram with negative sampling (HOSGNS) is able to disentangle the role of nodes and time, with a small fraction of the number of parameters needed by other approaches. We empirically evaluate our approach using time-resolved face-to-face proximity data, showing that the learned representations outperform state-of-the-art methods when used to solve downstream tasks such as network reconstruction. Good performance on predicting the outcome of dynamical processes such as disease spreading shows the potential of this new method to estimate contagion risk, providing early risk awareness based on contact tracing data.
****

Unsupervised Discovery of 3D Physical Objects from Video
Yilun Du,Kevin A. Smith,Tomer Ullman,Joshua B. Tenenbaum,Jiajun Wu
We study the problem of unsupervised physical object discovery. While existing f
rameworks aim to decompose scenes into 2D segments based off each object's appea
rance, we explore how physics, especially object interactions, facilitates disen
tangling of 3D geometry and position of objects from video, in an unsupervised m
anner. Drawing inspiration from developmental psychology, our Physical Object Di
scovery Network (POD-Net) uses both multi-scale pixel cues and physical motion c
ues to accurately segment observable and partially occluded objects of varying s
izes, and infer properties of those objects. Our model reliably segments objects
 on both synthetic and real scenes. The discovered object properties can also b
e used to reason about physical events.
****
NASOA: Towards Faster Task-oriented Online Fine-tuning
Hang Xu,Ning Kang,Gengwei Zhang,Xiaodan Liang,Zhenguo Li
Fine-tuning from pre-trained ImageNet models has been a simple, effective, and p
opular approach for various computer vision tasks. The common practice of fine-t
uning is to adopt a default hyperparameter setting with a fixed pre-trained mode
l, while both of them are not optimized for specific tasks and time constraints.
 Moreover, in cloud computing or GPU clusters where the tasks arrive sequentiall
y in a stream, faster online fine-tuning is a more desired and realistic strateg
y for saving money, energy consumption, and CO2 emission. In this paper, we prop
ose a joint Neural Architecture Search and Online Adaption framework named NASOA
 towards a faster task-oriented fine-tuning upon the request of users. Specifica
lly, NASOA first adopts an offline NAS to identify a group of training-efficient
 networks to form a pretrained model zoo. We propose a novel joint block and mac
ro-level search space to enable a flexible and efficient search. Then, by estima
ting fine-tuning performance via an adaptive model by accumulating experience fr
om the past tasks, an online schedule generator is proposed to pick up the most
suitable model and generate a personalized training regime with respect to each
desired task in a one-shot fashion. The resulting model zoo is more training eff
icient than SOTA NAS models, e.g. 6x faster than RegNetY-16GF, and 1.7x faster t
han EfficientNetB3. Experiments on multiple datasets also show that NASOA achiev
es much better fine-tuning results, i.e. improving around 2.1% accuracy than the
 best performance in RegNet series under various time constraints and tasks; 40x
 faster compared to the BOHB method.
****
Exploring Balanced Feature Spaces for Representation Learning
Bingyi Kang,Yu Li,Sa Xie,Zehuan Yuan,Jiashi Feng
Existing self-supervised learning (SSL) methods are mostly applied for training
representation models from artificially balanced datasets (e.g., ImageNet). It i
s unclear how well they will perform in the practical scenarios where datasets a
re often imbalanced w.r.t. the classes. Motivated by this question, we conduct a
 series of studies on the performance of self-supervised contrastive learning an
d supervised learning methods over multiple datasets where training instance dis
tributions vary from a balanced one to a long-tailed one. Our findings are quite
 intriguing. Different from supervised methods with large performance drop, the
self-supervised contrastive learning methods perform stably well even when the d
atasets are heavily imbalanced. This motivates us to explore the balanced featur
e spaces learned by contrastive learning, where the feature representations pres
ent similar linear separability w.r.t. all the classes. Our further experiments
reveal that a representation model generating a balanced feature space can gener
alize better than that yielding an imbalanced one across multiple settings. Insp
ired by these insights, we develop a novel representation learning method, calle
d k-positive contrastive learning. It effectively combines strengths of the su
pervised method and the contrastive learning method to learn representations tha
t are both discriminative and balanced. Extensive experiments demonstrate its su
periority on multiple recognition tasks. Remarkably, it achieves new state-of-th
e-art on challenging long-tailed recognition benchmarks. Code and models will be
 released.

****
Non-Linear Rewards For Successor Features
Norman L Tasfi,Miriam Capretz
Reinforcement Learning algorithms have reached new heights in performance, often overtaking humans on several challenging tasks such as Atari and Go. However, the resulting models learn fragile 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. Under this framework, transfer between related tasks requires only training the reward component. However, successor features builds upon the limiting assumption that the current reward can be predicted from a linear combination of state features. This paper proposes a novel improvement to the successor feature framework, where we instead assume that the reward function is a non-linear function of the state features, thereby increasing its representational power. After derivation of the new state-action value function, the decomposition includes a second term that learns the auto-correlation matrix between state features. Experimentally, we show 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 improvements to the successor feature framework is validated empirically on navigation tasks and control of a simulated robotic arm.
****
Auxiliary Learning by Implicit Differentiation
Aviv Navon,Idan Achituve,Haggai Maron,Gal Chechik,Ethan Fetaya
Training neural networks with auxiliary tasks is a common practice for improving the performance on a main task of interest.
Two main challenges arise in this multi-task learning setting: (i) designing useful auxiliary tasks; and (ii) combining auxiliary tasks into a single coherent loss. Here, we propose a novel framework, AuxiLearn, that targets both challenges based on implicit differentiation. First, when useful auxiliaries are known, we propose learning a network that combines all losses into a single coherent objective function. This network can learn non-linear interactions between tasks. Second, when no useful auxiliary task is known, we describe how to learn a network that generates a meaningful, novel auxiliary task. We evaluate AuxiLearn in a series of tasks and domains, including image segmentation and learning with attributes in the low data regime, and find that it consistently outperforms competing methods.
****
Exploiting Safe Spots in Neural Networks for Preemptive Robustness and Out-of-Distribution Detection
Seungyong Moon,Gaon An,Hyun Oh Song
Recent advances on adversarial defense mainly focus on improving the classifier’s robustness against adversarially perturbed inputs. In this paper, we turn our attention from classifiers to inputs and explore if there exist safe spots in the vicinity of natural images that are robust to adversarial attacks. In this regard, we introduce a novel bi-level optimization algorithm that can find safe spots on over 90% of the correctly classified images for adversarially trained classifiers on CIFAR-10 and ImageNet datasets. Our experiments also show that they can be used to improve both the empirical and certified robustness on smoothed classifiers. Furthermore, by exploiting a novel safe spot inducing model training scheme and our safe spot generation method, we propose a new out-of-distribution detection algorithm which achieves the state of the art results on near-distribution outliers.
****
Gradient Descent Ascent for Min-Max Problems on Riemannian Manifolds
Feihu Huang,Shangqian Gao,Heng Huang
In the paper, we study a class of useful non-convex minimax optimization problems on Riemanian manifolds and propose a class of Riemanian gradient descent ascent algorithms to solve these minimax problems. Specifically, we propose a new Riemannian gradient descent ascent (RGDA) algorithm for the deterministic minima

x optimization.
Moreover, we prove that the RGDA has a sample complexity of $O(\kappa^2\epsilon^{-2})$ for finding an ϵ-stationary point of the nonconvex strongly-concave minimax problems, where κ denotes the condition number.
At the same time, we introduce a Riemannian stochastic gradient descent ascent (RSGDA) algorithm for the stochastic minimax optimization. In the theoretical analysis, we prove that the RSGDA can achieve a sample complexity of $O(\kappa^4\epsilon^{-4})$.
To further reduce the sample complexity, we propose a novel momentum variance-reduced Riemannian stochastic gradient descent ascent (MVR-RSGDA) algorithm based on a new momentum variance-reduced technique of STORM. We prove that the MVR-RSGDA algorithm achieves a lower sample complexity of $\tilde{O}(\kappa^{4}\epsilon^{-3})$ without large batches, which reaches near the best known sample complexity for its Euclidean counterparts. Extensive experimental results on the robust deep neural networks training over Stiefel manifold demonstrate the efficiency of our proposed algorithms.
****

ProxylessKD: Direct Knowledge Distillation with Inherited Classifier for Face Recognition
Weidong Shi,Guanghui Ren,Yunpeng Chen,Shuicheng Yan
Knowledge Distillation (KD) refers to transferring knowledge from a large model to a smaller one, which is widely used to enhance model performance in machine learning. It tries to align embedding spaces generated from the teacher and the student model (i.e. to make images corresponding to the same semantics share the same embedding across different models). In this work, we focus on its application in face recognition. We observe that existing knowledge distillation models optimize the proxy tasks that force the student to mimic the teacher’s behavior, instead of directly optimizing the face recognition accuracy. Consequently, the obtained student models are not guaranteed to be optimal on the target task or able to benefit from advanced constraints, such as the large margin constraint (e.g. margin-based softmax). We then propose a novel method named ProxylessKD that directly optimizes face recognition accuracy by inheriting the teacher's classifier as the student's classifier to guide the student to learn discriminative embeddings in the teacher's embedding space. The proposed ProxylessKD is very easy to implement and sufficiently generic to be extended to other tasks beyond face recognition. We conduct extensive experiments on standard face recognition benchmarks,
and the results demonstrate that ProxylessKD achieves superior performance over existing knowledge distillation methods.
****

On the Bottleneck of Graph Neural Networks and its Practical Implications
Uri Alon,Eran Yahav
Since the proposal of the graph neural network (GNN) by Gori et al. (2005) and Scarselli et al. (2008), one of the major problems in training GNNs was their struggle to propagate information between distant nodes in the graph.
We propose a new explanation for this problem: GNNs are susceptible to a bottleneck when aggregating messages across a long path. This bottleneck causes the over-squashing of exponentially growing information into fixed-size vectors.
As a result, GNNs fail to propagate messages originating from distant nodes and perform poorly when the prediction task depends on long-range interaction.
In this paper, we highlight the inherent problem of over-squashing in GNNs:
we demonstrate that the bottleneck hinders popular GNNs from fitting long-range signals in the training data;
we further show that GNNs that absorb incoming edges equally, such as GCN and GIN, are more susceptible to over-squashing than GAT and GGNN;
finally, we show that prior work, which extensively tuned GNN models of long-range problems, suffers from over-squashing, and that breaking the bottleneck improves their state-of-the-art results without any tuning or additional weights.
Our code is available at https://github.com/tech-srl/bottleneck/ .
****

Progressively Stacking 2.0: A Multi-stage Layerwise Training Method for BERT Training Speedup
Cheng Yang,Shengnan Wang,Chao Yang,Yuechuan Li,Ru He,Jingqiao Zhang
Pre-trained language models, such as BERT, have achieved significant accuracy gain in many natural language processing tasks. Despite its effectiveness, the huge number of parameters makes training a BERT model computationally very challenging. In this paper, we propose an efficient multi-stage layerwise training (MSLT) approach to reduce the training time of BERT. We decompose the whole training process into several stages. The training is started from a small model with only a few encoder layers and we gradually increase the depth of the model by adding new encoder layers. At each stage, we only train the top (near the output layer) few encoder layers which are newly added. The parameters of the other layers which have been trained in the previous stages will not be updated in the current stage. In BERT training, the backward calculation is much more time-consuming than the forward calculation, especially in the distributed training setting in which the backward calculation time further includes the communication time for gradient synchronization. In the proposed training strategy, only top few layers participate backward calculation, while most layers only participate forward calculation. Hence both the computation and communication efficiencies are greatly improved. Experimental results show that the proposed method can greatly reduce the training time without significant performance degradation.
****
Bayesian Neural Networks with Variance Propagation for Uncertainty Evaluation
Yuki Mae,Wataru Kumagai,Takafumi Kanamori
Uncertainty evaluation is a core technique when deep neural networks (DNNs) are used in real-world problems. In practical applications, we often encounter unexpected samples that have not seen in the training process. Not only achieving the high-prediction accuracy but also detecting uncertain data is significant for safety-critical systems. In statistics and machine learning, Bayesian inference has been exploited for uncertainty evaluation. The Bayesian neural networks (BNNs) have recently attracted considerable attention in this context, as the DNN trained using dropout is interpreted as a Bayesian method. Based on this interpretation, several methods to calculate the Bayes predictive distribution for DNNs have been developed. Though the Monte-Carlo method called MC dropout is a popular method for uncertainty evaluation, it requires a number of repeated feed-forward calculations of DNNs with randomly sampled weight parameters. To overcome the computational issue, we propose a sampling-free method to evaluate uncertainty. Our method converts a neural network trained using the dropout to the corresponding Bayesian neural network with variance propagation. Our method is available not only to feed-forward NNs but also to recurrent NNs including LSTM. We report the computational efficiency and statistical reliability of our method in numerical experiments of the language modeling using RNNs, and the out-of-distribution detection with DNNs.
****
Adaptive Multi-model Fusion Learning for Sparse-Reward Reinforcement Learning
Giseung Park,Whiyoung Jung,Sungho Choi,Youngchul Sung
In this paper, we consider intrinsic reward generation for sparse-reward reinforcement learning based on model prediction errors. In typical model-prediction-error-based intrinsic reward generation, an agent has a learning model for the underlying environment. Then intrinsic reward is designed as the error between the model prediction and the actual outcome of the environment, based on the fact that for less-visited or non-visited states, the learned model yields larger prediction errors, promoting exploration helpful for reinforcement learning. This paper generalizes this model-prediction-error-based intrinsic reward generation method to multiple prediction models. We propose a new adaptive fusion method relevant to the multiple-model case, which learns optimal prediction-error fusion across the learning phase to enhance the overall learning performance. Numerical results show that for representative locomotion tasks, the proposed intrinsic reward generation method outperforms most of the previous methods, and the gain is significant in some tasks.

****
The Role of Momentum Parameters in the Optimal Convergence of Adaptive Polyak's Heavy-ball Methods
Wei Tao,Sheng Long,Gaowei Wu,Qing Tao
The adaptive stochastic gradient descent (SGD) with momentum has been widely adopted in deep learning as well as convex optimization. In practice, the last iterate is commonly used as the final solution. However, the available regret analysis and the setting of constant momentum parameters only guarantee the optimal convergence of the averaged solution. In this paper, we fill this theory-practice gap by investigating the convergence of the last iterate (referred to as {\it individual convergence}), which is a more difficult task than convergence analysis of the averaged solution. Specifically, in the constrained convex cases, we prove that the adaptive Polyak's Heavy-ball (HB) method, in which the step size is only updated using the exponential moving average strategy, attains an individual convergence rate of $O(\frac{1}{\sqrt{t}})$, as opposed to that of $O(\frac{\log t}{\sqrt {t}})$ of SGD, where t is the number of iterations. Our new analysis not only shows how the HB momentum and its time-varying weight help us to achieve the acceleration in convex optimization but also gives valuable hints how the momentum parameters should be scheduled in deep learning. Empirical results validate the correctness of our convergence analysis in optimizing convex functions and demonstrate the improved performance of the adaptive HB methods in training deep networks.
****
How Benign is Benign Overfitting ?
Amartya Sanyal,Puneet K. Dokania,Varun Kanade,Philip Torr
We investigate two causes for adversarial vulnerability in deep neural networks: bad data and (poorly) trained models. When trained with SGD, deep neural networks essentially achieve zero training error, even in the presence of label noise, while also exhibiting good generalization on natural test data, something referred to as benign overfitting (Bartlett et al., 2020; Chatterji & Long, 2020). However, these models are vulnerable to adversarial attacks. We identify label noise as one of the causes for adversarial vulnerability, and provide theoretical and empirical evidence in support of this. Surprisingly, we find several instances of label noise in datasets such as MNIST and CIFAR, and that robustly trained models incur training error on some of these, i.e. they don’t fit the noise. However, removing noisy labels alone does not suffice to achieve adversarial robustness. We conjecture that in part sub-optimal representation learning is also responsible for adversarial vulnerability. By means of simple theoretical setups, we show how the choice of representation can drastically affect adversarial robustness.
****
Entropic gradient descent algorithms and wide flat minima
Fabrizio Pittorino,Carlo Lucibello,Christoph Feinauer,Gabriele Perugini,Carlo Baldassi,Elizaveta Demyanenko,Riccardo Zecchina
The properties of flat minima in the empirical risk landscape of neural networks have been debated for some time. Increasing evidence suggests they possess better generalization capabilities with respect to sharp ones. In this work we first discuss the relationship between alternative measures of flatness: The local entropy, which is useful for analysis and algorithm development, and the local energy, which is easier to compute and was shown empirically in extensive tests on state-of-the-art networks to be the best predictor of generalization capabilities. We show semi-analytically in simple controlled scenarios that these two measures correlate strongly with each other and with generalization. Then, we extend the analysis to the deep learning scenario by extensive numerical validations. We study two algorithms, Entropy-SGD and Replicated-SGD, that explicitly include the local entropy in the optimization objective. We devise a training schedule by which we consistently find flatter minima (using both flatness measures), and improve the generalization error for common architectures (e.g. ResNet, EfficientNet).
****

Variational Dynamic Mixtures
Chen Qiu,Stephan Mandt,Maja Rudolph
Deep probabilistic time series forecasting models have become an integral part o
f machine learning. While several powerful generative models have been proposed,
 we provide evidence that their associated inference models are oftentimes too l
imited and cause the generative model to predict mode-averaged dynamics. Mode-av
eraging is problematic since many real-world sequences are highly multi-modal, a
nd their averaged dynamics are unphysical (e.g., predicted taxi trajectories mig
ht run through buildings on the street map). To better capture multi-modality, w
e develop variational dynamic mixtures (VDM): a new variational family to infer
sequential latent variables. The VDM approximate posterior at each time step is
a mixture density network, whose parameters come from propagating multiple sampl
es through a recurrent architecture. This results in an expressive multi-modal
posterior approximation. In an empirical study, we show that VDM outperforms com
peting approaches on highly multi-modal datasets from different domains.
****

Abductive Knowledge Induction from Raw Data
Wang-Zhou Dai,Stephen Muggleton
For many reasoning-heavy tasks, it is challenging to find an appropriate end-to-
end differentiable approximation to domain-specific inference mechanisms. Neural
-Symbolic (NeSy) AI divides the end-to-end pipeline into neural perception and s
ymbolic reasoning, which can directly exploit general domain knowledge such as a
lgorithms and logic rules. However, it suffers from the exponential computationa
l complexity caused by the interface between the two components, where the neura
l model lacks direct supervision, and the symbolic model lacks accurate input fa
cts. As a result, they usually focus on learning the neural model with a sound a
nd complete symbolic knowledge base while avoiding a crucial problem: where does
 the knowledge come from? In this paper, we present Abductive Meta-Interpretive
Learning ($Meta_{Abd}$), which unites abduction and induction to learn perceptua
l neural network and first-order logic theories simultaneously from raw data. Gi
ven the same amount of domain knowledge, we demonstrate that $Meta_{Abd}$ not on
ly outperforms the compared end-to-end models in predictive accuracy and data ef
ficiency but also induces logic programs that can be re-used as background knowl
edge in subsequent learning tasks. To the best of our knowledge, $Meta_{Abd}$ is
 the first system that can jointly learn neural networks and recursive first-ord
er logic theories with predicate invention.
****

Learning Predictive Communication by Imagination in Networked System Control
Yali Du,Yifan Zhao,Meng Fang,Jun Wang,Gangyan Xu,Haifeng Zhang
Dealing with multi-agent control in networked systems is one of the biggest chal
lenges in Reinforcement Learning (RL) and limited success has been presented com
pared to recent deep reinforcement learning in single-agent domain. However, obs
tacles remain in addressing the delayed global information where each agent lear
ns a decentralized control policy based on local observations and messages from
connected neighbors. This paper first considers delayed global information shari
ng by combining the delayed global information and latent imagination of farsigh
ted states in differentiable communication. Our model allows an agent to imagine
 its future states and communicate that with its neighbors. The predictive messa
ge sent to the connected neighbors reduces the delay in global information. On
the tasks of networked multi-agent traffic control, experimental results show th
at our model helps stabilize the training of each local agent and outperforms ex
isting algorithms for networked system control.
****

Signed Graph Diffusion Network
Jinhong Jung,Jaemin Yoo,U Kang
Given a signed social graph, how can we learn appropriate node representations t
o infer the signs of missing edges?
Signed social graphs have received considerable attention to model trust relatio
nships.
Learning node representations is crucial to effectively analyze graph data, and

various techniques such as network embedding and graph convolutional network (GCN) have been proposed for learning signed graphs.
However, traditional network embedding methods are not end-to-end for a specific task such as link sign prediction, and GCN-based methods suffer from a performance degradation problem when their depth increases.
In this paper, we propose Signed Graph Diffusion Network (SGDNet), a novel graph neural network that achieves end-to-end node representation learning for link sign prediction in signed social graphs.
We propose a random walk technique specially designed for signed graphs so that SGDNet effectively diffuses hidden node features.
Through extensive experiments, we demonstrate that SGDNet outperforms state-of-the-art models in terms of link sign prediction accuracy.
****

Explicit Connection Distillation
Lujun Li,Yikai Wang,Anbang Yao,Yi Qian,Xiao Zhou,Ke He
One effective way to ease the deployment of deep neural networks on resource constrained devices is Knowledge Distillation (KD), which boosts the accuracy of a low-capacity student model by mimicking the learnt information of a high-capacity teacher (either a single model or a multi-model ensemble). Although great progress has been attained on KD research, existing efforts are primarily invested to design better distillation losses by using soft logits or intermediate feature representations of the teacher as the extra supervision. In this paper, we present Explicit Connection Distillation (ECD), a new KD framework, which addresses the knowledge distillation problem in a novel perspective of bridging dense intermediate feature connections between a student network and its corresponding teacher generated automatically in the training, achieving knowledge transfer goal via direct cross-network layer-to-layer gradients propagation. ECD has two interdependent modules. In the first module, given a student network, an auxiliary teacher architecture is temporarily generated conditioned on strengthening feature representations of basic convolutions of the student network via replacing them with dynamic additive convolutions and keeping the other layers unchanged in structure. The teacher generated in this way guarantees its superior capacity and makes a perfect feature alignment (both in input and output dimensions) to the student at every convolutional layer. In the second module, dense feature connections between the aligned convolutional layers from the student to its auxiliary teacher are introduced, which allows explicit layer-to-layer gradients propagation from the teacher to the student via the merged model training from scratch. Intriguingly, as feature connection direction is one-way, all feature connections together with the auxiliary teacher merely exist during training phase. Experiments on popular image classification tasks validate the effectiveness of our method. Code will be made publicly available.
****

ADIS-GAN: Affine Disentangled GAN
Letao Liu,Martin Saerbeck,Justin Dauwels
This paper proposes Affine Disentangled GAN (ADIS-GAN), which is a Generative Adversarial Network that can explicitly disentangle affine transformations in a self-supervised and rigorous manner. The objective is inspired by InfoGAN, where an additional affine regularizer acts as the inductive bias. The affine regularizer is rooted in the affine transformation properties of images, changing some properties of the underlying images, while leaving all other properties invariant. We derive the affine regularizer by decomposing the affine matrix into separate transformation matrices and inferring the transformation parameters by maximum-likelihood estimation. Unlike the disentangled representations learned by existing approaches, the features learned by ADIS-GAN are axis-aligned and scalable, where transformations such as rotation, horizontal and vertical zoom, horizontal and vertical skew, horizontal and vertical translation can be explicitly selected and learned. ADIS-GAN successfully disentangles these features on the MNIST, CelebA, and dSprites datasets.
****

SEDONA: Search for Decoupled Neural Networks toward Greedy Block-wise Learning

Myeongjang Pyeon,Jihwan Moon,Taeyoung Hahn,Gunhee Kim
Backward locking and update locking are well-known sources of inefficiency in backpropagation that prevent from concurrently updating layers. Several works have recently suggested using local error signals to train network blocks asynchronously to overcome these limitations. However, they often require numerous iterations of trial-and-error to find the best configuration for local training, including how to decouple network blocks and which auxiliary networks to use for each block. In this work, we propose a differentiable search algorithm named SEDONA to automate this process. Experimental results show that our algorithm can consistently discover transferable decoupled architectures for VGG and ResNet variants, and significantly outperforms the ones trained with end-to-end backpropagation and other state-of-the-art greedy-leaning methods in CIFAR-10, Tiny-ImageNet and ImageNet.
****
Coupled Oscillatory Recurrent Neural Network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies
T. Konstantin Rusch,Siddhartha Mishra
Circuits of biological neurons, such as in the functional parts of the brain can be modeled as networks of coupled oscillators. Inspired by the ability of these systems to express a rich set of outputs while keeping (gradients of) state variables bounded, we propose a novel architecture for recurrent neural networks. Our proposed RNN is based on a time-discretization of a system of second-order ordinary differential equations, modeling networks of controlled nonlinear oscillators. We prove precise bounds on the gradients of the hidden states, leading to the mitigation of the exploding and vanishing gradient problem for this RNN. Experiments show that the proposed RNN is comparable in performance to the state of the art on a variety of benchmarks, demonstrating the potential of this architecture to provide stable and accurate RNNs for processing complex sequential data.
****
MVP: Multivariate polynomials for conditional generation
Grigorios Chrysos,Yannis Panagakis
Conditional Generative Adversarial Nets (cGANs) have been widely adopted for image generation. cGANs take i) a noise vector and ii) a conditional variable as input. The conditional variable can be discrete (e.g., a class label) or continuous (e.g., an input image) resulting into class-conditional (image) generation and image-to-image translation models, respectively. However, depending on whether the conditional variable is discrete or continuous, various cGANs employ substantially different deep architectures and loss functions for their training. In this paper, we propose a novel framework, called MVP, for conditional data generation. MVP resorts to multivariate polynomials of higher-order and treats in a unified way both discrete and continuous conditional variables. MVP is highly expressive, capturing higher-order auto- and cross-correlations of input variables (noise vector and conditional variable). Tailored sharing schemes are designed between the polynomial’s parameter tensors, which result in simple recursive formulas. MVP can synthesize realistic images in both class-conditional and image-to-image translation tasks even in the absence of activation functions between the layers.
****
not-MIWAE: Deep Generative Modelling with Missing not at Random Data
Niels Bruun Ipsen,Pierre-Alexandre Mattei,Jes Frellsen
When a missing process depends on the missing values themselves, it needs to be explicitly modelled and taken into account while doing likelihood-based inference. We present an approach for building and fitting deep latent variable models (DLVMs) in cases where the missing process is dependent on the missing data. Specifically, a deep neural network enables us to flexibly model the conditional distribution of the missingness pattern given the data. This allows for incorporating prior information about the type of missingness (e.g.~self-censoring) into the model. Our inference technique, based on importance-weighted variational inference, involves maximising a lower bound of the joint likelihood. Stochastic grad

ients of the bound are obtained by using the reparameterisation trick both in la tent space and data space. We show on various kinds of data sets and missingness patterns that explicitly modelling the missing process can be invaluable.
****

Reweighting Augmented Samples by Minimizing the Maximal Expected Loss
Mingyang Yi,Lu Hou,Lifeng Shang,Xin Jiang,Qun Liu,Zhi-Ming Ma
Data augmentation is an effective technique to improve the generalization of dee p neural networks. However, previous data augmentation methods usually treat the augmented samples equally without considering their individual impacts on the m odel. To address this, for the augmented samples from the same training example, we propose to assign different weights to them. We construct the maximal expect ed loss which is the supremum over any reweighted loss on augmented samples. Ins pired by adversarial training, we minimize this maximal expected loss (MMEL) and obtain a simple and interpretable closed-form solution: more attention should b e paid to augmented samples with large loss values (i.e., harder examples). Mini mizing this maximal expected loss enables the model to perform well under any re weighting strategy. The proposed method can generally be applied on top of any d ata augmentation methods. Experiments are conducted on both natural language und erstanding tasks with token-level data augmentation, and image classification ta sks with commonly-used image augmentation techniques like random crop and horizo ntal flip. Empirical results show that the proposed method improves the generali zation performance of the model.
****

Learning to Represent Action Values as a Hypergraph on the Action Vertices
Arash Tavakoli,Mehdi Fatemi,Petar Kormushev
Action-value estimation is a critical component of many reinforcement learning (RL) methods whereby sample complexity relies heavily on how fast a good estimato r for action value can be learned. By viewing this problem through the lens of r epresentation learning, good representations of both state and action can facili tate action-value estimation. While advances in deep learning have seamlessly dr iven progress in learning state representations, given the specificity of the no tion of agency to RL, little attention has been paid to learning action represen tations. We conjecture that leveraging the combinatorial structure of multi-dime nsional action spaces is a key ingredient for learning good representations of a ction. To test this, we set forth the action hypergraph networks framework---a c lass of functions for learning action representations in multi-dimensional discr ete action spaces with a structural inductive bias. Using this framework we real ise an agent class based on a combination with deep Q-networks, which we dub hyp ergraph Q-networks. We show the effectiveness of our approach on a myriad of dom ains: illustrative prediction problems under minimal confounding effects, Atari 2600 games, and discretised physical control benchmarks.
****

Local Information Opponent Modelling Using Variational Autoencoders
Georgios Papoudakis,Filippos Christianos,Stefano V Albrecht
Modelling the behaviours of other agents (opponents) is essential for understand ing how agents interact and making effective decisions. Existing methods for opp onent modelling commonly assume knowledge of the local observations and chosen a ctions of the modelled opponents, which can significantly limit their applicabil ity. We propose a new modelling technique based on variational autoencoders, whi ch are trained to reconstruct the local actions and observations of the opponent based on embeddings which depend only on the local observations of the modellin g agent (its observed world state, chosen actions, and received rewards). The em beddings are used to augment the modelling agent's decision policy which is trai ned via deep reinforcement learning; thus the policy does not require access to opponent observations. We provide a comprehensive evaluation and ablation study in diverse multi-agent tasks, showing that our method achieves comparable perfo rmance to an ideal baseline which has full access to opponent's information, and significantly higher returns than a baseline method which does not use the lear ned embeddings.
****

Robust Learning Rate Selection for Stochastic Optimization via Splitting Diagnostic
Matteo Sordello,Hangfeng He,Weijie J Su
This paper proposes SplitSGD, a new dynamic learning rate schedule for stochastic optimization. This method decreases the learning rate for better adaptation to the local geometry of the objective function whenever a stationary phase is detected, that is, the iterates are likely to bounce at around a vicinity of a local minimum. The detection is performed by splitting the single thread into two and using the inner product of the gradients from the two threads as a measure of stationarity. Owing to this simple yet provably valid stationarity detection, SplitSGD is easy- to-implement and essentially does not incur additional computational cost than standard SGD. Through a series of extensive experiments, we show that this method is appropriate for both convex problems and training (non-convex) neural networks, with performance compared favorably to other stochastic optimization methods. Importantly, this method is observed to be very robust with a set of default parameters for a wide range of problems and, moreover, yields better generalization performance than other adaptive gradient methods such as Adam.

****

Quantitative Understanding of VAE as a Non-linearly Scaled Isometric Embedding
Akira Nakagawa,Keizo Kato
Variational autoencoder (VAE) estimates the posterior parameters (mean and variance) of latent variables corresponding to each input data. While it is used for many tasks, the transparency of the model is still an underlying issue. This paper provides a quantitative understanding of VAE property by interpreting VAE as a non-linearly scaled isometric embedding. According to the Rate-distortion theory, the optimal transform coding is achieved by using a PCA-like orthonormal transform where the transform space is isometric to the input. From this analogy, we show theoretically and experimentally that VAE can be mapped to an implicit isometric embedding with a scale factor derived from the posterior parameter. As a result, we can estimate the data probabilities in the input space from the prior, loss metrics, and corresponding posterior parameters. In addition, the quantitative importance of each latent variable can be evaluated like the eigenvalue of PCA.

****

Learning Deeply Shared Filter Bases for Efficient ConvNets
Woochul Kang,Daeyeon Kim
Recently, inspired by repetitive block structure of modern ConvNets, such as ResNets, parameter-sharing among repetitive convolution layers has been proposed to reduce the size of parameters. However, naive sharing of convolution filters poses many challenges such as overfitting and vanishing/exploding gradients, resulting in worse performance than non-shared counterpart models. Furthermore, sharing parameters often increases computational complexity due to additional operations for re-parameterization. In this work, we propose an efficient parameter-sharing structure and an effective training mechanism of deeply shared parameters. In the proposed ConvNet architecture, convolution layers are decomposed into a filter basis, that can be shared recursively, and layer-specific parts. We conjecture that a shared filter basis combined with a small amount of layer-specific parameters can retain, or further enhance, the representation power of individual layers, if a proper training method is applied. We show both theoretically and empirically that potential vanishing/exploding gradients problems can be mitigated by enforcing orthogonality to the shared filter bases. Experimental results demonstrate that our scheme effectively reduces redundancy by saving up to 63.8% of parameters while consistently outperforming non-shared counterpart networks even when a filter basis is deeply shared by up to 10 repetitive convolution layers.

****

Towards Practical Second Order Optimization for Deep Learning
Rohan Anil,Vineet Gupta,Tomer Koren,Kevin Regan,Yoram Singer

Optimization in machine learning, both theoretical and applied, is presently dominated by first-order gradient methods such as stochastic gradient descent. Second-order optimization methods, that involve second derivatives and/or second order statistics of the data, are far less prevalent despite strong theoretical properties, due to their prohibitive computation, memory and communication costs. In an attempt to bridge this gap between theoretical and practical optimization, we present a scalable implementation of a second-order preconditioned method (concretely, a variant of full-matrix Adagrad), that along with several critical algorithmic and numerical improvements, provides significant convergence and wall-clock time improvements compared to conventional first-order methods on state-of-the-art deep models. Our novel design effectively utilizes the prevalent heterogeneous hardware architecture for training deep models, consisting of a multicore CPU coupled with multiple accelerator units. We demonstrate superior performance compared to state-of-the-art on very large learning tasks such as machine translation with Transformers, language modeling with BERT, click-through rate prediction on Criteo, and image classification on ImageNet with ResNet-50.
****

Autoregressive Entity Retrieval
Nicola De Cao,Gautier Izacard,Sebastian Riedel,Fabio Petroni
Entities are at the center of how we represent and aggregate knowledge. For instance, Encyclopedias such as Wikipedia are structured by entities (e.g., one per Wikipedia article). The ability to retrieve such entities given a query is fundamental for knowledge-intensive tasks such as entity linking and open-domain question answering. One way to understand current approaches is as classifiers among atomic labels, one for each entity. Their weight vectors are dense entity representations produced by encoding entity meta information such as their descriptions. This approach leads to several shortcomings: (i) context and entity affinity is mainly captured through a vector dot product, potentially missing fine-grained interactions between the two; (ii) a large memory footprint is needed to store dense representations when considering large entity sets; (iii) an appropriately hard set of negative data has to be subsampled at training time. In this work, we propose GENRE, the first system that retrieves entities by generating their unique names, left to right, token-by-token in an autoregressive fashion and conditioned on the context. This enables us to mitigate the aforementioned technical issues since: (i) the autoregressive formulation allows us to directly capture relations between context and entity name, effectively cross encoding both; (ii) the memory footprint is greatly reduced because the parameters of our encoder-decoder architecture scale with vocabulary size, not entity count; (iii) the exact softmax loss can be efficiently computed without the need to subsample negative data. We show the efficacy of the approach, experimenting with more than 20 datasets on entity disambiguation, end-to-end entity linking and document retrieval tasks, achieving new state-of-the-art or very competitive results while using a tiny fraction of the memory footprint of competing systems. Finally, we demonstrate that new entities can be added by simply specifying their unambiguous name. Code and pre-trained models at https://github.com/facebookresearch/GENRE.
****

Shapley Explanation Networks
Rui Wang,Xiaoqian Wang,David I. Inouye
Shapley values have become one of the most popular feature attribution explanation methods. However, most prior work has focused on post-hoc Shapley explanations, which can be computationally demanding due to its exponential time complexity and preclude model regularization based on Shapley explanations during training. Thus, we propose to incorporate Shapley values themselves as latent representations in deep models■ thereby making Shapley explanations first-class citizens in the modeling paradigm. This intrinsic explanation approach enables layer-wise explanations, explanation regularization of the model during training, and fast explanation computation at test time. We define the Shapley transform that transforms the input into a Shapley representation given a specific function. We operationalize the Shapley transform as a neural network module and construct both shallow and deep networks, called ShapNets, by composing Shapley modules. We prov

e that our Shallow ShapNets compute the exact Shapley values and our Deep ShapNe
ts maintain the missingness and accuracy properties of Shapley values. We demons
trate on synthetic and real-world datasets that our ShapNets enable layer-wise S
hapley explanations, novel Shapley regularizations during training, and fast com
putation while maintaining reasonable performance. Code is available at https://
github.com/inouye-lab/ShapleyExplanationNetworks.
****

Can We Use Gradient Norm as a Measure of Generalization Error for Model Selectio
n in Practice?
Haozhe An,Haoyi Xiong,Xuhong Li,Xingjian Li,Dejing Dou,Zhanxing Zhu
The recent theoretical investigation (Li et al., 2020) on the upper bound of gen
eralization error of deep neural networks (DNNs) demonstrates the potential of u
sing the gradient norm as a measure that complements validation accuracy for mod
el selection in practice. In this work, we carry out empirical studies using se
veral commonly-used neural network architectures and benchmark datasets to under
stand the effectiveness and efficiency of using gradient norm as the model selec
tion criterion, especially in the settings of hyper-parameter optimization. Whil
e strong correlations between the generalization error and the gradient norm mea
sures have been observed, we find the computation of gradient norm is time consu
ming due to the high gradient complexity. To balance the trade-off between effic
iency and effectiveness, we propose to use an accelerated approximation (Goodfel
low, 2015) of gradient norm that only computes the loss gradient in the Fully-C
onnected Layer (FC Layer) of DNNs with significantly reduced computation cost (2
00~20,000 times faster). Our empirical studies clearly find that the use of appr
oximated gradient norm, as one of the hyper-parameter search objectives, can sel
ect the models with lower generalization error, but the efficiency is still low
(marginal accuracy improvement but with high computation overhead). Our results
also show that the bandit-based or population-based algorithms, such as BOHB, pe
rform poorer with gradient norm objectives, since the correlation between gradie
nt norm and generalization error is not always consistent across phases of the t
raining process. Finally, gradient norm also fails to predict the generalization
 performance of models based on different architectures, in comparison with stat
e of the art algorithms and metrics.
****

Empirical Studies on the Convergence of Feature Spaces in Deep Learning
Haoran Liu,Haoyi Xiong,Yaqing Wang,Haozhe An,Dongrui Wu,Dejing Dou
While deep learning is effective to learn features/representations from data, th
e distributions of samples in feature spaces learned by various architectures fo
r different training tasks (e.g., latent layers of AEs and feature vectors in CN
N classifiers) have not been well-studied or compared. We hypothesize that the f
eature spaces of networks trained by various architectures (AEs or CNNs) and tas
ks (supervised, unsupervised, or self-supervised learning) share some common sub
spaces, no matter what types of DNN architectures or whether the labels have bee
n used in feature learning. To test our hypothesis, through Singular Value Decom
position (SVD) of feature vectors, we demonstrate that one could linearly projec
t the feature vectors of the same group of samples to a similar distribution, wh
ere the distribution is represented as the top left singular vector (i.e., princ
ipal subspace of feature vectors), namely $\mathcal{P}$-vectors. We further asse
ss the convergence of feature space learning using angles between $\mathcal{P}$-
vectors obtained from the well-trained model and its checkpoint per epoch during
 the learning procedure, where a quasi-monotonic trend of convergence to small a
ngles has been observed. Finally, we carry out case studies to connect $\mathcal
{P}$-vectors to the data distribution, and generalization performance. Extensive
 experiments with practically-used MLP, AE and CNN architectures for classificat
ion, image reconstruction, and self-supervised learning tasks on MNIST, CIFAR-10
 and CIFAR-100 datasets have been done to support our claims with solid evidence
s.
****

Iterated graph neural network system
Hanju Li

We present Iterated Graph Neural Network System (IGNNS), a new framework of Grap
h Neural Networks (GNNs), which can deal with undirected graph and directed grap
h in a unified way. The core component of IGNNS is the Iterated Function System
(IFS), which is an important research field in fractal geometry. The key idea of
 IGNNS is to use a pair of affine transformations to characterize the process of
 message passing between graph nodes and assign an adjoint probability vector to
 them to form an IFS layer with probability. After embedding in the latent space
, the node features are sent to IFS layer for iterating, and then obtain the hig
h-level representation of graph nodes. We also analyze the geometric properties
of IGNNS from the perspective of dynamical system. We prove that if the IFS indu
ced by IGNNS is contractive, then the fractal representation of graph nodes conv
erges to the fractal set of IFS in Hausdorff distance and the ergodic representa
tion of that converges to a constant matrix in Frobenius norm. We have carried o
ut a series of semi supervised node classification experiments on citation netwo
rk datasets such as citeser, Cora and PubMed. The experimental results show that
 the performance of our method is obviously better than the related methods.
****

Factoring out Prior Knowledge from Low-Dimensional Embeddings
Edith Heiter,Jonas Fischer,Jilles Vreeken
Low-dimensional embedding techniques such as tSNE and UMAP allow visualizing hig
h-dimensional data and therewith facilitate the discovery of interesting structu
re. Although they are widely used, they visualize data as is, rather than in lig
ht of the background knowledge we have about the data. What we already know, how
ever, strongly determines what is novel and hence interesting. In this paper we
propose two methods for factoring out prior knowledge in the form of distance ma
trices from low-dimensional embeddings. To factor out prior knowledge from tSNE
embeddings, we propose JEDI that adapts the tSNE objective in a principled way u
sing Jensen-Shannon divergence. To factor out prior knowledge from any downstrea
m embedding approach, we propose CONFETTI , in which we directly operate on the
input distance matrices. Extensive experiments on both synthetic and real world
data show that both methods work well, providing embeddings that exhibit meaning
ful structure that would otherwise remain hidden.
****

Neural Approximate Sufficient Statistics for Implicit Models
Yanzhi Chen,Dinghuai Zhang,Michael U. Gutmann,Aaron Courville,Zhanxing Zhu
We consider the fundamental problem of how to automatically construct summary st
atistics for implicit generative models where the evaluation of the likelihood f
unction is intractable but sampling data from the model is possible. The idea is
 to frame the task of constructing sufficient statistics as learning mutual info
rmation maximizing representations of the data with the help of deep neural netw
orks. The infomax learning procedure does not need to estimate any density or de
nsity ratio. We apply our approach to both traditional approximate Bayesian comp
utation and recent neural likelihood methods, boosting their performance on a ra
nge of tasks.
****

Introducing Sample Robustness
Monty Maximilian Zühlke
Choosing the right data and model for a pre-defined task is one of the critical
competencies in machine learning. Investigating what features of a dataset and i
ts underlying distribution a model decodes may enlighten the mysterious "black b
ox" and guide us to a deeper and more profound understanding of the ongoing proc
esses. Furthermore, it will help to improve the quality of models which directly
 depend on data or learn from it through training. In this work, we introduce th
e dataset-dependent concept of sample robustness, which is based on a point-wise
 Lipschitz constant of the label map. For a particular sample, it measures how s
mall of a perturbation is required to cause a label-change relative to the magni
tude of the label map. We introduce theory to motivate the concept and to analys
e the effects of having similar robustness distributions for the training- and t
est data. Afterwards, we conduct various experiments using different datasets an
d (non-)deterministic models. In some cases, we can boost performance by choosin

g specifically tailored training(sub)sets and hyperparameters depending on the r
obustness distribution of the test(sub)sets.
****

Einstein VI: General and Integrated Stein Variational Inference in NumPyro
Ahmad Salim Al-Sibahi,Ola Rønning,Christophe Ley,Thomas Wim Hamelryck
Stein Variational Inference is a technique for approximate Bayesian inferencetha
t is recently gaining popularity since it combines the scalability of traditiona
lVariational Inference (VI) with the flexibility of non-parametric particle base
dinference methods. While there has been considerable progress in developmentof
 algorithms, integration in existing probabilistic programming languages (PPLs)w
ith an easy-to-use interface is currently lacking. EinStein VI is a lightweight
composable library that integrates the latest Stein Variational Inference method
swith the NumPyro PPL. Inference with EinStein VI relies on ELBO-within-Stein to
support use of custom inference programs (guides), non-linear scaling of repulsi
onforce, second-order gradient updates using matrix-valued kernels and parameter
transforms. We demonstrate the achieved synergy of the different Stein technique
sand the versatility of EinStein VI library by applying it on examples. Compared
to traditional Stochastic VI, EinStein VI is better at capturing uncertainty and
representing richer posteriors. We use several applications to show how one canu
se Neural Transforms (NeuTra) and second-order optimization to provide betterinf
erence using EinStein VI. We show how EinStein VI can be used to infer theparame
ters of a Latent Dirichlet Allocation model with a neural guide. The resultsindi
cate that Einstein VI can be combined with NumPyro’s support for automaticmargin
alization to do inference over models with discrete latent variables. Finally,we
 introduce an example with a novel extension to Deep Markov Models, calledthe St
ein Mixture Deep Markov Model (SM-DMM), which shows that EinStein VIcan be scale
d to reasonably large models with over 500.000 parameters
****

A Chain Graph Interpretation of Real-World Neural Networks
Yuesong Shen,Daniel Cremers
The last decade has witnessed a boom of deep learning research and applications
achieving state-of-the-art results in various domains. However, most advances ha
ve been established empirically, and their theoretical analysis remains lacking.
 One major issue is that our current interpretation of neural networks (NNs) as
function approximators is too generic to support in-depth analysis. In this pape
r, we remedy this by proposing an alternative interpretation that identifies NNs
 as chain graphs (CGs) and feed-forward as an approximate inference procedure. T
he CG interpretation specifies the nature of each NN component within the rich t
heoretical framework of probabilistic graphical models, while at the same time r
emains general enough to cover real-world NNs with arbitrary depth, multi-branch
ing and varied activations, as well as common structures including convolution /
 recurrent layers, residual block and dropout. We demonstrate with concrete exam
ples that the CG interpretation can provide novel theoretical support and insigh
ts for various NN techniques, as well as derive new deep learning approaches suc
h as the concept of partially collapsed feed-forward inference. It is thus a pro
mising framework that deepens our understanding of neural networks and provides
a coherent theoretical formulation for future deep learning research.
****

Decentralized Knowledge Graph Representation Learning
Lingbing Guo,Weiqing Wang,Zequn Sun,Chenghao Liu,Wei Hu
Knowledge graph (KG) representation learning methods have achieved competitive p
erformance in many KG-oriented tasks, among which the best ones are usually base
d on graph neural networks (GNNs), a powerful family of networks that learns the
 representation of an entity by aggregating the features of its neighbors and it
self. However, many KG representation learning scenarios only provide the struct
ure information that describes the relationships among entities, causing that en
tities have no input features. In this case, existing aggregation mechanisms are
 incapable of inducing embeddings of unseen entities as these entities have no p
re-defined features for aggregation. In this paper, we present a decentralized K
G representation learning approach, decentRL, which encodes each entity from and

only from the embeddings of its neighbors. For optimization, we design an algorithm to distill knowledge from the model itself such that the output embeddings can continuously gain knowledge from the corresponding original embeddings. Extensive experiments show that the proposed approach performed better than many cutting-edge models on the entity alignment task, and achieved competitive performance on the entity prediction task. Furthermore, under the inductive setting, it significantly outperformed all baselines on both tasks.
****

Integrating linguistic knowledge into DNNs: Application to online grooming detection

Jay Morgan,Adeline Paiement,Nuria Lorenzo-Dus,Anina Kinzel,Matteo Di Cristofaro

Online grooming (OG) of children is a pervasive issue in an increasingly interconnected world. We explore various complementary methods to incorporate Corpus Linguistics (CL) knowledge into accurate and interpretable Deep Learning (DL) models. They provide an implicit text normalisation that adapts embedding spaces to the groomers' usage of language, and they focus the DNN's attention onto the expressions of OG strategies. We apply these integration to two architecture types and improve on the state-of-the-art on a new OG corpus.
****

Increasing the Coverage and Balance of Robustness Benchmarks by Using Non-Overlapping Corruptions

Alfred LAUGROS,Alice Caplier,Matthieu Ospici

Neural Networks are sensitive to various corruptions that usually occur in real-world applications such as low-lighting conditions, blurs, noises, etc. To estimate the robustness of neural networks to these common corruptions, we generally use a group of modeled corruptions gathered into a benchmark. We argue that corruption benchmarks often have a poor coverage: being robust to them only implies being robust to a narrow range of corruptions. They are also often unbalanced: they give too much importance to some corruptions compared to others. In this paper, we propose to build corruption benchmarks with only non-overlapping corruptions, to improve their coverage and their balance. Two corruptions overlap when the robustnesses of neural networks to these corruptions are correlated. We propose the first metric to measure the overlapping between two corruptions. We provide an algorithm that uses this metric to build benchmarks of Non-Overlapping Corruptions. Using this algorithm, we build from ImageNet a new corruption benchmark called ImageNet-NOC. We show that ImageNet-NOC is balanced and covers several kinds of corruptions that are not covered by ImageNet-C.
****

Implicit Regularization Effects of Unbiased Random Label Noises with SGD

Haoyi Xiong,Xuhong Li,Boyang Yu,Dejing Dou,Dongrui Wu,Zhanxing Zhu

Random label noises (or observational noises) widely exist in practical machinelearning settings. we analyze the learning dynamics of stochastic gradient descent(SGD) over the quadratic loss with unbiased label noises, and investigate a new noise term of dynamics, which is dynamized and influenced by mini-batch sam-pling and random label noises, as an implicit regularizer. Our theoretical analysisfinds such implicit regularizer would favor some convergence points that could stabilize model outputs against perturbation of parameters. To validate our analy-sis, we use our theorems to estimate the closed-form solution of the implicit reg-ularizer over continuous-time SGD dynamics for Ordinary Least-Square (OLS), where the numerical simulation backups our estimates. We further extend our proposals to interpret the newly-fashioned noisy self-distillation tricks for deep learning, where the implicit regularizer demonstrates a unique capacity of selecting models with improved output stability through learning from well-trained teach-ers with additive unbiased random label noises
****

Improving Random-Sampling Neural Architecture Search by Evolving the Proxy Search Space

Yuhong Li,Cong Hao,Xiaofan Zhang,Jinjun Xiong,Wen-mei Hwu,Deming Chen

Random-sampling Neural Architecture Search (RandomNAS) has recently become a prevailing NAS approach because of its search efficiency and simplicity. There are

two main steps in RandomNAS: the training step that randomly samples the weight-sharing architectures from a supernet and iteratively updates their weights, and the search step that ranks architectures by their respective validation performance. Key to both steps is the assumption of a high correlation between estimated performance(i.e., accuracy) for weight-sharing architectures and their respective achievable accuracy (i.e., ground truth) when trained from scratch. We examine such a phenomenon via NASBench-201, whose ground truth is known for its entire NAS search space. We observe that existing RandomNAS can rank a set of architectures uniformly sampled from the entire global search space(GS), that correlates well with its ground-truth ranking. However, if we only focus on the top-performing architectures (such as top 20\% according to the ground truth) in the GS, such a correlation drops dramatically. This raises the question of whether we can find an effective proxy search space (PS) that is only a small subset of GS to dramatically improve RandomNAS's search efficiency while at the same time keeping a good correlation for the top-performing architectures. This paper proposes a new RandomNAS-based approach called EPS (Evolving the Proxy Search Space) to address this problem. We show that, when applied to NASBench-201, EPS can achieve near-optimal NAS performance and beat all existing state-of-the-art. When applied to different-variants of DARTS-like search spaces for tasks such as image classification and natural language processing, EPS is able to robustly achieve superior performance with shorter or similar search time compared to some leading NAS works. The code is available at https://github.com/IcLr2020SuBmIsSiOn/EPS
****
Is deeper better? It depends on locality of relevant features
Takashi Mori,Masahito Ueda
It has been recognized that a heavily overparameterized artificial neural network exhibits surprisingly good generalization performance in various machine-learning tasks. Recent theoretical studies have made attempts to unveil the mystery of the overparameterization. In most of those previous works, the overparameterization is achieved by increasing the width of the network, while the effect of increasing the depth has been less well understood. In this work, we investigate the effect of increasing the depth within an overparameterized regime. To gain an insight into the advantage of depth, we introduce local and global labels as abstract but simple classification rules. It turns out that the locality of the relevant feature for a given classification rule plays an important role; our experimental results suggest that deeper is better for local labels, whereas shallower is better for global labels. We also compare the results of finite networks with those of the neural tangent kernel (NTK), which is equivalent to an infinitely wide network with a proper initialization and an infinitesimal learning rate. It is shown that the NTK does not correctly capture the depth dependence of the generalization performance, which indicates the importance of the feature learning, rather than the lazy learning.
****
Automated Concatenation of Embeddings for Structured Prediction
Xinyu Wang,Yong Jiang,Nguyen Bach,Tao Wang,Zhongqiang Huang,Fei Huang,Kewei Tu
Pretrained contextualized embeddings are powerful word representations for structured prediction tasks. Recent work found that better word representations can be obtained by concatenating different types of embeddings. However, the selection of embeddings to form the best concatenated representation usually varies depending on the task and the collection of candidate embeddings, and the ever-increasing number of embedding types makes it a more difficult problem. In this paper, we propose Automated Concatenation of Embeddings (ACE) to automate the process of finding better concatenations of embeddings for structured prediction tasks, based on a formulation inspired by recent progress on neural architecture search. Specifically, a controller alternately samples a concatenation of embeddings, according to its current belief of the effectiveness of individual embedding types in consideration for a task, and updates the belief based on a reward. We follow strategies in reinforcement learning to optimize the parameters of the controller and compute the reward based on the accuracy of a task model, which is fed with the sampled concatenation as input and trained on a task dataset. Empiric

al results on 6 tasks and 21 datasets show that our approach outperforms strong baselines and achieves state-of-the-art performance with fine-tuned embeddings in the vast majority of evaluations.
****

Quantifying Statistical Significance of Neural Network Representation-Driven Hypotheses by Selective Inference
Vo Nguyen Le Duy,Shogo Iwazaki,Ichiro Takeuchi
In the past few years, various approaches have been developed to explain and interpret deep neural network (DNN) representations, but it has been pointed out that these representations are sometimes unstable and not reproducible. In this paper, we interpret these representations as hypotheses driven by DNN (called DNN-driven hypotheses) and propose a method to quantify the reliability of these hypotheses in statistical hypothesis testing framework. To this end, we introduce Selective Inference (SI) framework, which has received much attention in the past few years as a new statistical inference framework for data-driven hypotheses. The basic idea of SI is to make conditional inferences on the selected hypotheses under the condition that they are selected. In order to use SI framework for DNN representations, we develop a new SI algorithm based on homotopy method which enables us to derive the exact (non-asymptotic) conditional sampling distribution of the DNN-driven hypotheses. We conduct experiments on both synthetic and real-world datasets, through which we offer evidence that our proposed method can successfully control the false positive rate, has decent performance in terms of computational efficiency, and provides good results in practical applications.
****

Multi-EPL: Accurate Multi-source Domain Adaptation
Seongmin Lee,Hyunsik Jeon,U Kang
Given multiple source datasets with labels, how can we train a target model with no labeled data? Multi-source domain adaptation (MSDA) aims to train a model using multiple source datasets different from a target dataset in the absence of target data labels. MSDA is a crucial problem applicable to many practical cases where labels for the target data are unavailable due to privacy issues. Existing MSDA frameworks are limited since they align data without considering conditional distributions $p(x|y)$ of each domain. They also do not fully utilize the target data without labels, and rely on limited feature extraction with a single extractor. In this paper, we propose Multi-EPL, a novel method for multi-source domain adaptation. Multi-EPL exploits label-wise moment matching to align conditional distributions $p(x|y)$, uses pseudolabels for the unavailable target labels, and introduces an ensemble of multiple feature extractors for accurate domain adaptation. Extensive experiments show that Multi-EPL provides the state-of-the-art performance for multi-source domain adaptation tasks in both of image domains and text domains.
****

Federated Mixture of Experts
Matthias Reisser,Christos Louizos,Efstratios Gavves,Max Welling
Federated learning (FL) has emerged as the predominant approach for collaborative training of neural network models across multiple users, without the need to gather the data at a central location. One of the important challenges in this setting is data heterogeneity; different users have different data characteristics. For this reason, training and using a single global model might be suboptimal when considering the performance of each of the individual user’s data. In this work, we tackle this problem via Federated Mixture of Experts, FedMix, a framework that allows us to train an ensemble of specialized models. FedMix adaptively selects and trains a user-specific selection of the ensemble members. We show that users with similar data characteristics select the same members and therefore share statistical strength while mitigating the effect of non-i.i.d data. Empirically, we show through an extensive experimental evaluation that FedMix improves performance compared to using a single global model while requiring similar or less communication costs.
****

AUBER: Automated BERT Regularization

Hyun Dong Lee,Seongmin Lee,U Kang
How can we effectively regularize BERT? Although BERT proves its effectiveness i n various downstream natural language processing tasks, it often over■ts when th ere are only a small number of training instances. A promising direction to regu larize BERT is based on pruning its attention heads based on a proxy score for h ead importance. However, heuristic-based methods are usually suboptimal since th ey predetermine the order by which attention heads are pruned. In order to overc ome such a limitation, we propose AUBER, an effective regularization method that leverages reinforcement learning to automatically prune attention heads from BE RT. Instead of depending on heuristics or rule-based policies, AUBER learns a pr uning policy that determines which attention heads should or should not be prune d for regularization. Experimental results show that AUBER outperforms existing pruning methods by achieving up to 10% better accuracy. In addition, our ablatio n study empirically demonstrates the effectiveness of our design choices for AUB ER.
****

NBDT: Neural-Backed Decision Tree
Alvin Wan,Lisa Dunlap,Daniel Ho,Jihan Yin,Scott Lee,Suzanne Petryk,Sarah Adel Ba rgal,Joseph E. Gonzalez
Machine learning applications such as finance and medicine demand accurate and j ustifiable predictions, barring most deep learning methods from use. In response , previous work combines decision trees with deep learning, yielding models that (1) sacrifice interpretability for accuracy or (2) sacrifice accuracy for inter pretability. We forgo this dilemma by jointly improving accuracy and interpretab ility using Neural-Backed Decision Trees (NBDTs). NBDTs replace a neural network 's final linear layer with a differentiable sequence of decisions and a surrogat e loss. This forces the model to learn high-level concepts and lessens reliance on highly-uncertain decisions, yielding (1) accuracy: NBDTs match or outperform modern neural networks on CIFAR, ImageNet and better generalize to unseen classe s by up to 16%. Furthermore, our surrogate loss improves the original model's ac curacy by up to 2%. NBDTs also afford (2) interpretability: improving human trus tby clearly identifying model mistakes and assisting in dataset debugging. Code and pretrained NBDTs are at https://github.com/alvinwan/neural-backed-decision-t rees.
****

DiffWave: A Versatile Diffusion Model for Audio Synthesis
Zhifeng Kong,Wei Ping,Jiaji Huang,Kexin Zhao,Bryan Catanzaro
In this work, we propose DiffWave, a versatile diffusion probabilistic model for conditional and unconditional waveform generation. The model is non-autoregress ive, and converts the white noise signal into structured waveform through a Mark ov chain with a constant number of steps at synthesis. It is efficiently trained by optimizing a variant of variational bound on the data likelihood. DiffWave p roduces high-fidelity audios in different waveform generation tasks, including n eural vocoding conditioned on mel spectrogram, class-conditional generation, and unconditional generation. We demonstrate that DiffWave matches a strong WaveNet vocoder in terms of speech quality (MOS: 4.44 versus 4.43), while synthesizing orders of magnitude faster. In particular, it significantly outperforms autoregr essive and GAN-based waveform models in the challenging unconditional generation task in terms of audio quality and sample diversity from various automatic and human evaluations.
****

Action Concept Grounding Network for Semantically-Consistent Video Generation
Wei Yu,Wenxin Chen,Animesh Garg
Recent works in self-supervised video prediction have mainly focused on passive forecasting and low-level action-conditional prediction, which sidesteps the pro blem of semantic learning. We introduce the task of semantic action-conditional video prediction, which can be regarded as an inverse problem of action recognit ion. The challenge of this new task primarily lies in how to effectively inform the model of semantic action information. To bridge vision and language, we util ize the idea of capsule and propose a novel video prediction model Action Concep

t Grounding Network (AGCN). Our method is evaluated on two newly designed synthe
tic datasets, CLEVR-Building-Blocks and Sapien-Kitchen, and experiments show tha
t given different action labels, our ACGN can correctly condition on instruction
s and generate corresponding future frames without need of bounding boxes. We fu
rther demonstrate our trained model can make out-of-distribution predictions for
 concurrent actions, be quickly adapted to new object categories and exploit its
 learnt features for object detection. Additional visualizations can be found at
 https://iclr-acgn.github.io/ACGN/.
****

RSO: A Gradient Free Sampling Based Approach For Training Deep Neural Networks
Rohun Tripathi,Bharat Singh
We propose RSO (random search optimization), a gradient free, sampling based app
roach for training deep neural networks. To this end, RSO adds a perturbation to
 a weight in a deep neural network and tests if it reduces the loss on a mini-ba
tch. If this reduces the loss, the weight is updated, otherwise the existing wei
ght is retained. Surprisingly, we find that repeating this process a few times f
or each weight is sufficient to train a deep neural network. The number of weigh
t updates for RSO is an order of magnitude lesser when compared to backpropagati
on with SGD. RSO can make aggressive weight updates in each step as there is no
concept of learning rate. The weight update step for individual layers is also n
ot coupled with the magnitude of the loss. RSO is evaluated on classification ta
sks on MNIST and CIFAR-10 datasets with deep neural networks of 6 to 10 layers w
here it achieves an accuracy of 99.1% and 81.8% respectively. We also find that
after updating the weights just 5 times, the algorithm obtains a classification
accuracy of 98% on MNIST.
****

What Preserves the Emergence of Language?
Ziluo Ding,Tiejun Huang,Zongqing Lu
The emergence of language is a mystery. One dominant theory is that cooperation
boosts language to emerge. However, as a means of giving out information, langua
ge seems not to be an evolutionarily stable strategy. To ensure the survival adv
antage of many competitors, animals are selfish in nature. From the perspective
of Darwinian, if an individual can obtain a higher benefit by deceiving the othe
r party, why not deceive? For those who are cheated, once bitten and twice shy,
cooperation will no longer be a good option. As a result, motivation for communi
cation, as well as the emergence of language would perish. Then, what preserves
the emergence of language? We aim to answer this question in a brand new framewo
rk of agent community, reinforcement learning, and natural selection. Empiricall
y, we reveal that lying indeed dispels cooperation. Even with individual resista
nce to lying behaviors, liars can easily defeat truth tellers and survive during
 natural selection. However, social resistance eventually constrains lying and m
akes the emergence of language possible.
****

Denoising Diffusion Implicit Models
Jiaming Song,Chenlin Meng,Stefano Ermon
Denoising diffusion probabilistic models (DDPMs) have achieved high quality imag
e generation without adversarial training, yet they require simulating a Markov
chain for many steps in order to produce a sample. To accelerate sampling, we pr
esent denoising diffusion implicit models (DDIMs), a more efficient class of ite
rative implicit probabilistic models with the same training procedure as DDPMs.
In DDPMs, the generative process is defined as the reverse of a particular Marko
vian diffusion process. We generalize DDPMs via a class of non-Markovian diffusi
on processes that lead to the same training objective. These non-Markovian proce
sses can correspond to generative processes that are deterministic, giving rise
to implicit models that produce high quality samples much faster. We empirically
 demonstrate that DDIMs can produce high quality samples $10 \times$ to $50 \tim
es$ faster in terms of wall-clock time compared to DDPMs, allow us to trade off
computation for sample quality, perform semantically meaningful image interpolat
ion directly in the latent space, and reconstruct observations with very low err
or.

****
A Unified Framework for Convolution-based Graph Neural Networks
Xuran Pan,Shiji Song,Gao Huang
Graph Convolutional Networks (GCNs) have attracted a lot of research interest in the machine learning community in recent years. Although many variants have been proposed, we still lack a systematic view of different GCN models and deep understanding of the relations among them. In this paper, we take a step forward to establish a unified framework for convolution-based graph neural networks, by formulating the basic graph convolution operation as an optimization problem in the graph Fourier space. Under this framework, a variety of popular GCN models, including the vanilla-GCNs, attention-based GCNs and topology-based GCNs, can be interpreted as a same optimization problem but with different carefully designed regularizers. This novel perspective enables a better understanding of the similarities and differences among many widely used GCNs, and may inspire new approaches for designing better models. As a showcase, we also present a novel regularization technique under the proposed framework to tackle the oversmoothing problem in graph convolution. The effectiveness of the newly designed model is validated empirically.
****
Learning Deep Latent Variable Models via Amortized Langevin Dynamics
Shohei Taniguchi,Yusuke Iwasawa,Yutaka Matsuo
How can we perform posterior inference for deep latent variable models in an efficient and flexible manner? Markov chain Monte Carlo (MCMC) methods, such as Langevin dynamics, provide sample approximations of such posteriors with an asymptotic convergence guarantee. However, it is difficult to apply these methods to large-scale datasets owing to their slow convergence and datapoint-wise iterations. In this study, we propose amortized Langevin dynamics, wherein datapoint-wise MCMC iterations are replaced with updates of an inference model that maps observations into latent variables. The amortization enables scalable inference from large-scale datasets. Developing a latent variable model and an inference model with neural networks, yields Langevin autoencoders (LAEs), a novel Langevin-based framework for deep generative models. Moreover, if we define a latent prior distribution with an unnormalized energy function for more flexible generative modeling, LAEs are extended to a more general framework, which we refer to as contrastive Langevin autoencoders (CLAEs). We experimentally show that LAEs and CLAEs can generate sharp image samples. Moreover, we report their performance of unsupervised anomaly detection.
****
Suppressing Outlier Reconstruction in Autoencoders for Out-of-Distribution Detection
Sangwoong Yoon,Yung-Kyun Noh,Frank C. Park
While only trained to reconstruct training data, autoencoders may produce high-quality reconstructions of inputs that are well outside the training data distribution. This phenomenon, which we refer to as outlier reconstruction, has a detrimental effect on the use of autoencoders for outlier detection, as an autoencoder will misclassify a clear outlier as being in-distribution. In this paper, we introduce the Energy-Based Autoencoder (EBAE), an autoencoder that is considerably less susceptible to outlier reconstruction.
The core idea of EBAE is to treat the reconstruction error as an energy function of a normalized density and to strictly enforce the normalization constraint. We show that the reconstruction of non-training inputs can be suppressed, and the reconstruction error made highly discriminative to outliers, by enforcing this constraint. We empirically show that EBAE significantly outperforms both existing autoencoders and other generative models for several out-of-distribution detection tasks.
****
Fast Predictive Uncertainty for Classification with Bayesian Deep Networks
Marius Hobbhahn,Agustinus Kristiadi,Philipp Hennig
In Bayesian Deep Learning, distributions over the output of classification neural networks are approximated by first constructing a Gaussian distribution over t

he weights, then sampling from it to receive a distribution over the categorical output distribution. This is costly. We reconsider old work to construct a Dirichlet approximation of this output distribution, which yields an analytic map between Gaussian distributions in logit space and Dirichlet distributions (the conjugate prior to the categorical) in the output space. We argue that the resulting Dirichlet distribution has theoretical and practical advantages, in particular, more efficient computation of the uncertainty estimate, scaling to large datasets and networks like ImageNet and DenseNet. We demonstrate the use of this Dirichlet approximation by using it to construct a lightweight uncertainty-aware output ranking for the ImageNet setup.
****

Discriminative Representation Loss (DRL): A More Efficient Approach than Gradient Re-Projection in Continual Learning
Yu Chen,Tom Diethe,Peter Flach
The use of episodic memories in continual learning has been shown to be effective in terms of alleviating catastrophic forgetting. In recent studies, several gradient-based approaches have been developed to make more efficient use of compact episodic memories, which constrain the gradients resulting from new samples with those from memorized samples, aiming to reduce the diversity of gradients from different tasks. In this paper, we reveal the relation between diversity of gradients and discriminativeness of representations, demonstrating connections between Deep Metric Learning and continual learning. Based on these findings, we propose a simple yet efficient method -- Discriminative Representation Loss (DRL) -- for continual learning. In comparison with several state-of-the-art methods, this method shows effectiveness with low computational cost on multiple benchmark experiments in the setting of online continual learning.
****

Deep Reinforcement Learning with Causality-based Intrinsic Reward
Peng Zhang,Furui Liu,Zhitang Chen,Jianye HAO,Jun Wang
Reinforcement Learning (RL) has shown great potential to deal with sequential decision-making problems. However, most RL algorithms do not explicitly consider the relations between entities in the environment. This makes the policy learning suffer from the problems of efficiency, effectivity and interpretability. In this paper, we propose a novel deep reinforcement learning algorithm, which firstly learns the causal structure of the environment and then leverages the learned causal information to assist policy learning. The proposed algorithm learns a graph to encode the environmental structure by calculating Average Causal Effect (ACE) between different categories of entities, and an intrinsic reward is given to encourage the agent to interact more with entities belonging to top-ranked categories, which significantly boosts policy learning. Several experiments are conducted on a number of simulation environments to demonstrate the effectiveness and better interpretability of our proposed method.
****

Multi-Task Learning by a Top-Down Control Network
Hila Levi,Shimon Ullman
As the range of tasks performed by a general vision system expands, executing multiple tasks accurately and ef■ciently in a single network has become an important and still open problem. Recent computer vision approaches address this problem by branching networks, or by a channel-wise modulation of the network feature-maps with task speci■c vectors. We present a novel architecture that uses a dedicated top-down control network to modify the activation of all the units in the main recognition network in a manner that depends on the selected task, image content, and spatial location. We show the effectiveness of our scheme by achieving signi■cantly better results than alternative state-of-the-art approaches on four our datasets. We further demonstrate our advantages in terms of task selectivity, scaling the number of tasks and interpretability.
Code is supplied in the supplementary materials and will be publicly available.
****

Bridging the Imitation Gap by Adaptive Insubordination
Luca Weihs,Unnat Jain,Jordi Salvador,Svetlana Lazebnik,Aniruddha Kembhavi,Alex S

chwing
When expert supervision is available, practitioners often use imitation learning with varying degrees of success. We show that when an expert has access to privileged information that is unavailable to the student, this information is marginalized in the student policy during imitation learning resulting in an ''imitation gap'' and, potentially, poor results. Prior work bridges this gap via a progression from imitation learning to reinforcement learning. While often successful, gradual progression fails for tasks that require frequent switches between exploration and memorization skills. To better address these tasks and alleviate the imitation gap we propose 'Adaptive Insubordination' (ADVISOR), which dynamically weights imitation and reward-based reinforcement learning losses during training, enabling switching between imitation and exploration. On a suite of challenging didactic and MiniGrid tasks, we show that ADVISOR outperforms pure imitation, pure reinforcement learning, as well as their sequential and parallel combinations.
****
Beyond the Pixels: Exploring the Effects of Bit-Level Network and File Corruptions on Video Model Robustness
Trenton Chang,Daniel Yang Fu,Yixuan Li
We investigate the robustness of video machine learning models to bit-level network and file corruptions, which can arise from network transmission failures or hardware errors, and explore defenses against such corruptions. We simulate network and file corruptions at multiple corruption levels, and find that bit-level corruptions can cause substantial performance drops on common action recognition and multi-object tracking tasks. We explore two types of defenses against bit-level corruptions: corruption-agnostic and corruption-aware defenses. We find that corruption-agnostic defenses such as adversarial training have limited effectiveness, performing up to 11.3 accuracy points worse than a no-defense baseline. In response, we propose Bit-corruption Augmented Training (BAT), a corruption-aware baseline that exploits knowledge of bit-level corruptions to enforce model invariance to such corruptions. BAT outperforms corruption-agnostic defenses, recovering up to 7.1 accuracy points over a no-defense baseline on highly-corrupted videos while maintaining competitive performance on clean/near-clean data.
****
Extract Local Inference Chains of Deep Neural Nets
Haiyan Zhao,Tianyi Zhou,Guodong Long,Jing Jiang,Chengqi Zhang
We study how to explain the main steps/chains of inference that a deep neural net (DNN) relies on to produce predictions in a local region of data space. This problem is related to network pruning and interpretable machine learning but the highlighted differences are: (1) fine-tuning of neurons/filters is forbidden: only exact copies are allowed; (2) we target an extremely high pruning rate, e.g., $\geq 95\%$; (3) the interpretation is for the whole inference process in a local region rather than for individual neurons/filters or on a single sample. In this paper, we introduce an efficient method, \name, to extract the local inference chains by optimizing a differentiable sparse scoring for the filters and layers to preserve the outputs on given data from a local region. Thereby, \name~can extract an extremely small sub-network composed of filters exactly copied from the original DNN by removing the filters/layers with small scores. We then visualize the sub-network by applying existing interpretation technique to the retained layer/filter/neurons and on any sample from the local region. Its architecture reveals how the inference process stitches and integrates the information layer by layer and filter by filter. We provide detailed and insightful case studies together with three quantitative analyses over thousands of trials to demonstrate the quality, sparsity, fidelity and accuracy of the interpretation within the assigned local regions and over unseen data. In our empirical study, \name~significantly enriches the interpretation and makes the inner mechanism of DNNs more transparent than before.
****
Policy Gradient with Expected Quadratic Utility Maximization: A New Mean-Variance Approach in Reinforcement Learning

Masahiro Kato,Kei Nakagawa
In real-world decision-making problems, risk management is critical. Among various risk management approaches, the mean-variance criterion is one of the most widely used in practice. In this paper, we suggest expected quadratic utility maximization (EQUM) as a new framework for policy gradient style reinforcement learning (RL) algorithms with mean-variance control. The quadratic utility function is a common objective of risk management in finance and economics. The proposed EQUM framework has several interpretations, such as reward-constrained variance minimization and regularization, as well as agent utility maximization. In addition, the computation of the EQUM framework is easier than that of existing mean-variance RL methods, which require double sampling. In experiments, we demonstrate the effectiveness of the proposed framework in benchmark setting of RL and financial data.
****
Exploring Routing Strategies for Multilingual Mixture-of-Experts Models
Sneha Kudugunta,Yanping Huang,Ankur Bapna,Maxim Krikun,Dmitry Lepikhin,Thang Luong,Orhan Firat
Sparsely-Gated Mixture-of-Experts (MoE) has been a successful approach for scaling multilingual translation models to billions of parameters without a proportional increase in training computation. These models, however, are prohibitively large for serving deployment and there is no easy way to extract a sub-network to decode for a particular language pair. This work proposes improved strategies to route MoE models by tasks instead of tokens, thus enabling separation of network structures at decoding time while enjoying the benefits of scale and task sharing at training time.
We compare routing strategies at multiple levels (token, sentence, task) in both, the encoder and the decoder, and conduct extensive experiments on two benchmarks: the public WMT dataset of 30 language pairs and an in-house web-scale dataset of 200 language pairs. On WMT, with a Transformer base model with 32 experts, our task-level MoE outperforms the best performing token-level MoE model by +1.0 BLEU on average over all language pairs. When scaling up to Transformer big model with 128 experts on the large-scale massively multilingual benchmark, our task-level MoE is competitive with token-level MoE while being able to reduce the decoder model size by a factor of 32.34 and increase peak throughput by 2.6 times at inference.
****
Learning What Not to Model: Gaussian Process Regression with Negative Constraints
Gaurav Shrivastava,Harsh Shrivastava,Abhinav Shrivastava
Gaussian Process (GP) regression fits a curve on a set of datapairs, with each pair consisting of an input point '$\mathbf{x}$' and its corresponding target regression value '$y(\mathbf{x})$' (a positive datapair). But, what if for an input point '$\bar{\mathbf{x}}$', we want to constrain the GP to avoid a target regression value '$\bar{y}(\bar{\mathbf{x}})$' (a negative datapair)? This requirement can often appear in real-world navigation tasks, where an agent would want to avoid obstacles, like furniture items in a room when planning a trajectory to navigate. In this work, we propose to incorporate such negative constraints in a GP regression framework. Our approach, 'GP-NC' or Gaussian Process with Negative Constraints, fits over the positive datapairs while avoiding the negative datapairs. Specifically, our key idea is to model the negative datapairs using small blobs of Gaussian distribution and maximize its KL divergence from the GP. We jointly optimize the GP-NC for both the positive and negative datapairs. We empirically demonstrate that our GP-NC framework performs better than the traditional GP learning and that our framework does not affect the scalability of Gaussian Process regression and helps the model converge faster as the size of the data increases.
****
Efficient Exploration for Model-based Reinforcement Learning with Continuous States and Actions
Ying Fan,Yifei Ming

Balancing exploration and exploitation is crucial in reinforcement learning (RL) . In this paper, we study the model-based posterior sampling algorithm in contin uous state-action spaces theoretically and empirically. First, we improve the re gret bound: with the assumption that reward and transition functions can be mode led as Gaussian Processes with linear kernels, we develop a Bayesian regret boun d of $\tilde{O}(H^{3/2}d\sqrt{T})$, where H is the episode length, d is the dimension of the state-action space, and T indicates the total time steps. Our bound can be extended to nonlinear cases as well: using linear kernels on the f eature representation ϕ, the Bayesian regret bound becomes $\tilde{O}(H^{3/ 2}d_{\phi}\sqrt{T})$, where d_ϕ is the dimension of the representation spac e. Moreover, we present MPC-PSRL, a model-based posterior sampling algorithm wit h model predictive control for action selection. To capture the uncertainty in m odels and realize posterior sampling, we use Bayesian linear regression on the p enultimate layer (the feature representation layer ϕ) of neural networks. E mpirical results show that our algorithm achieves the best sample efficiency in benchmark control tasks compared to prior model-based algorithms, and matches th e asymptotic performance of model-free algorithms.

****

Shape Matters: Understanding the Implicit Bias of the Noise Covariance

Jeff Z. HaoChen,Colin Wei,Jason D. Lee,Tengyu Ma

The noise in stochastic gradient descent (SGD) provides a crucial implicit regul arization effect for training overparameterized models. Prior theoretical work l argely focuses on spherical Gaussian noise, whereas empirical studies demonstrat e the phenomenon that parameter-dependent noise --- induced by mini-batches or l abel perturbation --- is far more effective than Gaussian noise.

This paper theoretically characterizes this phenomenon on a quadratically-parame terized model introduced by Vaskevicius et al. and Woodworth et al. We show tha t in an over-parameterized setting, SGD with label noise recovers the sparse gro und-truth with an arbitrary initialization, whereas SGD with Gaussian noise or g radient descent overfits to dense solutions with large norms. Our analysis revea ls that parameter-dependent noise introduces a bias towards local minima with sm aller noise variance, whereas spherical Gaussian noise does not.

****

Large Scale Image Completion via Co-Modulated Generative Adversarial Networks

Shengyu Zhao,Jonathan Cui,Yilun Sheng,Yue Dong,Xiao Liang,Eric I-Chao Chang,Yan Xu

Numerous task-specific variants of conditional generative adversarial networks h ave been developed for image completion. Yet, a serious limitation remains that all existing algorithms tend to fail when handling large-scale missing regions. To overcome this challenge, we propose a generic new approach that bridges the g ap between image-conditional and recent modulated unconditional generative archi tectures via co-modulation of both conditional and stochastic style representati ons. Also, due to the lack of good quantitative metrics for image completion, we propose the new Paired/Unpaired Inception Discriminative Score (P-IDS/U-IDS), w hich robustly measures the perceptual fidelity of inpainted images compared to r eal images via linear separability in a feature space. Experiments demonstrate s uperior performance in terms of both quality and diversity over state-of-the-art methods in free-form image completion and easy generalization to image-to-image translation. Code is available at https://github.com/zsyzzsoft/co-mod-gan.

****

GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding

Dmitry Lepikhin,HyoukJoong Lee,Yuanzhong Xu,Dehao Chen,Orhan Firat,Yanping Huang ,Maxim Krikun,Noam Shazeer,Zhifeng Chen

Neural network scaling has been critical for improving the model quality in many real-world machine learning applications with vast amounts of training data and compute. Although this trend of scaling is affirmed to be a sure-fire approach for better model quality, there are challenges on the path such as the computati on cost,ease of programming, and efficient implementation on parallel devices.

In this paper we demonstrate conditional computation as a remedy to the above me ntioned impediments, and demonstrate its efficacy and utility. We make extensiv

e use of GShard, a module composed of a set of lightweight annotation APIs and a n extension to the XLA compiler to enable large scale models with up to trillion s of parameters. GShard and conditional computation enable us to scale up multil ingual neural machine translation Transformer model with Sparsely-Gated Mixture- of-Experts. We demonstrate that such a giant model with 600 billion parameters c an efficiently be trained on 2048 TPU v3 cores in 4 days to achieve far superior quality for translation from 100 languages to English compared to the prior art .
****
Non-Negative Bregman Divergence Minimization for Deep Direct Density Ratio Estim ation
Masahiro Kato,Takeshi Teshima
The estimation of the ratio of two probability densities has garnered attention as the density ratio is useful in various machine learning tasks, such as anomal y detection and domain adaptation. To estimate the density ratio, methods collec tively known as direct density ratio estimation (DRE) have been explored. These methods are based on the minimization of the Bregman (BR) divergence between a d ensity ratio model and the true density ratio. However, existing direct DRE suff ers from serious overfitting when using flexible models such as neural networks. In this paper, we introduce a non-negative correction for empirical risk using only the prior knowledge of the upper bound of the density ratio. This correctio n makes a DRE method more robust against overfitting and enables the use of flex ible models. In the theoretical analysis, we discuss the consistency of the empi rical risk. In our experiments, the proposed estimators show favorable performan ce in inlier-based outlier detection and covariate shift adaptation.
****
What Should Not Be Contrastive in Contrastive Learning
Tete Xiao,Xiaolong Wang,Alexei A Efros,Trevor Darrell
Recent self-supervised contrastive methods have been able to produce impressive transferable visual representations by learning to be invariant to different dat a augmentations. However, these methods implicitly assume a particular set of re presentational invariances (e.g., invariance to color), and can perform poorly w hen a downstream task violates this assumption (e.g., distinguishing red vs. yel low cars). We introduce a contrastive learning framework which does not require prior knowledge of specific, task-dependent invariances. Our model learns to cap ture varying and invariant factors for visual representations by constructing se parate embedding spaces, each of which is invariant to all but one augmentation. We use a multi-head network with a shared backbone which captures information a cross each augmentation and alone outperforms all baselines on downstream tasks. We further find that the concatenation of the invariant and varying spaces perf orms best across all tasks we investigate, including coarse-grained, fine-graine d, and few-shot downstream classification tasks, and various data corruptions.
****
Learning Cross-Domain Correspondence for Control with Dynamics Cycle-Consistency
Qiang Zhang,Tete Xiao,Alexei A Efros,Lerrel Pinto,Xiaolong Wang
At the heart of many robotics problems is the challenge of learning corresponden ces across domains. For instance, imitation learning requires obtaining correspo ndence between humans and robots; sim-to-real requires correspondence between ph ysics simulators and real hardware; transfer learning requires correspondences b etween different robot environments. In this paper, we propose to learn correspo ndence across such domains emphasizing on differing modalities (vision and inter nal state), physics parameters (mass and friction), and morphologies (number of limbs). Importantly, correspondences are learned using unpaired and randomly col lected data from the two domains. We propose dynamics cycles that align dynamic robotic behavior across two domains using a cycle consistency constraint. Once t his correspondence is found, we can directly transfer the policy trained on one domain to the other, without needing any additional fine-tuning on the second do main. We perform experiments across a variety of problem domains, both in simula tion and on real robots. Our framework is able to align uncalibrated monocular v ideo of a real robot arm to dynamic state-action trajectories of a simulated arm

without paired data. Video demonstrations of our results are available at: https://sites.google.com/view/cycledynamics .
****
Mixture Representation Learning with Coupled Autoencoding Agents
Yeganeh Marghi,Rohan Gala,Uygar Sümbül
Jointly identifying a mixture of discrete and continuous factors of variability can help unravel complex phenomena. We study this problem by proposing an unsupervised framework called coupled mixture VAE (cpl-mixVAE), which utilizes multiple interacting autoencoding agents. The individual agents operate on augmented copies of training samples to learn mixture representations, while being encouraged to reach consensus on the categorical assignments. We provide theoretical justification to motivate the use of a multi-agent framework, and formulate it as a variational inference problem. We benchmark our approach on MNIST and dSprites, achieving state-of-the-art categorical assignments while preserving interpretability of the continuous factors. We then demonstrate the utility of this approach in jointly identifying cell types and type-specific, activity-regulated genes for a single-cell gene expression dataset profiling over 100 cortical neuron types.
****
On the Latent Space of Flow-based Models
Mingtian Zhang,Yitong Sun,Steven McDonagh,Chen Zhang
Flow-based generative models typically define a latent space with dimensionality identical to the observational space. In many problems, however, the data does not populate the full ambient data-space that they natively reside in, but rather inhabit a lower-dimensional manifold. In such scenarios, flow-based models are unable to represent data structures exactly as their density will always have support off the data manifold, potentially resulting in degradation of model performance. In addition, the requirement for equal latent and data space dimensionality can unnecessarily increase model complexity for contemporary flow models. Towards addressing these problems, we propose to learn a manifold prior that affords benefits to both the tasks of sample generation and representation quality. An auxiliary product of our approach is that we are able to identify the intrinsic dimension of the data distribution.
****
On Data-Augmentation and Consistency-Based Semi-Supervised Learning
Atin Ghosh,Alexandre H. Thiery
Recently proposed consistency-based Semi-Supervised Learning (SSL) methods such as the Pi-model, temporal ensembling, the mean teacher, or the virtual adversarial training, achieve the state of the art results in several SSL tasks. These methods can typically reach performances that are comparable to their fully supervised counterparts while using only a fraction of labelled examples. Despite these methodological advances, the understanding of these methods is still relatively limited. To make progress, we analyse (variations of) the Pi-model in settings where analytically tractable results can be obtained. We establish links with Manifold Tangent Classifiers and demonstrate that the quality of the perturbations is key to obtaining reasonable SSL performances. Furthermore, we propose a simple extension of the Hidden Manifold Model that naturally incorporates data-augmentation schemes and offers a tractable framework for understanding SSL methods.
****
SpreadsheetCoder: Formula Prediction from Semi-structured Context
Xinyun Chen,Petros Maniatis,Rishabh Singh,Charles Sutton,Hanjun Dai,Max Lin,Denny Zhou
Spreadsheet formula prediction has been an important program synthesis problem with many real-world applications. Previous works typically utilize input-output examples as the specification for spreadsheet formula synthesis, where each input-output pair simulates a separate row in the spreadsheet. However, such a formulation does not fully capture the rich context in real-world spreadsheets. First, spreadsheet data entries are organized as tables, thus rows and columns are not necessarily independent from each other. In addition, many spreadsheet tables include headers, which provide high-level descriptions of the cell data. However

, previous synthesis approaches do not consider headers as part of the specifica tion. In this work, we present the first approach for synthesizing spreadsheet f ormulas from tabular context, which includes both headers and semi-structured ta bular data. In particular, we propose SpreadsheetCoder, a BERT-based model archi tecture to represent the tabular context in both row-based and column-based form ats. We train our model on a large dataset of spreadsheets, and demonstrate that SpreadsheetCoder achieves top-1 prediction accuracy of 42.51%, which is a co nsiderable improvement over baselines that do not employ rich tabular context.
****
Informative Outlier Matters: Robustifying Out-of-distribution Detection Using Ou tlier Mining
Jiefeng Chen,Yixuan Li,Xi Wu,Yingyu Liang,Somesh Jha
Detecting out-of-distribution (OOD) inputs is critical for safely deploying deep learning models in an open-world setting. However, existing OOD detection solut ions can be brittle in the open world, facing various types of adversarial OOD i nputs. While methods leveraging auxiliary OOD data have emerged, our analysis re veals a key insight that the majority of auxiliary OOD examples may not meaningf ully improve the decision boundary of the OOD detector. In this paper, we provid e a theoretically motivated method, Adversarial Training with informative Outlie r Mining (ATOM), which improves the robustness of OOD detection. We show that, b y mining informative auxiliary OOD data, one can significantly improve OOD detec tion performance, and somewhat surprisingly, generalize to unseen adversarial at tacks. ATOM achieves state-of-the-art performance under a broad family of classi c and adversarial OOD evaluation tasks. For example, on the CIFAR-10 in-distribu tion dataset, ATOM reduces the FPR95 by up to 57.99\% under adversarial OOD inpu ts, surpassing the previous best baseline by a large margin.
****
DDPNOpt: Differential Dynamic Programming Neural Optimizer
Guan-Horng Liu,Tianrong Chen,Evangelos Theodorou
Interpretation of Deep Neural Networks (DNNs) training as an optimal control pro blem with nonlinear dynamical systems has received considerable attention recent ly, yet the algorithmic development remains relatively limited. In this work, we make an attempt along this line by reformulating the training procedure from th e trajectory optimization perspective. We first show that most widely-used algor ithms for training DNNs can be linked to the Differential Dynamic Programming (D DP), a celebrated second-order method rooted in the Approximate Dynamic Programm ing. In this vein, we propose a new class of optimizer, DDP Neural Optimizer (DD PNOpt), for training feedforward and convolution networks. DDPNOpt features laye r-wise feedback policies which improve convergence and reduce sensitivity to hyp er-parameter over existing methods. It outperforms other optimal-control inspire d training methods in both convergence and complexity, and is competitive agains t state-of-the-art first and second order methods. We also observe DDPNOpt has s urprising benefit in preventing gradient vanishing. Our work opens up new avenue s for principled algorithmic design built upon the optimal control theory.
****
Differentiate Everything with a Reversible Domain-Specific Language
JinGuo Liu,Taine Zhao
Reverse-mode automatic differentiation (AD) suffers from the issue of having too much space overhead to trace back intermediate computational states for backpro pagation.
The traditional method to trace back states is called checkpointing that stores intermediate states into a global stack and restore state through either stack p op or re-computing.
The overhead of stack manipulations and re-computing makes the general purposed (or not tensor-based) AD engines unable to meet many industrial needs.
Instead of checkpointing, we propose to use reverse computing to trace back stat es by designing and implementing a reversible programming eDSL, where a program can be executed bi-directionally without implicit stack operations. The absence of implicit stack operations makes the program compatible with existing compiler features, including utilizing existing optimization passes and compiling the co

de as GPU kernels.
We implement AD for sparse matrix operations and some machine learning applicati ons to show that our framework has state-of-the-art performance.
****

Diverse Video Generation using a Gaussian Process Trigger
Gaurav Shrivastava,Abhinav Shrivastava
Generating future frames given a few context (or past) frames is a challenging t ask. It requires modeling the temporal coherence of videos as well as multi-moda lity in terms of diversity in the potential future states. Current variational a pproaches for video generation tend to marginalize over multi-modal future outco mes. Instead, we propose to explicitly model the multi-modality in the future ou tcomes and leverage it to sample diverse futures. Our approach, Diverse Video Ge nerator, uses a GP to learn priors on future states given the past and maintains a probability distribution over possible futures given a particular sample. We leverage the changes in this distribution over time to control the sampling of d iverse future states by estimating the end of on-going sequences. In particular, we use the variance of GP over the output function space to trigger a change in the action sequence. We achieve state-of-the-art results on diverse future fram e generation in terms of reconstruction quality and diversity of the generated s equences.
****

RRL: A Scalable Classifier for Interpretable Rule-Based Representation Learning
Zhuo Wang,Wei Zhang,Ning Liu,Jianyong Wang
Rule-based models, e.g., decision trees, are widely used in scenarios demanding high model interpretability for their transparent inner structures and good mode l expressivity. However, rule-based models are hard to optimize, especially on l arge data sets, due to their discrete parameters and structures. Ensemble method s and fuzzy/soft rules are commonly used to tackle these issues, but they sacrif ice the model interpretability. In this paper, we propose a new classifier, name d Rule-based Representation Learner (RRL), that automatically learns interpretab le non-fuzzy rules for data representation. To train the non-differentiable RRL effectively, we project it to a continuous space and propose a novel training me thod, called Gradient Grafting, that can directly optimize the discrete model us ing gradient descent. An improved design of logical activation functions is also devised to increase the scalability of RRL and enable it to discretize the cont inuous features end-to-end. Exhaustive experiments on 9 small and 4 large data s ets show that RRL outperforms the competitive approaches, has low complexity clo se to the simple decision trees, and is rational for its main technical contribu tions.
****

Monotonic Kronecker-Factored Lattice
William Taylor Bakst,Nobuyuki Morioka,Erez Louidor
It is computationally challenging to learn flexible monotonic functions that gua rantee model behavior and provide interpretability beyond a few input features, and in a time where minimizing resource use is increasingly important, we must b e able to learn such models that are still efficient. In this paper we show how to effectively and efficiently learn such functions using Kronecker-Factored Lat tice (KFL), an efficient reparameterization of flexible monotonic lat tice regression via Kronecker product. Both computational and storage costs scal e linearly in the number of input features, which is a significant improvement o ver existing methods that grow exponentially. We also show that we can still pro perly enforce monotonicity and other shape constraints. The KFL funct ion class consists of products of piecewise-linear functions, and the size of th e function class can be further increased through ensembling. We prove that the function class of an ensemble of M base KFL models strictly increas es as M increases up to a certain threshold. Beyond this threshold, every mult ilinear interpolated lattice function can be expressed. Our experimental results demonstrate that KFL trains faster with fewer parameters while still achieving accuracy and evaluation speeds comparable to or better than the basel ine methods and preserving monotonicity guarantees on the learned model.

****
Training By Vanilla SGD with Larger Learning Rates
Yueyao Yu,Jie Wang,Wenye Li,Yin Zhang
The stochastic gradient descent (SGD) method, first proposed in 1950's, has been the foundation for deep-neural-network (DNN) training with numerous enhancements including adding a momentum or adaptively selecting learning rates, or using both strategies and more. A common view for SGD is that the learning rate should be eventually made small in order to reach sufficiently good approximate solutions. Another widely held view is that the vanilla SGD is out of fashion in comparison to many of its modern variations. In this work, we provide a contrarian claim that, when training over-parameterized DNNs, the vanilla SGD can still compete well with, and oftentimes outperform, its more recent variations by simply using learning rates significantly larger than commonly used values. We establish theoretical results to explain this local convergence behavior of SGD on nonconvex functions, and also present computational evidence, across multiple tasks including image classification, speech recognition and natural language processing, to support the practice of using larger learning rates.

****
Auto Seg-Loss: Searching Metric Surrogates for Semantic Segmentation
Hao Li,Chenxin Tao,Xizhou Zhu,Xiaogang Wang,Gao Huang,Jifeng Dai
Designing proper loss functions is essential in training deep networks. Especially in the field of semantic segmentation, various evaluation metrics have been proposed for diverse scenarios. Despite the success of the widely adopted cross-entropy loss and its variants, the mis-alignment between the loss functions and evaluation metrics degrades the network performance. Meanwhile, manually designing loss functions for each specific metric requires expertise and significant manpower. In this paper, we propose to automate the design of metric-specific loss functions by searching differentiable surrogate losses for each metric. We substitute the non-differentiable operations in the metrics with parameterized functions, and conduct parameter search to optimize the shape of loss surfaces. Two constraints are introduced to regularize the search space and make the search efficient. Extensive experiments on PASCAL VOC and Cityscapes demonstrate that the searched surrogate losses outperform the manually designed loss functions consistently. The searched losses can generalize well to other datasets and networks. Code shall be released at https://github.com/fundamentalvision/Auto-Seg-Loss.
****
Deformable DETR: Deformable Transformers for End-to-End Object Detection
Xizhou Zhu,Weijie Su,Lewei Lu,Bin Li,Xiaogang Wang,Jifeng Dai
DETR has been recently proposed to eliminate the need for many hand-designed components in object detection while demonstrating good performance. However, it suffers from slow convergence and limited feature spatial resolution, due to the limitation of Transformer attention modules in processing image feature maps. To mitigate these issues, we proposed Deformable DETR, whose attention modules only attend to a small set of key sampling points around a reference. Deformable DETR can achieve better performance than DETR (especially on small objects) with 10 $\times$ less training epochs. Extensive experiments on the COCO benchmark demonstrate the effectiveness of our approach. Code is released at https://github.com/fundamentalvision/Deformable-DETR.
****
Interpretable Sequence Classification Via Prototype Trajectory
Dat Hong,Stephen Baek,Tong Wang
We propose a novel interpretable recurrent neural network (RNN) model, called ProtoryNet, in which we introduce a new concept of prototype trajectories. Motivated by the prototype theory in modern linguistics, ProtoryNet makes a prediction by finding the most similar prototype for each sentence in a text sequence and feeding an RNN backbone with the proximity of each of the sentences to the prototypes. The RNN backbone then captures the temporal pattern of the prototypes, to which we refer as prototype trajectories. The prototype trajectories enable intuitive, fine-grained interpretation of how the model reached to the final predict

ion, resembling the process of how humans analyze paragraphs. Experiments conduc
ted on multiple public data sets reveal that the proposed method not only is mor
e interpretable but also is more accurate than the current state-of-the-art prot
otype-based method. Furthermore, we report a survey result indicating that human
 users find ProtoryNet more intuitive and easier to understand, compared to the
other prototype-based methods.
****

Whitening and second order optimization both destroy information about the datas
et, and can make generalization impossible
Neha S. Wadia,Daniel Duckworth,Samuel Stern Schoenholz,Ethan Dyer,Jascha Sohl-Di
ckstein
Machine learning is predicated on the concept of generalization: a model achievi
ng low error on a sufficiently large training set should also perform well on no
vel samples from the same distribution. We show that both data whitening and sec
ond order optimization can harm or entirely prevent generalization. In general,
model training harnesses information contained in the sample-sample second momen
t matrix of a dataset. For a general class of models, namely models with a fully
 connected first layer, we prove that the information contained in this matrix i
s the only information which can be used to generalize. Models trained using whi
tened data, or with certain second order optimization schemes, have less access
to this information; in the high dimensional regime they have no access at all,
resulting in poor or nonexistent generalization ability. We experimentally verif
y these predictions for several architectures, and further demonstrate that gene
ralization continues to be harmed even when theoretical requirements are relaxed
. However, we also show experimentally that regularized second order optimizatio
n can provide a practical tradeoff, where training is accelerated but less infor
mation is lost, and generalization can in some circumstances even improve.
****

Imbalanced Gradients: A New Cause of Overestimated Adversarial Robustness
Linxi Jiang,Xingjun Ma,Zejia Weng,James Bailey,Yu-Gang Jiang
Evaluating the robustness of a defense model is a challenging task in adversaria
l robustness research. Obfuscated gradients, a type of gradient masking, have pr
eviously been found to exist in many defense methods and cause a false signal of
 robustness. In this paper, we identify a more subtle situation called \emph{Imb
alanced Gradients} that can also cause overestimated adversarial robustness. The
 phenomenon of imbalanced gradients occurs when the gradient of one term of the
margin loss dominates and pushes the attack towards to a suboptimal direction. T
o exploit imbalanced gradients, we formulate a \emph{Margin Decomposition (MD)}
attack that decomposes a margin loss into individual terms and then explores the
 attackability of these terms separately via a two-stage process. We examine 12
state-of-the-art defense models, and find that models exploiting label smoothing
 easily cause imbalanced gradients, and on which our MD attacks can decrease the
ir PGD robustness (evaluated by PGD attack) by over 23\%. For 6 out of the 12 de
fenses, our attack can reduce their PGD robustness by at least 9\%. The results
suggest that imbalanced gradients need to be carefully addressed for more reliab
le adversarial robustness.
****

Democratizing Evaluation of Deep Model Interpretability through Consensus
Xuhong Li,Haoyi Xiong,Siyu Huang,Shilei Ji,Yanjie Fu,Dejing Dou
Deep learning interpretability tools, such as (Bau et al., 2017; Ribeiro et al.,
 2016; Smilkov et al., 2017), have been proposed to explain and visualize the wa
ys that deep neural networks make predictions. The success of these methods high
ly relies on human subjective interpretations, i.e., the ground truth of interpr
etations, such as feature importance ranking or locations of visual objects, whe
n evaluating the interpretability of the deep models on a speci■c task. For task
s that the ground truth of interpretations is not available, we propose a novel
framework Consensus incorporating an ensemble of deep models as the committee fo
r interpretability evaluation. Given any task/dataset, Consensus ■rst obtains th
e interpretation results using existing tools, e.g., LIME (Ribeiro et al., 2016)
, for every model in the committee, then aggregates the results from the entire

committee and approximates the “ground truth” of interpretations through voting. With such approximated ground truth, Consensus evaluates the interpretability of a model through matching its interpretation result and the approximated one, and ranks the matching scores together with committee members, so as to pursue the absolute and relative interpretability evaluation results. We carry out extensive experiments to validate Consensus on various datasets. The results show that Consensus can precisely identify the interpretability for a wide range of models on ubiquitous datasets that the ground truth is not available. Robustness analyses further demonstrate the advantage of the proposed framework to reach the consensus of interpretations through simple voting and evaluate the interpretability of deep models. Through the proposed Consensus framework, the interpretability evaluation has been democratized without the need of ground truth as criterion.

****

Iterative Graph Self-Distillation

Hanlin Zhang,Shuai Lin,Weiyang Liu,Pan Zhou,Jian Tang,Xiaodan Liang,Eric Xing

How to discriminatively vectorize graphs is a fundamental challenge that attracts increasing attentions in recent years. Motivated by the recent success of unsupervised contrastive learning, we aim to learn graph-level representation in an unsupervised manner. Specifically, we propose a novel unsupervised graph learning paradigm called Iterative Graph Self-Distillation (IGSD) which iteratively performs the teacher-student distillation with graph augmentations. Different from conventional knowledge distillation, IGSD constructs the teacher with an exponential moving average of the student model and distills the knowledge of itself. The intuition behind IGSD is to predict the teacher network representation of the graph pairs under different augmented views. As a natural extension, we also apply IGSD to semi-supervised scenarios by jointly regularizing the network with both supervised and unsupervised contrastive loss. Finally, we show that finetuning the IGSD-trained models with self-training can further improve the graph representation power. Empirically, we achieve significant and consistent performance gain on various graph datasets in both unsupervised and semi-supervised settings, which well validates the superiority of IGSD.

****

Combining Physics and Machine Learning for Network Flow Estimation

Arlei Lopes da Silva,Furkan Kocayusufoglu,Saber Jafarpour,Francesco Bullo,Ananthram Swami,Ambuj Singh

The flow estimation problem consists of predicting missing edge flows in a network (e.g., traffic, power, and water) based on partial observations. These missing flows depend both on the underlying \textit{physics} (edge features and a flow conservation law) as well as the observed edge flows. This paper introduces an optimization framework for computing missing edge flows and solves the problem using bilevel optimization and deep learning. More specifically, we learn regularizers that depend on edge features (e.g., number of lanes in a road, the resistance of a power line) using neural networks. Empirical results show that our method accurately predicts missing flows, outperforming the best baseline, and is able to capture relevant physical properties in traffic and power networks.

****

NeMo: Neural Mesh Models of Contrastive Features for Robust 3D Pose Estimation

Angtian Wang,Adam Kortylewski,Alan Yuille

3D pose estimation is a challenging but important task in computer vision. In this work, we show that standard deep learning approaches to 3D pose estimation are not robust to partial occlusion. Inspired by the robustness of generative vision models to partial occlusion, we propose to integrate deep neural networks with 3D generative representations of objects into a unified neural architecture that we term NeMo. In particular, NeMo learns a generative model of neural feature activations at each vertex on a dense 3D mesh. Using differentiable rendering we estimate the 3D object pose by minimizing the reconstruction error between NeMo and the feature representation of the target image. To avoid local optima in the reconstruction loss, we train the feature extractor to maximize the distance between the individual feature representations on the mesh using contrastive lea

rning. Our extensive experiments on PASCAL3D+, occluded-PASCAL3D+ and ObjectNet3D show that NeMo is much more robust to partial occlusion compared to standard deep networks, while retaining competitive performance on non-occluded data. Interestingly, our experiments also show that NeMo performs reasonably well even when the mesh representation only crudely approximates the true object geometry with a cuboid, hence revealing that the detailed 3D geometry is not needed for accurate 3D pose estimation.
****
How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision
Dongkwan Kim,Alice Oh
Attention mechanism in graph neural networks is designed to assign larger weights to important neighbor nodes for better representation. However, what graph attention learns is not understood well, particularly when graphs are noisy. In this paper, we propose a self-supervised graph attention network (SuperGAT), an improved graph attention model for noisy graphs. Specifically, we exploit two attention forms compatible with a self-supervised task to predict edges, whose presence and absence contain the inherent information about the importance of the relationships between nodes. By encoding edges, SuperGAT learns more expressive attention in distinguishing mislinked neighbors. We find two graph characteristics influence the effectiveness of attention forms and self-supervision: homophily and average degree. Thus, our recipe provides guidance on which attention design to use when those two graph characteristics are known. Our experiment on 17 real-world datasets demonstrates that our recipe generalizes across 15 datasets of them, and our models designed by recipe show improved performance over baselines.
****
MiCE: Mixture of Contrastive Experts for Unsupervised Image Clustering
Tsung Wei Tsai,Chongxuan Li,Jun Zhu
We present Mixture of Contrastive Experts (MiCE), a unified probabilistic clustering framework that simultaneously exploits the discriminative representations learned by contrastive learning and the semantic structures captured by a latent mixture model. Motivated by the mixture of experts, MiCE employs a gating function to partition an unlabeled dataset into subsets according to the latent semantics and multiple experts to discriminate distinct subsets of instances assigned to them in a contrastive learning manner. To solve the nontrivial inference and learning problems caused by the latent variables, we further develop a scalable variant of the Expectation-Maximization (EM) algorithm for MiCE and provide proof of the convergence. Empirically, we evaluate the clustering performance of MiCE on four widely adopted natural image datasets. MiCE achieves significantly better results than various previous methods and a strong contrastive learning baseline.
****
Anytime Sampling for Autoregressive Models via Ordered Autoencoding
Yilun Xu,Yang Song,Sahaj Garg,Linyuan Gong,Rui Shu,Aditya Grover,Stefano Ermon
Autoregressive models are widely used for tasks such as image and audio generation. The sampling process of these models, however, does not allow interruptions and cannot adapt to real-time computational resources. This challenge impedes the deployment of powerful autoregressive models, which involve a slow sampling process that is sequential in nature and typically scales linearly with respect to the data dimension. To address this difficulty, we propose a new family of autoregressive models that enables anytime sampling. Inspired by Principal Component Analysis, we learn a structured representation space where dimensions are ordered based on their importance with respect to reconstruction. Using an autoregressive model in this latent space, we trade off sample quality for computational efficiency by truncating the generation process before decoding into the original data space. Experimentally, we demonstrate in several image and audio generation tasks that sample quality degrades gracefully as we reduce the computational budget for sampling. The approach suffers almost no loss in sample quality (measured by FID) using only 60\% to 80\% of all latent dimensions for image data. Code is available at https://github.com/Newbeeer/Anytime-Auto-Regressive-Model.

****
Rethinking Convolution: Towards an Optimal Efficiency
Tao Wei,Yonghong Tian,Chang Wen Chen
In this paper, we present our recent research about the computational efficiency in convolution. Convolution operation is the most critical component in recent surge of deep learning research. Conventional 2D convolution takes $O(C^{2}K^{2}HW)$ to calculate, where C is the channel size, K is the kernel size, while H and W are the output height and width. Such computation has become really costly considering that these parameters increased over the past few years to meet the needs of demanding applications. Among various implementation of the convolution, separable convolution has been proven to be more efficient in reducing the computational demand. For example, depth separable convolution reduces the complexity to $O(CHW\cdot(C+K^{2}))$ while spatial separable convolution reduces the complexity to $O(C^{2}KHW)$. However, these are considered an ad hoc design which cannot ensure that they can in general achieve optimal separation. In this research, we propose a novel operator called \emph{optimal separable convolution} which can be calculated at $O(C^{\frac{3}{2}}KHW)$ by optimal design for the internal number of groups and kernel sizes for general separable convolutions. When there is no restriction in the number of separated convolutions, an even lower complexity at $O(CHW\cdot\log(CK^{2}))$ can be achieved. Experimental results demonstrate that the proposed optimal separable convolution is able to achieve an improved accuracy-FLOPs and accuracy-#Params trade-offs over both conventional and depth/spatial separable convolutions.
****
Initialization and Regularization of Factorized Neural Layers
Mikhail Khodak,Neil A. Tenenholtz,Lester Mackey,Nicolo Fusi
Factorized layers—operations parameterized by products of two or more matrices—occur in a variety of deep learning contexts, including compressed model training, certain types of knowledge distillation, and multi-head self-attention architectures. We study how to initialize and regularize deep nets containing such layers, examining two simple, understudied schemes, spectral initialization and Frobenius decay, for improving their performance. The guiding insight is to design optimization routines for these networks that are as close as possible to that of their well-tuned, non-decomposed counterparts; we back this intuition with an analysis of how the initialization and regularization schemes impact training with gradient descent, drawing on modern attempts to understand the interplay of weight-decay and batch-normalization. Empirically, we highlight the benefits of spectral initialization and Frobenius decay across a variety of settings. In model compression, we show that they enable low-rank methods to significantly outperform both unstructured sparsity and tensor methods on the task of training low-memory residual networks; analogs of the schemes also improve the performance of tensor decomposition techniques. For knowledge distillation, Frobenius decay enables a simple, overcomplete baseline that yields a compact model from over-parameterized training without requiring retraining with or pruning a teacher network. Finally, we show how both schemes applied to multi-head attention lead to improved performance on both translation and unsupervised pre-training.
****
CO2: Consistent Contrast for Unsupervised Visual Representation Learning
Chen Wei,Huiyu Wang,Wei Shen,Alan Yuille
Contrastive learning has recently been a core for unsupervised visual representation learning. Without human annotation, the common practice is to perform an instance discrimination task: Given a query image crop, label crops from the same image as positives, and crops from other randomly sampled images as negatives. An important limitation of this label assignment is that it can not reflect the heterogeneous similarity of the query crop to crops from other images, but regarding them as equally negative. To address this issue, inspired by consistency regularization in semi-supervised learning, we propose Consistent Contrast (CO2), which introduces a consistency term into unsupervised contrastive learning framework. The consistency term takes the similarity of the query crop to crops from other images as unlabeled, and the corresponding similarity of a positive crop as

a pseudo label. It then encourages consistency between these two similarities. Empirically, CO2 improves Momentum Contrast (MoCo) by 2.9% top-1 accuracy on ImageNet linear protocol, 3.8% and 1.1% top-5 accuracy on 1% and 10% labeled semi-supervised settings. It also transfers to image classification, object detection, and semantic segmentation on PASCAL VOC. This shows that CO2 learns better visual representations for downstream tasks.
****

Geometry-Aware Gradient Algorithms for Neural Architecture Search
Liam Li,Mikhail Khodak,Nina Balcan,Ameet Talwalkar
Recent state-of-the-art methods for neural architecture search (NAS) exploit gradient-based optimization by relaxing the problem into continuous optimization over architectures and shared-weights, a noisy process that remains poorly understood. We argue for the study of single-level empirical risk minimization to understand NAS with weight-sharing, reducing the design of NAS methods to devising optimizers and regularizers that can quickly obtain high-quality solutions to this problem. Invoking the theory of mirror descent, we present a geometry-aware framework that exploits the underlying structure of this optimization to return sparse architectural parameters, leading to simple yet novel algorithms that enjoy fast convergence guarantees and achieve state-of-the-art accuracy on the latest NAS benchmarks in computer vision. Notably, we exceed the best published results for both CIFAR and ImageNet on both the DARTS search space and NAS-Bench-201; on the latter we achieve near-oracle-optimal performance on CIFAR-10 and CIFAR-100. Together, our theory and experiments demonstrate a principled way to co-design optimizers and continuous relaxations of discrete NAS search spaces.
****

Beyond Fully-Connected Layers with Quaternions: Parameterization of Hypercomplex Multiplications with $1/n$ Parameters
Aston Zhang,Yi Tay,SHUAI Zhang,Alvin Chan,Anh Tuan Luu,Siu Hui,Jie Fu
Recent works have demonstrated reasonable success of representation learning in hypercomplex space. Specifically, “fully-connected layers with quaternions” (quaternions are 4D hypercomplex numbers), which replace real-valued matrix multiplications in fully-connected layers with Hamilton products of quaternions, both enjoy parameter savings with only 1/4 learnable parameters and achieve comparable performance in various applications. However, one key caveat is that hypercomplex space only exists at very few predefined dimensions (4D, 8D, and 16D). This restricts the flexibility of models that leverage hypercomplex multiplications. To this end, we propose parameterizing hypercomplex multiplications, allowing models to learn multiplication rules from data regardless of whether such rules are predefined. As a result, our method not only subsumes the Hamilton product, but also learns to operate on any arbitrary nD hypercomplex space, providing more architectural flexibility using arbitrarily $1/n$ learnable parameters compared with the fully-connected layer counterpart. Experiments of applications to the LSTM and transformer models on natural language inference, machine translation, text style transfer, and subject verb agreement demonstrate architectural flexibility and effectiveness of the proposed approach.
****

Asynchronous Advantage Actor Critic: Non-asymptotic Analysis and Linear Speedup
Han Shen,Kaiqing Zhang,Mingyi Hong,Tianyi Chen
Asynchronous and parallel implementation of standard reinforcement learning (RL) algorithms is a key enabler of the tremendous success of modern RL.
Among many asynchronous RL algorithms, arguably the most popular and effective one is the asynchronous advantage actor-critic (A3C) algorithm.
Although A3C is becoming the workhorse of RL, its theoretical properties are still not well-understood, including the non-asymptotic analysis and the performance gain of parallelism (a.k.a. speedup).
This paper revisits the A3C algorithm with TD(0) for the critic update, termed A3C-TD(0), with provable convergence guarantees.
With linear value function approximation for the TD update, the convergence of A3C-TD(0) is established under both i.i.d. and Markovian sampling. Under i.i.d. sampling, A3C-TD(0) obtains sample complexity of $\mathcal{O}(\epsilon^{-2.5}/N)$

per worker to achieve ϵ accuracy, where N is the number of workers. Compared to the best-known sample complexity of $\mathcal{O}(\epsilon^{-2.5})$ for two-timescale AC, A3C-TD(0) achieves \emph{linear speedup}, which justifies the advantage of parallelism and asynchrony in AC algorithms theoretically for the first time.
Numerical tests on synthetically generated instances and OpenAI Gym environments have been provided to verify our theoretical analysis.
****
Tent: Fully Test-Time Adaptation by Entropy Minimization
Dequan Wang,Evan Shelhamer,Shaoteng Liu,Bruno Olshausen,Trevor Darrell
A model must adapt itself to generalize to new and different data during testing. In this setting of fully test-time adaptation the model has only the test data and its own parameters. We propose to adapt by test entropy minimization (tent): we optimize the model for confidence as measured by the entropy of its predictions. Our method estimates normalization statistics and optimizes channel-wise affine transformations to update online on each batch. Tent reduces generalization error for image classification on corrupted ImageNet and CIFAR-10/100 and reaches a new state-of-the-art error on ImageNet-C. Tent handles source-free domain adaptation on digit recognition from SVHN to MNIST/MNIST-M/USPS, on semantic segmentation from GTA to Cityscapes, and on the VisDA-C benchmark. These results are achieved in one epoch of test-time optimization without altering training.
****
Bidirectionally Self-Normalizing Neural Networks
Yao Lu,Stephen Gould,Thalaiyasingam Ajanthan
The problem of exploding and vanishing gradients has been a long-standing obstacle that hinders the effective training of neural networks. Despite various tricks and techniques that have been employed to alleviate the problem in practice, there still lacks satisfactory theories or provable solutions. In this paper, we address the problem from the perspective of high-dimensional probability theory. We provide a rigorous result that shows, under mild conditions, how the exploding/vanishing gradient problem disappears with high probability if the neural networks have sufficient width. Our main idea is to constrain both forward and backward signal propagation in a nonlinear neural network through a new class of activation functions, namely Gaussian-Poincaré normalized functions, and orthogonal weight matrices. Experiments on both synthetic and real-world data validate our theory and confirm its effectiveness on very deep neural networks when applied in practice.
****
Reducing Class Collapse in Metric Learning with Easy Positive Sampling
Elad Levi,Tete Xiao,Xiaolong Wang,trevor darrell
Metric learning seeks perceptual embeddings where visually similar instances are close and dissimilar instances are apart, but learned representation can be sub-optimal when the distribution of intra-class samples is diverse and distinct sub-clusters are present. We theoretically prove and empirically show that under reasonable noise assumptions, prevalent embedding losses in metric learning, e.g., triplet loss, tend to project all samples of a class with various modes onto a single point in the embedding space, resulting in a class collapse that usually renders the space ill-sorted for classification or retrieval. To address this problem, we propose a simple modification to the embedding losses such that each sample selects its nearest same-class counterpart in a batch as the positive element in the tuple/triplet. This allows for the presence of multiple sub-clusters within each class. The adaptation can be integrated into a wide range of metric learning losses. Our method demonstrates clear benefits on various fine-grained image retrieval datasets over a variety of existing losses; qualitative retrieval results show that samples with similar visual patterns are indeed closer in the embedding space.

****
Crowd-sourced Phrase-Based Tokenization for Low-Resourced Neural Machine Translation: The case of Fon Language

Bonaventure F. P. Dossou,Chris Chinenye Emezue
Building effective neural machine translation (NMT) models for very low-resource d and morphologically rich African indigenous languages is an open challenge. Be sides the issue of finding available resources for them, a lot of work is put in to preprocessing and tokenization. Recent studies have shown that standard token ization methods do not always adequately deal with the grammatical, diacritical, and tonal properties of some African languages. That, coupled with the extremel y low availability of training samples, hinders the production of reliable NMT m odels. In this paper, using Fon language as a case study, we revisit standard to kenization methods and introduce Word-Expressions-Based (WEB) tokenization, a hu man-involved super-words tokenization strategy to create a better representative vocabulary for training.
****
DC3: A learning method for optimization with hard constraints
Priya L. Donti,David Rolnick,J Zico Kolter
Large optimization problems with hard constraints arise in many settings, yet cl assical solvers are often prohibitively slow, motivating the use of deep network s as cheap "approximate solvers." Unfortunately, naive deep learning approaches typically cannot enforce the hard constraints of such problems, leading to infea sible solutions. In this work, we present Deep Constraint Completion and Correct ion (DC3), an algorithm to address this challenge. Specifically, this method enf orces feasibility via a differentiable procedure, which implicitly completes par tial solutions to satisfy equality constraints and unrolls gradient-based correc tions to satisfy inequality constraints. We demonstrate the effectiveness of DC3 in both synthetic optimization tasks and the real-world setting of AC optimal p ower flow, where hard constraints encode the physics of the electrical grid. In both cases, DC3 achieves near-optimal objective values while preserving feasibil ity.
****
Symmetric Wasserstein Autoencoders
Sun Sun,Hongyu Guo
Leveraging the framework of Optimal Transport, we introduce a new family of gene rative autoencoders with a learnable prior, called Symmetric Wasserstein Autoenc oders (SWAEs). We propose to symmetrically match the joint distributions of the observed data and the latent representation induced by the encoder and the decod er. The resultant algorithm jointly optimizes the modelling losses in both the d ata and the latent spaces with the loss in the data space leading to the denois ing effect. With the symmetric treatment of the data and the latent representati on, the algorithm implicitly preserves the local structure of the data in the la tent space. To further improve the latent representation, we incorporate a recon struction loss into the objective, which significantly benefits both the generat ion and reconstruction. We empirically show the superior performance of SWAEs ov er several state-of-the-art generative autoencoders in terms of classification, reconstruction, and generation.
****
ResPerfNet: Deep Residual Learning for Regressional Performance Modeling of Deep Neural Networks
Chuan-Chi Wang,Ying-Chiao Liao,Chia-Heng Tu,Ming-Chang Kao,Wen-Yew Liang,Shih-Ha o Hung
The rapid advancements of computing technology facilitate the development of div erse deep learning applications. Unfortunately, the efficiency of parallel compu ting infrastructures varies widely with neural network models, which hinders the exploration of the design space to find high-performance neural network archite ctures on specific computing platforms for a given application. To address such a challenge, we propose a deep learning-based method, ResPerfNet, which trains a residual neural network with representative datasets obtained on the target pla tform to predict the performance for a deep neural network. Our experimental res ults show that ResPerfNet can accurately predict the execution time of individua l neural network layers and full network models on a variety of platforms. In pa rticular, ResPerfNet achieves 8.4% of mean absolute percentage error for LeNet,

AlexNet and VGG16 on the NVIDIA GTX 1080Ti, which is substantially lower than th
e previously published works.
****

Neighborhood-Aware Neural Architecture Search
Xiaofang Wang,Shengcao Cao,Mengtian Li,Kris M. Kitani
Existing neural architecture search (NAS) methods often return an architecture w
ith good search performance but generalizes poorly to the test setting. To achie
ve better generalization, we propose a novel neighborhood-aware NAS formulation
to identify flat-minima architectures in the search space, with the assumption t
hat flat minima generalize better than sharp minima. The phrase ``flat-minima ar
chitecture'' refers to architectures whose performance is stable under small per
turbations in the architecture (\emph{e.g.}, replacing a convolution with a skip
 connection). Our formulation takes the ``flatness'' of an architecture into acc
ount by aggregating the performance over the neighborhood of this architecture.
We demonstrate a principled way to apply our formulation to existing search algo
rithms, including sampling-based algorithms and gradient-based algorithms. To fa
cilitate the application to gradient-based algorithms, we also propose a differe
ntiable representation for the neighborhood of architectures. Based on our formu
lation, we propose neighborhood-aware random search (NA-RS) and neighborhood-awa
re differentiable architecture search (NA-DARTS). Notably, by simply augmenting
DARTS~\cite{liu2018darts} with our formulation, NA-DARTS finds architectures tha
t perform better or on par with those found by state-of-the-art NAS methods on e
stablished benchmarks, including CIFAR-10, CIAFR-100 and ImageNet.
****

Enforcing robust control guarantees within neural network policies
Priya L. Donti,Melrose Roderick,Mahyar Fazlyab,J Zico Kolter
When designing controllers for safety-critical systems, practitioners often face
 a challenging tradeoff between robustness and performance. While robust control
 methods provide rigorous guarantees on system stability under certain worst-cas
e disturbances, they often yield simple controllers that perform poorly in the a
verage (non-worst) case. In contrast, nonlinear control methods trained using de
ep learning have achieved state-of-the-art performance on many control tasks, bu
t often lack robustness guarantees. In this paper, we propose a technique that c
ombines the strengths of these two approaches: constructing a generic nonlinear
control policy class, parameterized by neural networks, that nonetheless enforce
s the same provable robustness criteria as robust control. Specifically, our app
roach entails integrating custom convex-optimization-based projection layers int
o a neural network-based policy. We demonstrate the power of this approach on se
veral domains, improving in average-case performance over existing robust contro
l methods and in worst-case stability over (non-robust) deep RL methods.
****

Neural Topic Model via Optimal Transport
He Zhao,Dinh Phung,Viet Huynh,Trung Le,Wray Buntine
Recently, Neural Topic Models (NTMs) inspired by variational autoencoders have o
btained increasingly research interest due to their promising results on text an
alysis. However, it is usually hard for existing NTMs to achieve good document r
epresentation and coherent/diverse topics at the same time. Moreover, they often
 degrade their performance severely on short documents. The requirement of repar
ameterisation could also comprise their training quality and model flexibility.
To address these shortcomings, we present a new neural topic model via the theor
y of optimal transport (OT). Specifically, we propose to learn the topic distrib
ution of a document by directly minimising its OT distance to the document's wor
d distributions. Importantly, the cost matrix of the OT distance models the weig
hts between topics and words, which is constructed by the distances between topi
cs and words in an embedding space. Our proposed model can be trained efficientl
y with a differentiable loss. Extensive experiments show that our framework sign
ificantly outperforms the state-of-the-art NTMs on discovering more coherent and
 diverse topics and deriving better document representations for both regular an
d short texts.
****

Robust and Generalizable Visual Representation Learning via Random Convolutions
Zhenlin Xu,Deyi Liu,Junlin Yang,Colin Raffel,Marc Niethammer
While successful for various computer vision tasks, deep neural networks have sh
own to be vulnerable to texture style shifts and small perturbations to which hu
mans are robust. In this work, we show that the robustness of neural networks ca
n be greatly improved through the use of random convolutions as data augmentatio
n. Random convolutions are approximately shape-preserving and may distort local
textures. Intuitively, randomized convolutions create an infinite number of new
domains with similar global shapes but random local texture. Therefore, we explo
re using outputs of multi-scale random convolutions as new images or mixing them
 with the original images during training. When applying a network trained with
our approach to unseen domains, our method consistently improves the performance
 on domain generalization benchmarks and is scalable to ImageNet. In particular,
 in the challenging scenario of generalizing to the sketch domain in PACS and to
 ImageNet-Sketch, our method outperforms state-of-art methods by a large margin.
 More interestingly, our method can benefit downstream tasks by providing a more
 robust pretrained visual representation.
****

WrapNet: Neural Net Inference with Ultra-Low-Precision Arithmetic
Renkun Ni,Hong-min Chu,Oscar Castaneda,Ping-yeh Chiang,Christoph Studer,Tom Gold
stein
Low-precision neural networks represent both weights and activations with few bi
ts, drastically reducing the cost of multiplications. Meanwhile, these products
are accumulated using high-precision (typically 32-bit) additions. Additions do
minate the arithmetic complexity of inference in quantized (e.g., binary) nets,
and high precision is needed to avoid overflow. To further optimize inference, w
e propose WrapNet, an architecture that adapts neural networks to use low-precis
ion (8-bit) additions while achieving classification accuracy comparable to thei
r 32-bit counterparts. We achieve resilience to low-precision accumulation by in
serting a cyclic activation layer that makes results invariant to overflow. We d
emonstrate the efficacy of our approach using both software and hardware platfor
ms.
****

Bridging Graph Network to Lifelong Learning with Feature Interaction
Chen Wang,Yuheng Qiu,Sebastian Scherer
Graph neural networks (GNN) are powerful models for many graph-structured tasks.
 In this paper, we aim to bridge GNN to lifelong learning, which is to overcome
the effect of ``catastrophic forgetting" for continuously learning a sequence of
 graph-structured tasks. Although many lifelong learning techniques for convolut
ional neural networks (CNN) have been developed, lifelong learning for GNN is st
ill underexplored and suffers from incomplete graph structure during learning. T
his is because in lifelong learning the nodes increase dynamically and can only
be present to the model once, which makes many graph models and sampling strateg
ies inapplicable. To solve this problem, we propose a new graph topology based o
n feature interaction, called the feature graph. It takes features as new nodes
and turns nodes into independent graphs. This successfully converts the original
 problem of node classification to graph classification, in which the increasing
 nodes are turned into training samples. Therefore, the lifelong learning techni
ques developed for CNN become applicable to GNN for the first time. In the exper
iments, we demonstrate both the efficiency and effectiveness of feature graphs f
or lifelong learning tasks using a rehearsal method. We expect that it will have
 broad potential applications for graph-structured tasks in lifelong learning.
****

Score-based Causal Discovery from Heterogeneous Data
Chenwei Ding,Biwei Huang,Mingming Gong,Kun Zhang,Tongliang Liu,Dacheng Tao
Causal discovery has witnessed significant progress over the past decades. Most
algorithms in causal discovery consider a single domain with a fixed distributio
n. However, it is commonplace to encounter heterogeneous data (data from differe
nt domains with distribution shifts). Applying existing methods on such heteroge
neous data may lead to spurious edges or incorrect directions in the learned gra

ph. In this paper, we develop a novel score-based approach for causal discovery from heterogeneous data. Specifically, we propose a Multiple-Domain Score Search (MDSS) algorithm, which is guaranteed to find the correct graph skeleton asymptotically. Furthermore, benefiting from distribution shifts, MDSS enables the detection of more causal directions than previous algorithms designed for single domain data. The proposed MDSS can be readily incorporated into off-the-shelf search strategies, such as the greedy search and the policy-gradient-based search. Theoretical analyses and extensive experiments on both synthetic and real data demonstrate the efficacy of our method.
****

ForceNet: A Graph Neural Network for Large-Scale Quantum Chemistry Simulation
Weihua Hu,Muhammed Shuaibi,Abhishek Das,Siddharth Goyal,Anuroop Sriram,Jure Leskovec,Devi Parikh,Larry Zitnick
Machine Learning (ML) has a potential to dramatically accelerate large-scale physics-based simulations. However, practical models for real large-scale and complex problems remain out of reach. Here we present ForceNet, a model for accurate and fast quantum chemistry simulations to accelerate catalyst discovery for renewable energy applications. ForceNet is a graph neural network that uses surrounding 3D molecular structure to estimate per-atom forces---a central capability for performing atomic simulations. The key challenge is to accurately capture highly complex and non-linear quantum interactions of atoms in 3D space, on which forces are dependent. To this end, ForceNet adopts (1) expressive message passing architecture, (2) appropriate choice of basis and non-linear activation functions, and (3) model scaling in terms of network depth and width. We show ForceNet reduces the estimation error of atomic forces by 30% compared to existing ML models, and generalizes well to out-of-distribution structures. Finally, we apply ForceNet to the large-scale catalyst dataset, OC20. We use ForceNet to perform quantum chemistry simulations, where ForceNet is able to achieve 4x higher success rate than existing ML models. Overall, we demonstrate the potential for ML-based simulations to achieve practical usefulness while being orders of magnitude faster than physics-based simulations.
****

Bi-tuning of Pre-trained Representations
Jincheng Zhong,Ximei Wang,Zhi Kou,Jianmin Wang,Mingsheng Long
It is common within the deep learning community to first pre-train a deep neural network from a large-scale dataset and then fine-tune the pre-trained model to a specific downstream task. Recently, both supervised and unsupervised pre-training approaches to learning representations have achieved remarkable advances, which exploit the discriminative knowledge of labels and the intrinsic structure of data, respectively. It follows natural intuition that both discriminative knowledge and intrinsic structure of the downstream task can be useful for fine-tuning, however, existing fine-tuning methods mainly leverage the former and discard the latter. A question arises: How to fully explore the intrinsic structure of data for boosting fine-tuning? In this paper, we propose Bi-tuning, a general learning framework to fine-tuning both supervised and unsupervised pre-trained representations to downstream tasks. Bi-tuning generalizes the vanilla fine-tuning by integrating two heads upon the backbone of pre-trained representations: a classifier head with an improved contrastive cross-entropy loss to better leverage the label information in an instance-contrast way, and a projector head with a newly-designed categorical contrastive learning loss to fully exploit the intrinsic structure of data in a category-consistent way. Comprehensive experiments confirm that Bi-tuning achieves state-of-the-art results for fine-tuning tasks of both supervised and unsupervised pre-trained models by large margins (e.g.~10.7\% absolute rise in accuracy on CUB in low-data regime).
****

CAFE: Catastrophic Data Leakage in Federated Learning
Xiao Jin,Ruijie Du,Pin-Yu Chen,Tianyi Chen
Private training data can be leaked through the gradient sharing mechanism deployed in machine learning systems, such as federated learning (FL).
Increasing batch size is often viewed as a promising defense strategy against da

ta leakage. In this paper, we revisit this defense premise and propose an advanced data leakage attack to efficiently recover batch data from the shared aggregated gradients.
We name our proposed method as \textit{\underline{c}atastrophic d\underline{a}ta leakage in \underline{f}ederated l\underline{e}arning (CAFE)}.
Comparing to existing data leakage attacks, CAFE demonstrates the ability to perform large-batch data leakage attack with high data recovery quality.
Experimental results on vertical and horizontal FL settings have validated the effectiveness of CAFE in recovering private data from the shared aggregated gradients.
Our results suggest that data participated in FL, especially the vertical case, have a high risk of being leaked from the training gradients. Our analysis implies unprecedented and practical data leakage risks in those learning settings.
****

Guarantees for Tuning the Step Size using a Learning-to-Learn Approach
Xiang Wang,Shuai Yuan,Chenwei Wu,Rong Ge
Learning-to-learn---using optimization algorithms to learn a new optimizer---has successfully trained efficient optimizers in practice. This approach relies on meta-gradient descent on a meta-objective based on the trajectory that the optimizer generates. However, there were few theoretical guarantees on how to avoid meta-gradient explosion/vanishing problem, or how to train an optimizer with good generalization performance. In this paper, we study the learning-to-learn approach on a simple problem of tuning the step size for quadratic loss. Our results show that although there is a way to design the meta-objective so that the meta-gradient remains polynomially bounded, computing the meta-gradient directly using backpropagation leads to numerical issues that look similar to gradient explosion/vanishing problems. We also characterize when it is necessary to compute the meta-objective on a separate validation set instead of the original training set. Finally, we verify our results empirically and show that a similar phenomenon appears even for more complicated learned optimizers parametrized by neural networks.
****

Pre-Training by Completing Point Clouds
Hanchen Wang,Qi Liu,Xiangyu Yue,Joan Lasenby,Matt Kusner
There has recently been a flurry of exciting advances in deep learning models on point clouds. However, these advances have been hampered by the difficulty of creating labelled point cloud datasets: sparse point clouds often have unclear label identities for certain points, while dense point clouds are time-consuming to annotate. Inspired by mask-based pre-training in the natural language processing community, we propose a pre-training mechanism based point clouds completion. It works by masking occluded points that result from observations at different camera views. It then optimizes a completion model that learns how to reconstruct the occluded points, given the partial point cloud. In this way, our method learns a pre-trained representation that can identify the visual constraints inherently embedded in real-world point clouds. We call our method Occlusion Completion (OcCo). We demonstrate that OcCo learns representations that improve the semantic understandings as well as generalization on downstream tasks over prior methods, transfer to different datasets, reduce training time and improve label efficiency.
****

Rank the Episodes: A Simple Approach for Exploration in Procedurally-Generated Environments
Daochen Zha,Wenye Ma,Lei Yuan,Xia Hu,Ji Liu
Exploration under sparse reward is a long-standing challenge of model-free reinforcement learning. The state-of-the-art methods address this challenge by introducing intrinsic rewards to encourage exploration in novel states or uncertain environment dynamics. Unfortunately, methods based on intrinsic rewards often fall short in procedurally-generated environments, where a different environment is generated in each episode so that the agent is not likely to visit the same state more than once. Motivated by how humans distinguish good exploration behaviors

by looking into the entire episode, we introduce RAPID, a simple yet effective episode-level exploration method for procedurally-generated environments. RAPID regards each episode as a whole and gives an episodic exploration score from both per-episode and long-term views. Those highly scored episodes are treated as good exploration behaviors and are stored in a small ranking buffer. The agent then imitates the episodes in the buffer to reproduce the past good exploration behaviors. We demonstrate our method on several procedurally-generated MiniGrid environments, a first-person-view 3D Maze navigation task from MiniWorld, and several sparse MuJoCo tasks. The results show that RAPID significantly outperforms the state-of-the-art intrinsic reward strategies in terms of sample efficiency and final performance. The code is available at https://github.com/daochenzha/rapid

****

FASG: Feature Aggregation Self-training GCN for Semi-supervised Node Classification

Gongpei Zhao,Tao Wang,Yidong Li,Yi Jin

Recently, Graph Convolutioal Networks (GCNs) have achieved significant success in many graph-based learning tasks, especially for node classification, due to its excellent ability in representation learning. Nevertheless, it remains challenging for GCN models to obtain satisfying prediction on graphs where few nodes are with known labels. In this paper, we propose a novel self-training algorithm based on GCN to boost semi-supervised node classification on graphs with little supervised information. Inspired by self-supervision strategy, the proposed method introduces an ingenious checking part to add new nodes as supervision after each training epoch to enhance node prediction. In particular, the embedded checking part is designed based on aggregated features, which is more accurate than previous methods and boosts node classification significantly. The proposed algorithm is validated on three public benchmarks in comparison with several state-of-the-art baseline algorithms, and the results illustrate its excellent performance.

****

To be Robust or to be Fair: Towards Fairness in Adversarial Training

Han Xu,Xiaorui Liu,Yaxin Li,Jiliang Tang

Adversarial training algorithms have been proven to be reliable to improve machine learning models' robustness against adversarial examples. However, we find that adversarial training algorithms tend to introduce severe disparity of accuracy and robustness between different groups of data. For instance, PGD adversarially trained ResNet18 model on CIFAR-10 has 93% clean accuracy and 67% PGD l_{∞}-8 adversarial accuracy on the class ''automobile'' but only 59% and 17% on class ''cat''. This phenomenon happens in balanced datasets and does not exist in naturally trained models when only using clean samples. In this work, we theoretically show that this phenomenon can generally happen under adversarial training algorithms which minimize DNN models' robust errors. Motivated by these findings, we propose a Fair-Robust-Learning (FRL) framework to mitigate this unfairness problem when doing adversarial defenses and experimental results validate the effectiveness of FRL.

****

Generative Adversarial Neural Architecture Search with Importance Sampling

SEYED SAEED CHANGIZ REZAEI,Fred X. Han,Di Niu,Mohammad Salameh,Keith G Mills,Shangling Jui

Despite the empirical success of neural architecture search (NAS) in deep learning applications, the optimality, reproducibility and cost of NAS schemes remain hard to assess. The variation in search spaces adopted has further affected a fair comparison between search strategies. In this paper, we focus on search strategies in NAS and propose Generative Adversarial NAS (GA-NAS), promoting stable and reproducible neural architecture search. GA-NAS is theoretically inspired by importance sampling for rare event simulation, and iteratively refits a generator to previously discovered top architectures, thus increasingly focusing on important parts of the search space. We propose an efficient adversarial learning approach in GA-NAS, where the generator is not trained based on a large number of

observations on architecture performance, but based on the relative prediction made by a discriminator, thus significantly reducing the number of evaluations required.
Extensive experiments show that GA-NAS beats the best published results under several cases on the public NAS benchmarks including NAS-Bench-101, NAS-Bench-201, and NAS-Bench-301. We further show that GA-NAS can handle ad-hoc search constraints and search spaces. GA-NAS can find new architectures that enhance EfficientNet and ProxylessNAS in terms of ImageNet Top-1 accuracy and/or the number of parameters by searching in their original search spaces.
****

Temperature check: theory and practice for training models with softmax-cross-entropy losses
Atish Agarwala,Samuel Stern Schoenholz,Jeffrey Pennington,Yann Dauphin
The softmax function combined with a cross-entropy loss is a principled approach to modeling probability distributions that has become ubiquitous in deep learning. The softmax function is defined by a lone hyperparameter, the temperature, that is commonly set to one or regarded as a way to tune model confidence after training; however, less is known about how the temperature impacts training dynamics or generalization performance. In this work we develop a theory of early learning for models trained with softmax-cross-entropy loss and show that the learning dynamics depend crucially on the inverse-temperature β as well as the magnitude of the logits at initialization, $||\beta\bf{z}||_{2}$. We follow up these analytic results with a large-scale empirical study of a variety of model architectures trained on CIFAR10, ImageNet, and IMDB sentiment analysis. We find that generalization performance depends strongly on the temperature, but only weakly on the initial logit magnitude. We provide evidence that the dependence of generalization on β is not due to changes in model confidence, but is a dynamical phenomenon. It follows that the addition of β as a tunable hyperparameter is key to maximizing model performance. Although we find the optimal β to be sensitive to the architecture, our results suggest that tuning β over the range 10^{-2} to 10^1 improves performance over all architectures studied. We find that smaller β may lead to better peak performance at the cost of learning stability.
****

Dynamic Relational Inference in Multi-Agent Trajectories
Ruichao Xiao,Manish Kumar Singh,Rose Yu
Unsupervised learning of interactions from multi-agent trajectories has broad applications in physics, vision, and robotics. However, existing neural relational inference works are limited to static relations. We consider a more general setting of dynamic relational inference where interactions change over time. We propose DYnamic multi-Agent Relational Inference (DYARI) model, a deep generative model that can reason about dynamic relations. Using a simulated physics system, we study various dynamic relation scenarios, including periodic and additive dynamics. We perform a comprehensive study on the trade-off between dynamic and inference period, the impact of the training scheme, and model architecture on dynamic relational inference accuracy. We also showcase an application of our model to infer coordination and competition patterns from real-world multi-agent basketball trajectories.
****

Intrinsically Guided Exploration in Meta Reinforcement Learning
Jin Zhang,Jianhao Wang,Hao Hu,Tong Chen,Yingfeng Chen,Changjie Fan,Chongjie Zhang
Deep reinforcement learning algorithms generally require large amounts of data to solve a single task. Meta reinforcement learning (meta-RL) agents learn to adapt to novel unseen tasks with high sample efficiency by extracting useful prior knowledge from previous tasks. Despite recent progress, efficient exploration in meta-training and adaptation remains a key challenge in sparse-reward meta-RL tasks. We propose a novel off-policy meta-RL algorithm to address this problem, which disentangles exploration and exploitation policies and learns intrinsically motivated exploration behaviors. We design novel intrinsic rewards derived from

information gain to reduce task uncertainty and encourage the explorer to collect informative trajectories about the current task. Experimental evaluation shows that our algorithm achieves state-of-the-art performance on various sparse-reward MuJoCo locomotion tasks and more complex Meta-World tasks.
****
TextSETTR: Label-Free Text Style Extraction and Tunable Targeted Restyling
Parker Riley,Noah Constant,Mandy Guo,Girish Kumar,David Uthus,Zarana Parekh
We present a novel approach to the challenging problem of label-free text style transfer. Unlike previous approaches that use parallel or non-parallel labeled data, our technique removes the need for labels entirely, relying instead on the implicit connection in style between adjacent sentences in unlabeled text. We show that T5 (Raffel et al., 2020), a strong pretrained text-to-text model, can be adapted to extract a style vector from arbitrary text and use this vector to condition the decoder to perform style transfer. As the resulting learned style vector space encodes many facets of textual style, we recast transfers as "targeted restyling" vector operations that adjust specific attributes of the input text while preserving others. When trained over unlabeled Amazon reviews data, our resulting TextSETTR model is competitive on sentiment transfer, even when given only four exemplars of each class. Furthermore, we demonstrate that a single model trained on unlabeled Common Crawl data is capable of transferring along multiple dimensions including dialect, emotiveness, formality, politeness, and sentiment.
****
Online Testing of Subgroup Treatment Effects Based on Value Difference
Miao Yu,Wenbin Lu,Rui Song
Online A/B testing plays a critical role in high-tech industry to guide product development and accelerate innovation. It performs a null hypothesis statistical test to determine which variant is better. However, a typical A/B test presents two problems: (i) a fixed-horizon framework inflates the false positive errors under continuous monitoring; (ii) the homogeneous effects assumption fails to identify a subgroup with a beneficial treatment effect. In this paper, we propose a sequential test for subgroup treatment effects based on value difference, named SUBTLE, to address these two problems simultaneously. The SUBTLE allows the experimenters to "peek" the results during the experiment without harming the statistical guarantees. It assumes heterogeneous treatment effects and aims to test if some subgroup of the population will benefit from the investigative treatment. If the testing result indicates the existence of such subgroup, a subgroup will be identified using a readily available estimated optimal treatment rule. We examine the empirical performance of our proposed test on both simulations and a real data set. The results show that the SUBTLE has high detection power with controlled type I error at any time, is more robust to noise covariates, and can achieve early stopping compared with the corresponding fixed-horizon test.
****
Neural Point Process for Forecasting Spatiotemporal Events
Zihao Zhou,Xingyi Yang,Xinyi He,Ryan Rossi,Handong Zhao,Rose Yu
Forecasting events occurring in space and time is a fundamental problem. Existing neural point process models are only temporal and are limited in spatial inference. We propose a family of deep sequence models that integrate spatiotemporal point processes with deep neural networks. Our novel Neural Spatiotemporal Point Process model is flexible, efficient, and can accurately predict irregularly sampled events. The key construction of our approach is based on space-time separation of temporal intensity function and time-conditioned spatial density function, which is approximated by kernel density estimation. We validate our model on the synthetic spatiotemporal Hawkes process and self-correcting process. On many benchmark spatiotemporal event forecasting datasets, our model demonstrates superior performances. To the best of our knowledge, this is the first neural point process model that can jointly predict the continuous space and time of events.
****
Making Coherence Out of Nothing At All: Measuring Evolution of Gradient Alignmen

t
Satrajit Chatterjee,Piotr Zielinski
We propose a new metric (m-coherence) to experimentally study the alignment of per-example gradients during training. Intuitively, given a sample of size m, m-coherence is the number of examples in the sample that benefit from a small step along the gradient of any one example on average. We show that compared to other commonly used metrics, m-coherence is more interpretable, cheaper to compute ($O(m)$ instead of $O(m^2)$) and mathematically cleaner. (We note that m-coherence is closely connected to gradient diversity, a quantity previously used in some theoretical bounds.) Using m-coherence, we study the evolution of alignment of per-example gradients in ResNet and EfficientNet models on ImageNet and several variants with label noise, particularly from the perspective of the recently proposed Coherent Gradients (CG) theory that provides a simple, unified explanation for memorization and generalization [Chatterjee, ICLR 20]. Although we have several interesting takeaways, our most surprising result concerns memorization. Naively, one might expect that when training with completely random labels, each example is fitted independently, and so m-coherence should be close to 1. However, this is not the case: m-coherence reaches moderately high values during training (though still much smaller than real labels), indicating that over-parameterized neural networks find common patterns even in scenarios where generalization is not possible. A detailed analysis of this phenomenon provides both a deeper confirmation of CG, but at the same point puts into sharp relief what is missing from the theory in order to provide a complete explanation of generalization in neural networks.
****
Stochastic Security: Adversarial Defense Using Long-Run Dynamics of Energy-Based Models
Mitch Hill,Jonathan Craig Mitchell,Song-Chun Zhu
The vulnerability of deep networks to adversarial attacks is a central problem for deep learning from the perspective of both cognition and security. The current most successful defense method is to train a classifier using adversarial images created during learning. Another defense approach involves transformation or purification of the original input to remove adversarial signals before the image is classified. We focus on defending naturally-trained classifiers using Markov Chain Monte Carlo (MCMC) sampling with an Energy-Based Model (EBM) for adversarial purification. In contrast to adversarial training, our approach is intended to secure highly vulnerable pre-existing classifiers. To our knowledge, no prior defensive transformation is capable of securing naturally-trained classifiers, and our method is the first to validate a post-training defense approach that is distinct from current successful defenses which modify classifier training.

The memoryless behavior of long-run MCMC sampling will eventually remove adversarial signals, while metastable behavior preserves consistent appearance of MCMC samples after many steps to allow accurate long-run prediction. Balancing these factors can lead to effective purification and robust classification. We evaluate adversarial defense with an EBM using the strongest known attacks against purification. Our contributions are 1) an improved method for training EBM's with realistic long-run MCMC samples for effective purification, 2) an Expectation-Over-Transformation (EOT) defense that resolves ambiguities for evaluating stochastic defenses and from which the EOT attack naturally follows, and 3) state-of-the-art adversarial defense for naturally-trained classifiers and competitive defense compared to adversarial training on CIFAR-10, SVHN, and CIFAR-100. Our code and pre-trained models are available at https://github.com/point0bar1/ebm-defense.
****
Rewriter-Evaluator Framework for Neural Machine Translation
Yangming Li,Kaisheng Yao
Encoder-decoder architecture has been widely used in neural machine translation (NMT). A few methods have been proposed to improve it with multiple passes of decoding. However, their full potential is limited by a lack of appropriate termination policy. To address this issue, we present a novel framework, Rewriter-Eval

uator. It consists of a rewriter and an evaluator. Translating a source sentence involves multiple passes. At every pass, the rewriter produces a new translation to improve the past translation and the evaluator estimates the translation quality to decide whether to terminate the rewriting process. We also propose a prioritized gradient descent (PGD) method that facilitates training the rewriter and the evaluator jointly. Though incurring multiple passes of decoding, Rewriter-Evaluator with the proposed PGD method can be trained with similar time to that of training encoder-decoder models. We apply the proposed framework to improve the general NMT models (e.g., Transformer). We conduct extensive experiments on two translation tasks, Chinese-English and English-German, and show that the proposed framework notably improves the performances of NMT models and significantly outperforms previous baselines.

****

Defective Convolutional Networks

Tiange Luo,Tianle Cai,Mengxiao Zhang,Siyu Chen,Di He,Liwei Wang

Robustness of convolutional neural networks (CNNs) has gained in importance on account of adversarial examples, i.e., inputs added as well-designed perturbations that are imperceptible to humans but can cause the model to predict incorrectly. Recent research suggests that the noise in adversarial examples breaks the textural structure, which eventually leads to wrong predictions. To mitigate the threat of such adversarial attacks, we propose defective convolutional networks that make predictions relying less on textural information but more on shape information by properly integrating defective convolutional layers into standard CNNs. The defective convolutional layers contain defective neurons whose activations are set to be a constant function. As defective neurons contain no information and are far different from standard neurons in its spatial neighborhood, the textural features cannot be accurately extracted, and so the model has to seek other features for classification, such as the shape. We show extensive evidence to justify our proposal and demonstrate that defective CNNs can defend against black-box attacks better than standard CNNs. In particular, they achieve state-of-the-art performance against transfer-based attacks without any adversarial training being applied.

****

BayesAdapter: Being Bayesian, Inexpensively and Robustly, via Bayesian Fine-tuning

Zhijie Deng,Xiao Yang,Hao Zhang,Yinpeng Dong,Jun Zhu

Despite their theoretical appealingness, Bayesian neural networks (BNNs) are falling far behind in terms of adoption in real-world applications compared with normal NNs, mainly due to their limited scalability in training, and low fidelity in their uncertainty estimates. In this work, we develop a new framework, named BayesAdapter, to address these issues and bring Bayesian deep learning to the masses. The core notion of BayesAdapter is to adapt pre-trained deterministic NNs to be BNNs via Bayesian fine-tuning. We implement Bayesian fine-tuning with a plug-and-play instantiation of stochastic variational inference, and propose exemplar reparameterization to reduce gradient variance and stabilize the fine-tuning. Together, they enable training BNNs as if one were training deterministic NNs with minimal added overheads. During Bayesian fine-tuning, we further propose an uncertainty regularization to supervise and calibrate the uncertainty quantification of learned BNNs at low cost. To empirically evaluate BayesAdapter, we conduct extensive experiments on a diverse set of challenging benchmarks, and observe satisfactory training efficiency, competitive predictive performance, and calibrated and faithful uncertainty estimates.

****

Nearest Neighbor Machine Translation

Urvashi Khandelwal,Angela Fan,Dan Jurafsky,Luke Zettlemoyer,Mike Lewis

We introduce k-nearest-neighbor machine translation (kNN-MT), which predicts tokens with a nearest-neighbor classifier over a large datastore of cached examples, using representations from a neural translation model for similarity searc

h. This approach requires no additional training and scales to give the decoder direct access to billions of examples at test time, resulting in a highly expressive model that consistently improves performance across many settings. Simply adding nearest-neighbor search improves a state-of-the-art German-English translation model by 1.5 BLEU. kNN-MT allows a single model to be adapted to diverse domains by using a domain-specific datastore, improving results by an average of 9.2 BLEU over zero-shot transfer, and achieving new state-of-the-art results---without training on these domains. A massively multilingual model can also be specialized for particular language pairs, with improvements of 3 BLEU for translating from English into German and Chinese. Qualitatively, kNN-MT is easily interpretable; it combines source and target context to retrieve highly relevant examples.
****
Weak and Strong Gradient Directions: Explaining Memorization, Generalization, and Hardness of Examples at Scale
Piotr Zielinski,Shankar Krishnan,Satrajit Chatterjee
Coherent Gradients (CGH) [Chatterjee, ICLR 20] is a recently proposed hypothesis to explain why over-parameterized neural networks trained with gradient descent generalize well even though they have sufficient capacity to memorize the training set. The key insight of CGH is that, since the overall gradient for a single step of SGD is the sum of the per-example gradients, it is strongest in directions that reduce the loss on multiple examples if such directions exist. In this paper, we validate CGH on ResNet, Inception, and VGG models on ImageNet. Since the techniques presented in the original paper do not scale beyond toy models and datasets, we propose new methods. By posing the problem of suppressing weak gradient directions as a problem of robust mean estimation, we develop a coordinate-based median of means approach. We present two versions of this algorithm, M3, which partitions a mini-batch into 3 groups and computes the median, and a more efficient version RM3, which reuses gradients from previous two time steps to compute the median. Since they suppress weak gradient directions without requiring per-example gradients, they can be used to train models at scale. Experimentally, we find that they indeed greatly reduce overfitting (and memorization) and thus provide the first convincing evidence that CGH holds at scale. We also propose a new test of CGH that does not depend on adding noise to training labels or on suppressing weak gradient directions. Using the intuition behind CGH, we posit that the examples learned early in the training process (i.e., "easy" examples) are precisely those that have more in common with other training examples. Therefore, as per CGH, the easy examples should generalize better amongst themselves than the hard examples amongst themselves. We validate this hypothesis with detailed experiments, and believe that it provides further orthogonal evidence for CGH.
****
Towards Robust and Efficient Contrastive Textual Representation Learning
Liqun Chen,Yizhe Zhang,Dianqi Li,Chenyang Tao,Dong Wang,Lawrence Carin
There has been growing interest in representation learning for text data, based on theoretical arguments and empirical evidence. One important direction involves leveraging contrastive learning to improve learned representations. We propose an application of contrastive learning for intermediate textual feature pairs, to explicitly encourage the model to learn more distinguishable representations. To overcome the learner's degeneracy due to vanishing contrasting signals, we impose Wasserstein constraints on the critic via spectral regularization.

Finally, to moderate such an objective from overly regularized training and to enhance learning efficiency, with theoretical justification, we further leverage an active negative-sample-selection procedure to only use high-quality contrast examples. We evaluate the proposed method over a wide range of natural language processing applications, from the perspectives of both supervised and unsupervised learning. Empirical results show consistent improvement over baselines.
****
Deformable Capsules for Object Detection
Rodney LaLonde,Naji Khosravan,Ulas Bagci

Capsule networks promise significant benefits over convolutional networks by sto
ring stronger internal representations, and routing information based on the agr
eement between intermediate representations' projections. Despite this, their su
ccess has been mostly limited to small-scale classification datasets due to thei
r computationally expensive nature. Recent studies have partially overcome this
burden by locally-constraining the dynamic routing of features with convolutiona
l capsules. Though memory efficient, convolutional capsules impose geometric con
straints which fundamentally limit the ability of capsules to model the pose/def
ormation of objects. Further, they do not address the bigger memory concern of c
lass-capsules scaling-up to bigger tasks such as detection or large-scale classi
fication. In this study, we introduce deformable capsules (DeformCaps), a new ca
psule structure (SplitCaps), and a novel dynamic routing algorithm (SE-Routing)
to balance computational efficiency with the need for modeling a large number of
 objects and classes. We demonstrate that the proposed methods allow capsules to
 efficiently scale-up to large-scale computer vision tasks for the first time, a
nd create the first-ever capsule network for object detection in the literature.
 Our proposed architecture is a one-stage detection framework and obtains result
s on MS COCO which are on-par with state-of-the-art one-stage CNN-based methods,
 while producing fewer false positive detections.
****

Learning Hyperbolic Representations of Topological Features

Panagiotis Kyriakis,Iordanis Fostiropoulos,Paul Bogdan

Learning task-specific representations of persistence diagrams is an important p
roblem in topological data analysis and machine learning. However, current state
 of the art methods are restricted in terms of their expressivity as they are fo
cused on Euclidean representations. Persistence diagrams often contain features
of infinite persistence (i.e., essential features) and Euclidean spaces shrink t
heir importance relative to non-essential features because they cannot assign in
finite distance to finite points. To deal with this issue, we propose a method t
o learn representations of persistence diagrams on hyperbolic spaces, more speci
fically on the Poincare ball. By representing features of infinite persistence i
nfinitesimally close to the boundary of the ball, their distance to non-essentia
l features approaches infinity, thereby their relative importance is preserved.
This is achieved without utilizing extremely high values for the learnable param
eters, thus the representation can be fed into downstream optimization methods a
nd trained efficiently in an end-to-end fashion. We present experimental results
 on graph and image classification tasks and show that the performance of our me
thod is on par with or exceeds the performance of other state of the art methods
.

****

Learning Movement Strategies for Moving Target Defense

Sailik Sengupta,Subbarao Kambhampati

The field of cybersecurity has mostly been a cat-and-mouse game with the discove
ry of new attacks leading the way. To take away an attacker's advantage of recon
naissance, researchers have proposed proactive defense methods such as Moving Ta
rget Defense (MTD). To find good movement strategies, researchers have modeled M
TD as leader-follower games between the defender and a cyber-adversary. We argue
 that existing models are inadequate in sequential settings when there is incomp
lete information about rational adversary and yield sub-optimal movement strateg
ies. Further, while there exists an array of work on learning defense policies i
n sequential settings for cyber-security, they are either unpopular due to scala
bility issues arising out of incomplete information or tend to ignore the strate
gic nature of the adversary simplifying the scenario to use single-agent reinfor
cement learning techniques. To address these concerns, we propose (1) a unifying
 game-theoretic model, called the Bayesian Stackelberg Markov Games (BSMGs), tha
t can model uncertainty over attacker types and the nuances of an MTD system and
 (2) a Bayesian Strong Stackelberg Q-learning (BSS-Q) approach that can, via int
eraction, learn the optimal movement policy for BSMGs within a reasonable time.
We situate BSMGs in the landscape of incomplete-information Markov games and cha

racterize the notion of Strong Stackelberg Equilibrium (SSE) in them. We show th at our learning approach converges to an SSE of a BSMG and then highlight that t he learned movement policy (1) improves the state-of-the-art in MTD for web-appl ication security and (2) converges to an optimal policy in MTD domains with inco mplete information about adversaries even when prior information about rewards a nd transitions is absent.
****

Latent Programmer: Discrete Latent Codes for Program Synthesis
Joey Hong,David Dohan,Rishabh Singh,Charles Sutton,Manzil Zaheer
In many sequence learning tasks, such as program synthesis and document summariz ation, a key problem is searching over a large space of possible output sequence s. We propose to learn representations of the outputs that is specifically meant for search: rich enough to specify the desired output but compact enough to mak e search more efficient. An appealing realization of such representation are dis crete latent codes, as this naturally allows sophisticated combinatorial search strategies. The latent codes are learned using a self-supervised learning princi ple, in which first a discrete autoencoder is trained on the output sequences, a nd then the resulting latent codes are used as intermediate targets for the end- to-end sequence prediction task. Based on these insights, we introduce the Laten t Programmer, a program synthesis method that first predicts a discrete latent c odes from input/output examples, and then generates the program in the target la nguage. We evaluate the Latent Programmer on two domains: synthesis of string t ransformation programs, and generation of programs from natural language descrip tions. We demonstrate that the discrete latent representation significantly impr oves synthesis accuracy.
****

A Panda? No, It's a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Networ k Inference
Sanghyun Hong,Yigitcan Kaya,Ionu■-Vlad Modoranu,Tudor Dumitras
Recent increases in the computational demands of deep neural networks (DNNs), co mbined with the observation that most input samples require only simple models, have sparked interest in input-adaptive multi-exit architectures, such as MSDNet s or Shallow-Deep Networks. These architectures enable faster inferences and cou ld bring DNNs to low-power devices, e.g., in the Internet of Things (IoT). Howev er, it is unknown if the computational savings provided by this approach are rob ust against adversarial pressure. In particular, an adversary may aim to slowdow n adaptive DNNs by increasing their average inference time—a threat analogous to the denial-of-service attacks from the Internet. In this paper, we conduct a sy stematic evaluation of this threat by experimenting with three generic multi-exi t DNNs (based on VGG16, MobileNet, and ResNet56) and a custom multi-exit archite cture, on two popular image classification benchmarks (CIFAR-10 and Tiny ImageNe t). To this end, we show that adversarial example-crafting techniques can be mod ified to cause slowdown, and we propose a metric for comparing their impact on d ifferent architectures. We show that a slowdown attack reduces the efficacy of m ulti-exit DNNs by 90–100%, and it amplifies the latency by 1.5–5× in a typical I oT deployment. We also show that it is possible to craft universal, reusable per turbations and that the attack can be effective in realistic black-box scenarios , where the attacker has limited knowledge about the victim. Finally, we show th at adversarial training provides limited protection against slowdowns. These res ults suggest that further research is needed for defending multi-exit architectu res against this emerging threat. Our code is available at https://github.com/sa nghyun-hong/deepsloth.
****

Non-Attentive Tacotron: Robust and controllable neural TTS synthesis including u nsupervised duration modeling
Jonathan Shen,Ye Jia,Mike Chrzanowski,Yu Zhang,Isaac Elias,Heiga Zen,Yonghui Wu
This paper presents Non-Attentive Tacotron based on the Tacotron 2 text-to-speec h model, replacing the attention mechanism with an explicit duration predictor. This improves robustness significantly as measured by unaligned duration ratio a nd word deletion rate, two metrics introduced in this paper for large-scale robu

stness evaluation using a pre-trained speech recognition model. With the use of Gaussian upsampling, Non-Attentive Tacotron achieves a 5-scale mean opinion score for naturalness of 4.41, slightly outperforming Tacotron 2. The duration predictor enables both utterance-wide and per-phoneme control of duration at inference time. When accurate target durations are scarce or unavailable in the training data, we propose a method using a fine-grained variational auto-encoder to train the duration predictor in a semi-supervised or unsupervised manner, with results almost as good as supervised training.
****

NCP-VAE: Variational Autoencoders with Noise Contrastive Priors
Jyoti Aneja,Alex Schwing,Jan Kautz,Arash Vahdat
Variational autoencoders (VAEs) are one of the powerful likelihood-based generative models with applications in various domains. However, they struggle to generate high-quality images, especially when samples are obtained from the prior without any tempering. One explanation for VAEs' poor generative quality is the prior hole problem: the prior distribution fails to match the aggregate approximate posterior. Due to this mismatch, there exist areas in the latent space with high density under the prior that do not correspond to any encoded image. Samples from those areas are decoded to corrupted images. To tackle this issue, we propose an energy-based prior defined by the product of a base prior distribution and a reweighting factor, designed to bring the base closer to the aggregate posterior. We train the reweighting factor by noise contrastive estimation, and we generalize it to hierarchical VAEs with many latent variable groups. Our experiments confirm that the proposed noise contrastive priors improve the generative performance of state-of-the-art VAEs by a large margin on the MNIST, CIFAR-10, CelebA 64, and CelebA HQ 256 datasets.
****

Information Condensing Active Learning
Siddhartha Jain,Ge Liu,David Gifford
We introduce Information Condensing Active Learning (ICAL), a batch mode model agnostic Active Learning (AL) method targeted at Deep Bayesian Active Learning that focuses on acquiring labels for points which have as much information as possible about the still unacquired points. ICAL uses the Hilbert Schmidt Independence Criterion (HSIC) to measure the strength of the dependency between a candidate batch of points and the unlabeled set. We develop key optimizations that allow us to scale our method to large unlabeled sets. We show significant improvements in terms of model accuracy and negative log likelihood (NLL) on several image datasets compared to state of the art batch mode AL methods for deep learning.
****

Adaptive Discretization for Continuous Control using Particle Filtering Policy Network
Pei Xu,Ioannis Karamouzas
Controlling the movements of highly articulated agents and robots has been a long-standing challenge to model-free deep reinforcement learning. In this paper, we propose a simple, yet general, framework for improving the performance of policy gradient algorithms by discretizing the continuous action space. Instead of using a fixed set of predetermined atomic actions, we exploit particle filtering to adaptively discretize actions during training and track the posterior policy represented as a mixture distribution. The resulting policy can replace the original continuous policy of any given policy gradient algorithm without changing its underlying model architecture. We demonstrate the applicability of our approach to state-of-the-art on-policy and off-policy baselines in challenging control tasks. Baselines using our particle-based policies achieve better final performance and speed of convergence as compared to corresponding continuous implementations and implementations that rely on fixed discretization schemes.
****

Scalable Graph Neural Networks for Heterogeneous Graphs
Lingfan Yu,Jiajun Shen,Jinyang Li,Adam Lerer
Graph neural networks (GNNs) are a popular class of parametric model for learning over graph-structured data. Recent work has argued that GNNs primarily use the

graph for feature smoothing, and have shown competitive results on benchmark tasks by simply operating on graph-smoothed node features, rather than using end-to-end learned feature hierarchies that are challenging to scale to large graphs. In this work, we ask whether these results can be extended to heterogeneous graphs, which encode multiple types of relationship between different entities. We propose Neighbor Averaging over Relation Subgraphs (NARS), which trains a classifier on neighbor-averaged features for randomly-sampled subgraphs of the ‘metagraph‘ of relations. We describe optimizations to allow these sets of node features to be computed in a memory-efficient way, both at training and inference time. NARS achieves a new state of the art accuracy on several benchmark datasets, outperforming more expensive GNN-based methods.
****
Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs
Matthew L Leavitt,Ari S. Morcos
The properties of individual neurons are often analyzed in order to understand the biological and artificial neural networks in which they're embedded. Class selectivity—typically defined as how different a neuron's responses are across different classes of stimuli or data samples—is commonly used for this purpose. However, it remains an open question whether it is necessary and/or sufficient for deep neural networks (DNNs) to learn class selectivity in individual units. We investigated the causal impact of class selectivity on network function by directly regularizing for or against class selectivity. Using this regularizer to reduce class selectivity across units in convolutional neural networks increased test accuracy by over 2% in ResNet18 and 1% in ResNet50 trained on Tiny ImageNet. For ResNet20 trained on CIFAR10 we could reduce class selectivity by a factor of 2.5 with no impact on test accuracy, and reduce it nearly to zero with only a small (~2%) drop in test accuracy. In contrast, regularizing to increase class selectivity significantly decreased test accuracy across all models and datasets. These results indicate that class selectivity in individual units is neither sufficient nor strictly necessary, and can even impair DNN performance. They also encourage caution when focusing on the properties of single units as representative of the mechanisms by which DNNs function.
****
Integrating Categorical Semantics into Unsupervised Domain Translation
Samuel Lavoie-Marchildon,Faruk Ahmed,Aaron Courville
While unsupervised domain translation (UDT) has seen a lot of success recently, we argue that mediating its translation via categorical semantic features could broaden its applicability. In particular, we demonstrate that categorical semantics improves the translation between perceptually different domains sharing multiple object categories. We propose a method to learn, in an unsupervised manner, categorical semantic features (such as object labels) that are invariant of the source and target domains. We show that conditioning the style encoder of unsupervised domain translation methods on the learned categorical semantics leads to a translation preserving the digits on MNIST$\leftrightarrow$SVHN and to a more realistic stylization on Sketches$\to$Reals.
****
Are Neural Rankers still Outperformed by Gradient Boosted Decision Trees?
Zhen Qin,Le Yan,Honglei Zhuang,Yi Tay,Rama Kumar Pasumarthi,Xuanhui Wang,Michael Bendersky,Marc Najork
Despite the success of neural models on many major machine learning problems, their effectiveness on traditional Learning-to-Rank (LTR) problems is still not widely acknowledged. We first validate this concern by showing that most recent neural LTR models are, by a large margin, inferior to the best publicly available Gradient Boosted Decision Trees (GBDT) in terms of their reported ranking accuracy on benchmark datasets. This unfortunately was somehow overlooked in recent neural LTR papers. We then investigate why existing neural LTR models under-perform and identify several of their weaknesses. Furthermore, we propose a unified framework comprising of counter strategies to ameliorate the existing weaknesses of neural models. Our models are the first to be able to perform equally well, co

mparing with the best tree-based baseline, while outperforming recently publishe d neural LTR models by a large margin. Our results can also serve as a benchmark to facilitate future improvement of neural LTR models.
****

Zero-shot Synthesis with Group-Supervised Learning
Yunhao Ge,Sami Abu-El-Haija,Gan Xin,Laurent Itti
Visual cognition of primates is superior to that of artificial neural networks i n its ability to “envision” a visual object, even a newly-introduced one, in dif ferent attributes including pose, position, color, texture, etc. To aid neural networks to envision objects with different attributes, we propose a family of objective functions, expressed on groups of examples, as a novel learning framew ork that we term Group-Supervised Learning (GSL). GSL allows us to decompose inp uts into a disentangled representation with swappable components, that can be re combined to synthesize new samples. For instance, images of red boats & blue ca rs can be decomposed and recombined to synthesize novel images of red cars. We propose an implementation based on auto-encoder, termed group-supervised zero-s hot synthesis network (GZS-Net) trained with our learning framework, that can pr oduce a high-quality red car even if no such example is witnessed during trainin g. We test our model and learning framework on existing benchmarks, in addition to a new dataset that we open-source. We qualitatively and quantitatively demons trate that GZS-Net trained with GSL outperforms state-of-the-art methods
****

Learning Algebraic Representation for Abstract Spatial-Temporal Reasoning
Chi Zhang,Sirui Xie,Baoxiong Jia,Yixin Zhu,Ying Nian Wu,Song-Chun Zhu
Is intelligence realized by connectionist or classicist? While connectionist app roaches have achieved superhuman performance, there has been growing evidence th at such task-specific superiority is particularly fragile in systematic generali zation. This observation lies in the central debate (Fodor et al., 1988; Fodor & McLaughlin, 1990) between connectionist and classicist, wherein the latter conti nually advocates an algebraic treatment in cognitive architectures. In this work , we follow the classicist's call and propose a hybrid approach to improve syste matic generalization in reasoning. Specifically, we showcase a prototype with al gebraic representations for the abstract spatial-temporal reasoning task of Rave n’s Progressive Matrices (RPM) and present the ALgebra-Aware Neuro-Semi-Symbolic (ALANS^2) learner. The ALANS^2 learner is motivated by abstract algebra and the representation theory. It consists of a neural visual perception frontend a nd an algebraic abstract reasoning backend: the frontend summarizes the visual i nformation from object-based representations, while the backend transforms it in to an algebraic structure and induces the hidden operator on-the-fly. The induce d operator is later executed to predict the answer's representation, and the cho ice most similar to the prediction is selected as the solution. Extensive experi ments show that by incorporating an algebraic treatment, the ALANS^2 learner o utperforms various pure connectionist models in domains requiring systematic gen eralization. We further show that the algebraic representation learned can be de coded by isomorphism and used to generate an answer.
****

Learning Generalizable Visual Representations via Interactive Gameplay
Luca Weihs,Aniruddha Kembhavi,Kiana Ehsani,Sarah M Pratt,Winson Han,Alvaro Herra sti,Eric Kolve,Dustin Schwenk,Roozbeh Mottaghi,Ali Farhadi
A growing body of research suggests that embodied gameplay, prevalent not just i n human cultures but across a variety of animal species including turtles and ra vens, is critical in developing the neural flexibility for creative problem solv ing, decision making, and socialization. Comparatively little is known regarding the impact of embodied gameplay upon artificial agents. While recent work has p roduced agents proficient in abstract games, these environments are far removed the real world and thus these agents can provide little insight into the advanta ges of embodied play. Hiding games, such as hide-and-seek, played universally, p rovide a rich ground for studying the impact of embodied gameplay on representat ion learning in the context of perspective taking, secret keeping, and false bel ief understanding. Here we are the first to show that embodied adversarial reinf

orcement learning agents playing Cache, a variant of hide-and-seek, in a high fi
delity, interactive, environment, learn generalizable representations of their o
bservations encoding information such as object permanence, free space, and cont
ainment. Moving closer to biologically motivated learning strategies, our agents
' representations, enhanced by intentionality and memory, are developed through
interaction and play. These results serve as a model for studying how facets of
vision develop through interaction, provide an experimental framework for assess
ing what is learned by artificial agents, and demonstrates the value of moving f
rom large, static, datasets towards experiential, interactive, representation le
arning.
****

VA-RED^2: Video Adaptive Redundancy Reduction
Bowen Pan,Rameswar Panda,Camilo Luciano Fosco,Chung-Ching Lin,Alex J Andonian,Yu
e Meng,Kate Saenko,Aude Oliva,Rogerio Feris
Performing inference on deep learning models for videos remains a challenge due
to the large amount of computational resources required to achieve robust recogn
ition. An inherent property of real-world videos is the high correlation of info
rmation across frames which can translate into redundancy in either temporal or
spatial feature maps of the models, or both. The type of redundant features depe
nds on the dynamics and type of events in the video: static videos have more tem
poral redundancy while videos focusing on objects tend to have more channel redu
ndancy. Here we present a redundancy reduction framework, termed VA-RED^2, whi
ch is input-dependent. Specifically, our VA-RED^2 framework uses an input-depe
ndent policy to decide how many features need to be computed for temporal and ch
annel dimensions. To keep the capacity of the original model, after fully comput
ing the necessary features, we reconstruct the remaining redundant features from
 those using cheap linear operations. We learn the adaptive policy jointly with
the network weights in a differentiable way with a shared-weight mechanism, maki
ng it highly efficient. Extensive experiments on multiple video datasets and dif
ferent visual tasks show that our framework achieves $20\% - 40\%$ reduction in
computation (FLOPs) when compared to state-of-the-art methods without any perfor
mance loss. Project page: http://people.csail.mit.edu/bpan/va-red/.
****

BERTology Meets Biology: Interpreting Attention in Protein Language Models
Jesse Vig,Ali Madani,Lav R. Varshney,Caiming Xiong,richard socher,Nazneen Rajani
Transformer architectures have proven to learn useful representations for protei
n classification and generation tasks. However, these representations present ch
allenges in interpretability. In this work, we demonstrate a set of methods for
analyzing protein Transformer models through the lens of attention. We show that
 attention: (1) captures the folding structure of proteins, connecting amino aci
ds that are far apart in the underlying sequence, but spatially close in the thr
ee-dimensional structure, (2) targets binding sites, a key functional component
of proteins, and (3) focuses on progressively more complex biophysical propertie
s with increasing layer depth. We find this behavior to be consistent across thr
ee Transformer architectures (BERT, ALBERT, XLNet) and two distinct protein data
sets. We also present a three-dimensional visualization of the interaction betwe
en attention and protein structure. Code for visualization and analysis is avail
able at https://github.com/salesforce/provis.
****

Physics Informed Deep Kernel Learning
Zheng Wang,Wei Xing,Robert Kirby,Shandian Zhe
Deep kernel learning is a promising combination of deep neural networks and nonp
arametric function estimation. However, as a data-driven approach, the performan
ce of deep kernel learning can still be restricted by scarce or insufficient dat
a, especially in extrapolation tasks. To address these limitations, we propose
Physics Informed Deep Kernel Learning (PI-DKL) that exploits physics knowledge
represented by differential equations with latent sources. Specifically, we use
the posterior function sample of the Gaussian process as the surrogate for the
solution of the differential equation, and construct a generative component to
integrate the equation in a principled Bayesian hybrid framework. For efficient

and effective inference, we marginalize out the latent variables in the joint pr
obability and derive a simple model evidence lower bound (ELBO), based on which
we develop a stochastic collapsed inference algorithm. Our ELBO can be viewed as
 a nice, interpretable posterior regularization objective. On synthetic datasets
 and real-world applications, we show the advantage of our approach in both pred
iction accuracy and uncertainty quantification.
****

STRATA: Simple, Gradient-free Attacks for Models of Code
Jacob M. Springer,Bryn Marie Reinstadler,Una-May O'Reilly
Adversarial examples are imperceptible perturbations in the input to a neural mo
del that result in misclassification. Generating adversarial examples for source
 code poses an additional challenge compared to the domains of images and natura
l language, because source code perturbations must adhere to strict semantic gui
delines so the resulting programs retain the functional meaning of the code. We
propose a simple and efficient gradient-free method for generating state-of-the-
art adversarial examples on models of code that can be applied in a white-box or
 black-box setting. Our method generates untargeted and targeted attacks, and em
pirically outperforms competing gradient-based methods with less information and
 less computational effort.
****

Average Reward Reinforcement Learning with Monotonic Policy Improvement
Yiming Zhang,Keith W. Ross
In continuing control tasks, an agent’s average reward per time step is a more n
atural performance measure compared to the commonly used discounting framework s
ince it can better capture an agent’s long-term behavior. We derive a novel lowe
r bound on the difference of the long-term average reward for two policies. The
 lower bound depends on the average divergence between the policies and on the s
o-called Kemeny constant, which measures to what degree the unichain Markov chai
ns associated with the policies are well-connected. We also show that previous w
ork based on the discounted return (Schulman et al., 2015; Achiam et al., 2017)
results in a non-meaningful lower bound in the average reward setting. Based on
our lower bound, we develop an iterative procedure which produces a sequence of
monotonically improved policies for the average reward criterion. When combined
with Deep Reinforcement Learning (DRL) methods, the procedure leads to scalable
and efficient algorithms for maximizing the agent’s average reward performance.E
mpirically we demonstrate the effectiveness of our method on continuing control
tasks and show how discounting can lead to unsatisfactory performance.
****

Slot Machines: Discovering Winning Combinations of Random Weights in Neural Netw
orks
Maxwell Mbabilla Aladago,Lorenzo Torresani
In contrast to traditional weight optimization in a continuous space, we demonst
rate the existence of effective random networks whose weights are never updated.
 By selecting a weight among a fixed set of random values for each individual co
nnection, our method uncovers combinations of random weights that match the perf
ormance of trained networks of the same capacity. We refer to our networks as ``
slot machines'' where each reel (connection) contains a fixed set of symbols (ra
ndom values). Our backpropagation algorithm ``spins'' the reels to seek ``winnin
g'' combinations, i.e., selections of random weight values that minimize the gi
ven loss. Quite surprisingly, we find that allocating just a few random values t
o each connection (e.g., 8 values per connection) yields highly competitive comb
inations despite being dramatically more constrained compared to traditionally l
earned weights. Moreover, finetuning these combinations often improves performan
ce over the trained baselines. A randomly initialized VGG-19 with 8 values per
connection contains a combination that achieves 90% test accuracy on CIFAR-10.
Our method also achieves an impressive performance of 98.1% on MNIST for neural
networks containing only random weights.
****

Trajectory Prediction using Equivariant Continuous Convolution
Robin Walters,Jinxi Li,Rose Yu

Trajectory prediction is a critical part of many AI applications, for example, t
he safe operation of autonomous vehicles. However, current methods are prone to
making inconsistent and physically unrealistic predictions. We leverage insights
 from fluid dynamics to overcome this limitation by considering internal symmet
ry in real-world trajectories. We propose a novel model, Equivariant Continous C
Onvolution (ECCO) for improved trajectory prediction. ECCO uses rotationally-eq
uivariant continuous convolutions to embed the symmetries of the system. On both
 vehicle and pedestrian trajectory datasets, ECCO attains competitive accuracy
with significantly fewer parameters. It is also more sample efficient, generaliz
ing automatically from few data points in any orientation. Lastly, ECCO improve
s generalization with equivariance, resulting in more physically consistent pred
ictions. Our method provides a fresh perspective towards increasing trust and
transparency in deep learning models. Our code and data can be found at https://
github.com/Rose-STL-Lab/ECCO.
****

ERMAS: Learning Policies Robust to Reality Gaps in Multi-Agent Simulations
Eric Zhao,Alexander R Trott,Caiming Xiong,Stephan Zheng
Policies for real-world multi-agent problems, such as optimal taxation, can be l
earned in multi-agent simulations with AI agents that emulate humans. However, s
imulations can suffer from reality gaps as humans often act suboptimally or opti
mize for different objectives (i.e., bounded rationality). We introduce $\epsilo
n$-Robust Multi-Agent Simulation (ERMAS), a robust optimization framework to lea
rn AI policies that are robust to such multi-agent reality gaps. The objective o
f ERMAS theoretically guarantees robustness to the ϵ-Nash equilibria of
 other agents – that is, robustness to behavioral deviations with a regret of at
 most ϵ. ERMAS efficiently solves a first-order approximation of the ro
bustness objective using meta-learning methods. We show that ERMAS yields robust
 policies for repeated bimatrix games and optimal adaptive taxation in economic
simulations, even when baseline notions of robustness are uninformative or intra
ctable. In particular, we show ERMAS can learn tax policies that are robust to c
hanges in agent risk aversion, improving policy objectives (social welfare) by u
p to 15% in complex spatiotemporal simulations using the AI Economist (Zheng et
al., 2020).
****

Open-world Semi-supervised Learning
Kaidi Cao,Maria Brbic,Jure Leskovec
Supervised and semi-supervised learning methods have been traditionally designed
 for the closed-world setting which is based on the assumption that unlabeled te
st data contains only classes previously encountered in the labeled training dat
a. However, the real world is often open and dynamic, and thus novel previously
unseen classes may appear in the test data or during the model deployment. Here,
 we introduce a new open-world semi-supervised learning setting in which the mod
el is required to recognize previously seen classes, as well as to discover nove
l classes never seen in the labeled dataset. To tackle the problem, we propose O
RCA, an approach that jointly learns a feature representation and a classifier o
n the labeled and unlabeled subsets of the data. The key idea in ORCA is in intr
oducing uncertainty based adaptive margin that effectively circumvents the bias
caused by the imbalance of variance between seen and novel classes. We demonstra
te that ORCA accurately discovers novel classes and assigns samples to previousl
y seen classes on standard benchmark image classification datasets, including C
IFAR and ImageNet. Remarkably, despite solving the harder task ORCA outperforms
semi-supervised methods on seen classes, as well as novel class discovery method
s on unseen classes, achieving 7% and 151% improvements on seen and unseen class
es of the ImageNet dataset.
****

Target Training: Tricking Adversarial Attacks to Fail
Blerta Lindqvist
Recent adversarial defense approaches have failed. Untargeted gradient-based att
acks cause classifiers to choose any wrong class. Our novel white-box defense tr
icks untargeted attacks into becoming attacks targeted at designated target clas

ses. From these target classes, we derive the real classes. The Target Training defense tricks the minimization at the core of untargeted, gradient-based adversarial attacks: minimize the sum of (1) perturbation and (2) classifier adversarial loss. Target Training changes the classifier minimally, and trains it with additional duplicated points (at 0 distance) labeled with designated classes. These differently-labeled duplicated samples minimize both terms (1) and (2) of the minimization, steering attack convergence to samples of designated classes, from which correct classification is derived. Importantly, Target Training eliminates the need to know the attack and the overhead of generating adversarial samples of attacks that minimize perturbations. Without using adversarial samples and against an adaptive attack aware of our defense, Target Training exceeds even default, unsecured classifier accuracy of 84.3% for CIFAR10 with 86.6% against DeepFool attack; and achieves 83.2% against CW-L_2 (κ=0) attack. Using adversarial samples, we achieve 75.6% against CW-L_2 (κ=40). Due to our deliberate choice of low-capacity classifiers, Target Training does not withstand L_∞ adaptive attacks in CIFAR10 but withstands CW-L_∞ (κ=0) in MNIST. Target Training presents a fundamental change in adversarial defense strategy.
****

FactoredRL: Leveraging Factored Graphs for Deep Reinforcement Learning
Bharathan Balaji,Petros Christodoulou,Xiaoyu Lu,Byungsoo Jeon,Jordan Bell-Masterson

We propose a simple class of deep reinforcement learning (RL) methods, called FactoredRL, that can leverage factored environment structures to improve the sample efficiency of existing model-based and model-free RL algorithms. In tabular and linear approximation settings, the factored Markov decision process literature has shown exponential improvements in sample efficiency by leveraging factored environment structures. We extend this to deep RL algorithms that use neural networks. For model-based algorithms, we use the factored structure to inform the state transition network architecture and for model-free algorithms we use the factored structure to inform the Q network or the policy network architecture. We demonstrate that doing this significantly improves sample efficiency in both discrete and continuous state-action space settings.
****

Scheduled Restart Momentum for Accelerated Stochastic Gradient Descent
Bao Wang,Tan Minh Nguyen,Tao Sun,Andrea Bertozzi,Richard Baraniuk,Stanley Osher

Stochastic gradient descent (SGD) algorithms, with constant momentum and its variants such as Adam, are the optimization methods of choice for training deep neural networks (DNNs). There is great interest in speeding up the convergence of these methods due to their high computational expense. Nesterov accelerated gradient (NAG) with a time-varying momentum, denoted as NAG below, improves the convergence rate of gradient descent (GD) for convex optimization using a specially designed momentum; however, it accumulates error when an inexact gradient is used (such as in SGD), slowing convergence at best and diverging at worst. In this paper, we propose scheduled restart SGD (SRSGD), a new NAG-style scheme for training DNNs. SRSGD replaces the constant momentum in SGD by the increasing momentum in NAG but stabilizes the iterations by resetting the momentum to zero according to a schedule. Using a variety of models and benchmarks for image classification, we demonstrate that, in training DNNs, SRSGD significantly improves convergence and generalization; for instance, in training ResNet-200 for ImageNet classification, SRSGD achieves an error rate of 20.93% vs. the benchmark of 22.13%. These improvements become more significant as the network grows deeper. Furthermore, on both CIFAR and ImageNet, SRSGD reaches similar or even better error rates with significantly fewer training epochs compared to the SGD baseline.
****

Learning Collision-free Latent Space for Bayesian Optimization
Fengxue Zhang,Yair Altas,Louise Fan,Kaustubh Vinchure,Brian Nord,Yuxin Chen

Learning and optimizing a blackbox function is a common task in Bayesian optimization and experimental design. In real-world scenarios (e.g., tuning hyper-parameters for deep learning models, synthesizing a protein sequence, etc.), these functions tend to be expensive to evaluate and often rely on high-dimensional inpu

ts. While classical Bayesian optimization algorithms struggle in handling the sc
ale and complexity of modern experimental design tasks, recent works attempt to
get around this issue by applying neural networks ahead of the Gaussian process
to learn a (low-dimensional) latent representation. We show that such learned re
presentation often leads to collisions in the latent space: two points with sign
ificantly different observations collide in the learned latent space. Collisions
 could be regarded as additional noise introduced by the traditional neural netw
ork, leading to degraded optimization performance. To address this issue, we pro
pose Collision-Free Latent Space Optimization (CoFLO), which employs a novel reg
ularizer to reduce the collision in the learned latent space and encourage the m
apping from the latent space to objective value to be Lipschitz continuous. CoFL
O takes in pairs of data points and penalizes those too close in the latent spac
e compared to their target space distance. We provide a rigorous theoretical jus
tification for the regularizer by inspecting the regret of the proposed algorith
m. Our empirical results further demonstrate the effectiveness of CoFLO on sever
al synthetic and real-world Bayesian optimization tasks, including a case study
for computational cosmic experimental design.
****

Towards Robustness against Unsuspicious Adversarial Examples
Liang Tong,Minzhe Guo,Atul Prakash,Yevgeniy Vorobeychik
Despite the remarkable success of deep neural networks, significant concerns hav
e emerged about their robustness to adversarial perturbations to inputs. While m
ost attacks aim to ensure that these are imperceptible, physical perturbation at
tacks typically aim for being unsuspicious, even if perceptible. However, there
is no universal notion of what it means for adversarial examples to be unsuspici
ous. We propose an approach for modeling suspiciousness by leveraging cognitive
salience. Specifically, we split an image into foreground (salient region) and b
ackground (the rest), and allow significantly larger adversarial perturbations i
n the background, while ensuring that cognitive salience of background remains l
ow. We describe how to compute the resulting non-salience-preserving dual-pertur
bation attacks on classifiers. We then experimentally demonstrate that our attac
ks indeed do not significantly change perceptual salience of the background, but
 are highly effective against classifiers robust to conventional attacks. Furthe
rmore, we show that adversarial training with dual-perturbation attacks yields c
lassifiers that are more robust to these than state-of-the-art robust learning a
pproaches, and comparable in terms of robustness to conventional attacks.
****

Unsupervised Meta-Learning through Latent-Space Interpolation in Generative Mode
ls
Siavash Khodadadeh,Sharare Zehtabian,Saeed Vahidian,Weijia Wang,Bill Lin,Ladisla
u Boloni
Several recently proposed unsupervised meta-learning approaches rely on syntheti
c meta-tasks created using techniques such as random selection, clustering and/o
r augmentation. In this work, we describe a novel approach that generates meta-t
asks using generative models. The proposed family of algorithms generate pairs o
f in-class and out-of-class samples from the latent space in a principled way, a
llowing us to create synthetic classes forming the training and validation data
of a meta-task. We find that the proposed approach, LAtent Space Interpolation U
nsupervised Meta-learning (LASIUM), outperforms or is competitive with current u
nsupervised learning baselines on few-shot classification tasks on the most wide
ly used benchmark datasets.
****

Double Generative Adversarial Networks for Conditional Independence Testing
Chengchun Shi,Tianlin Xu,Wicher Pieter Bergsma,Lexin Li
In this article, we consider the problem of high-dimensional conditional indepen
dence testing, which is a key building block in statistics and machine learning.
 We propose a double generative adversarial networks (GAN)-based inference proce
dure. We first introduce a double GANs framework to learn two generators, and in
tegrate the two generators to construct a doubly-robust test statistic. We next
consider multiple generalized covariance measures, and take their maximum as our

test statistic. Finally, we obtain the empirical distribution of our test statistic through multiplier bootstrap. We show that our test controls type-I error, while the
power approaches one asymptotically. More importantly, these theoretical guarantees are obtained under much weaker and practically more feasible conditions compared to existing tests. We demonstrate the efficacy of our test through both synthetic and real datasets.
****

Analyzing and Improving Generative Adversarial Training for Generative Modeling and Out-of-Distribution Detection
Xuwang Yin,Shiying li,Gustavo Rohde
Generative adversarial training (GAT) is a recently introduced adversarial defense method. Previous works have focused on empirical evaluations of its application to training robust predictive models. In this paper we focus on theoretical understanding of the GAT method and extending its application to generative modeling and out-of-distribution detection. We analyze the optimal solutions of the maximin formulation employed by the GAT objective, and make a comparative analysis of the minimax formulation employed by GANs. We use theoretical analysis and 2D simulations to understand the convergence property of the training algorithm. Based on these results, we develop an unconstrained GAT algorithm, and conduct comprehensive evaluations of the algorithm's application to image generation and adversarial out-of-distribution detection. Our results suggest that generative adversarial training is a promising new direction for the above applications.
****

Heteroskedastic and Imbalanced Deep Learning with Adaptive Regularization
Kaidi Cao,Yining Chen,Junwei Lu,Nikos Arechiga,Adrien Gaidon,Tengyu Ma
Real-world large-scale datasets are heteroskedastic and imbalanced --- labels have varying levels of uncertainty and label distributions are long-tailed. Heteroskedasticity and imbalance challenge deep learning algorithms due to the difficulty of distinguishing among mislabeled, ambiguous, and rare examples. Addressing heteroskedasticity and imbalance simultaneously is under-explored. We propose a data-dependent regularization technique for heteroskedastic datasets that regularizes different regions of the input space differently. Inspired by the theoretical derivation of the optimal regularization strength in a one-dimensional nonparametric classification setting, our approach adaptively regularizes the data points in higher-uncertainty, lower-density regions more heavily. We test our method on several benchmark tasks, including a real-world heteroskedastic and imbalanced dataset, WebVision. Our experiments corroborate our theory and demonstrate a significant improvement over other methods in noise-robust deep learning.
****

Towards Understanding Fast Adversarial Training
Bai Li,Shiqi Wang,Suman Jana,Lawrence Carin
Current neural-network-based classifiers are susceptible to adversarial examples. The most empirically successful approach to defending against such adversarial examples is adversarial training, which incorporates a strong self-attack during training to enhance its robustness. This approach, however, is computationally expensive and hence is hard to scale up. A recent work, called fast adversarial training, has shown that it is possible to markedly reduce computation time without sacrificing significant performance. This approach incorporates simple self-attacks, yet it can only run for a limited number of training epochs, resulting in sub-optimal performance. In this paper, we conduct experiments to understand the behavior of fast adversarial training and show the key to its success is the ability to recover from overfitting to weak attacks. We then extend our findings to improve fast adversarial training, demonstrating superior robust accuracy to strong adversarial training, with much-reduced training time.
****

Distribution Embedding Network for Meta-Learning with Variable-Length Input
Lang Liu,Mahdi Milani Fard,Sen Zhao
We propose Distribution Embedding Network (DEN) for meta-learning, which is designed for applications where both the distribution and the number of features cou

ld vary across tasks. DEN first transforms features using a learned piecewise li
near function, then learns an embedding of the underlying data distribution afte
r the transformation, and finally classifies examples based on the distribution
embedding. We show that the parameters of the distribution embedding and the cla
ssification modules can be shared across tasks. We propose a novel methodology t
o mass-simulate binary classification training tasks, and demonstrate that DEN o
utperforms existing methods in a number of test tasks in numerical studies.
****
Deep Partial Updating
Zhongnan Qu,Cong Liu,Junfeng Guo,Lothar Thiele
Emerging edge intelligence applications require the server to continuously retra
in and update deep neural networks deployed on remote edge nodes to leverage new
ly collected data samples. Unfortunately, it may be impossible in practice to co
ntinuously send fully updated weights to these edge nodes due to the highly cons
trained communication resource. In this paper, we propose the weight-wise deep p
artial updating paradigm, which smartly selects only a subset of weights to upda
te at each server-to-edge communication round, while achieving a similar perform
ance compared to full updating. Our method is established through analytically u
pper-bounding the loss difference between partial updating and full updating, an
d only updates the weights which make the largest contributions to the upper bou
nd. Extensive experimental results demonstrate the efficacy of our partial updat
ing methodology which achieves a high inference accuracy while updating a rather
small number of weights.
****
Clearing the Path for Truly Semantic Representation Learning
Dominik Zietlow,Michal Rolinek,Georg Martius
The performance of β-Variational-Autoencoders (β-VAEs) and their var
iants on learning semantically meaningful, disentangled representations is unpar
alleled. On the other hand, there are theoretical arguments suggesting impossibi
lity of unsupervised disentanglement. In this work, we show that small perturbat
ions of existing datasets hide the convenient correlation structure that is easi
ly exploited by VAE-based architectures. To demonstrate this, we construct modif
ied versions of the standard datasets on which (i) the generative factors are pe
rfectly preserved; (ii) each image undergoes a transformation barely visible to
the human eye; (iii) the leading disentanglement architectures fail to produce d
isentangled representations. We intend for these datasets to play a role in sepa
rating correlation-based models from those that discover the true causal structu
re.

The construction of the modifications is non-trivial and relies on recent progre
ss on mechanistic understanding of β-VAEs and their connection to PCA, whi
le also providing additional insights that might be of stand-alone interest.
****
Neighbourhood Distillation: On the benefits of non end-to-end distillation
Laëtitia Shao,Elad Eban,Yair Movshovitz-Attias
End-to-end training with back propagation is the standard method for training de
ep neural networks. However, as networks become deeper and bigger, end-to-end tr
aining becomes more challenging: highly non-convex models gets stuck easily in l
ocal optima, gradients signals are prone to vanish or explode during backpropaga
tion, training requires computational resources and time.

In this work, we propose to break away from the end-to-end paradigm in the conte
xt of Knowledge Distillation. Instead of distilling a model end-to-end, we propo
se to split it into smaller sub-networks - also called neighbourhoods - that are
then trained independently. We empirically show that distilling networks in a n
on end-to-end fashion can be beneficial in a diverse range of use cases. First,
we show that it speeds up Knowledge Distillation by exploiting parallelism and t
raining on smaller networks. Second, we show that independently distilled neighb
ourhoods may be efficiently re-used for Neural Architecture Search. Finally, bec
ause smaller networks model simpler functions, we show that they are easier to t

rain with synthetic data than their deeper counterparts.
****
Graph Permutation Selection for Decoding of Error Correction Codes using Self-Attention
Nir Raviv,Avi Caciularu,Tomer Raviv,Jacob Goldberger,Yair Be'ery
Error correction codes are an integral part of communication applications and boost the reliability of transmission. The optimal decoding of transmitted codewords is the maximum likelihood rule, which is NP-hard. For practical realizations, suboptimal decoding algorithms are employed; however, the lack of theoretical insights currently impedes the exploitation of the full potential of these algorithms. One key insight is the choice of permutation in permutation decoding. We present a data-driven framework for permutation selection combining domain knowledge with machine learning concepts such as node embedding and self-attention. Significant and consistent improvements in the bit error rate are shown for the simulated Bose Chaudhuri Hocquenghem (BCH) code as compared to the baseline decoders. To the best of our knowledge, this work is the first to leverage the benefits of self-attention networks in physical layer communication systems.

****
Improved Denoising Diffusion Probabilistic Models
Alexander Quinn Nichol,Prafulla Dhariwal
We explore denoising diffusion probabilistic models, a class of generative models which have recently been shown to produce excellent samples in the image and audio domains. While these models produce excellent samples, it has yet to be shown that they can achieve competitive log-likelihoods. We show that, with several small modifications, diffusion models can achieve competitive log-likelihoods in the image domain while maintaining high sample quality. Additionally, our models allow for sampling with an order of magnitude fewer diffusion steps with only a modest difference in sample quality. Finally, we explore how sample quality and log-likelihood scale with the number of diffusion steps and the amount of model capacity. We conclude that denoising diffusion probabilistic models are a promising class of generative models with excellent scaling properties and sample quality.
****
Implicit Under-Parameterization Inhibits Data-Efficient Deep Reinforcement Learning
Aviral Kumar,Rishabh Agarwal,Dibya Ghosh,Sergey Levine
We identify an implicit under-parameterization phenomenon in value-based deep RL methods that use bootstrapping: when value functions, approximated using deep neural networks, are trained with gradient descent using iterated regression onto target values generated by previous instances of the value network, more gradient updates decrease the expressivity of the current value network. We char- acterize this loss of expressivity via a drop in the rank of the learned value net- work features, and show that this typically corresponds to a performance drop. We demonstrate this phenomenon on Atari and Gym benchmarks, in both offline and online RL settings. We formally analyze this phenomenon and show that it results from a pathological interaction between bootstrapping and gradient-based optimization. We further show that mitigating implicit under-parameterization by controlling rank collapse can improve performance.
****
Failure Modes of Variational Autoencoders and Their Effects on Downstream Tasks
Yaniv Yacoby,Weiwei Pan,Finale Doshi-Velez
Variational Auto-encoders (VAEs) are deep generative latent variable models that are widely used for a number of downstream tasks. While it has been demonstrated that VAE training can suffer from a number of pathologies, existing literature lacks characterizations of exactly when these pathologies occur and how they impact down-stream task performance. In this paper we concretely characterize conditions under which VAE training exhibits pathologies and connect these failure modes to undesirable effects on specific downstream tasks, such as learning compressed and disentangled representations, adversarial robustness and semi-supervis

ed learning.
****
On Batch-size Selection for Stochastic Training for Graph Neural Networks
Yaochen Hu,Amit Levi,Ishaan Kumar,Yingxue Zhang,Mark Coates
Batch size is an important hyper-parameter for training deep learning models wit h stochastic gradient decent (SGD) method, and it has great influence on the tra ining time and model performance. We study the batch size selection problem for training graph neural network (GNN) with SGD method.
To reduce the training time while keeping a decent model performance, we propose a metric that combining both the variance of gradients and compute time for eac h mini-batch. We theoretically analyze how batch-size influence such a metric an d propose the formula to evaluate some rough range of optimal batch size.
In GNN, gradients evaluated on samples in a mini-batch are not independent and i t is challenging to evaluate the exact variance of gradients. To address the dep endency, we analyze an estimator for gradients that considers the randomness ari sing from two consecutive layers in GNN, and suggest a guideline for picking the appropriate scale of the batch size.
We complement our theoretical results with extensive empirical experiments for C lusterGCN, FastGCN and GraphSAINT on 4 datasets: Ogbn-products, Ogbn-arxiv, Redd it and Pubmed. We demonstrate that in contrast to conventional deep learning mod els, GNNs benefit from large batch sizes.
****
Measuring and Harnessing Transference in Multi-Task Learning
Chris Fifty,Ehsan Amid,Zhe Zhao,Tianhe Yu,Rohan Anil,Chelsea Finn
Multi-task learning can leverage information learned by one task to benefit the training of other tasks. Despite this capacity, naive formulations often degrade performance and in particular, identifying the tasks that would benefit from co -training remains a challenging design question. In this paper, we analyze the d ynamics of information transfer, or transference, across tasks throughout traini ng. Specifically, we develop a similarity measure that can quantify transference among tasks and use this quantity to both better understand the optimization dy namics of multi-task learning as well as improve overall learning performance. I n the latter case, we propose two methods to leverage our transference metric. T he first operates at a macro-level by selecting which tasks should train togethe r while the second functions at a micro-level by determining how to combine task gradients at each training step. We find these methods can lead to significant improvement over prior work on three supervised multi-task learning benchmarks a nd one multi-task reinforcement learning paradigm.
****
Parameterized Pseudo-Differential Operators for Graph Convolutional Neural Netwo rks
Kevin M. Potter,Steven Richard Sleder,Matthew David Smith,John Tencer
We present a novel graph convolutional layer that is fast, conceptually simple, and provides high accuracy with reduced overfitting. Based on pseudo-differentia l operators, our layer operates on graphs with relative position information ava ilable for each pair of connected nodes. We evaluate our method on a variety of supervised learning tasks, including superpixel image classification using the M NIST, CIFAR10, and CIFAR100 superpixel datasets, node correspondence using the F AUST dataset, and shape classification using the ModelNet10 dataset. The new la yer outperforms multiple recent architectures on superpixel image classification tasks using the MNIST and CIFAR100 superpixel datasets and performs comparably with recent results on the CIFAR10 superpixel dataset. We measure test accuracy without bias to the test set by selecting the model with the best training accur acy. The new layer achieves a test error rate of 0.80% on the MNIST superpixel d ataset, beating the closest reported rate of 0.95% by a factor of more than 15%. After dropping roughly 70% of the edge connections from the input by performing a Delaunay triangulation, our model still achieves a competitive error rate of 1.04%.
****
Bowtie Networks: Generative Modeling for Joint Few-Shot Recognition and Novel-Vi

ew Synthesis
Zhipeng Bao,Yu-Xiong Wang,Martial Hebert
We propose a novel task of joint few-shot recognition and novel-view synthesis: given only one or few images of a novel object from arbitrary views with only category annotation, we aim to simultaneously learn an object classifier and generate images of that type of object from new viewpoints. While existing work copes with two or more tasks mainly by multi-task learning of shareable feature representations, we take a different perspective. We focus on the interaction and cooperation between a generative model and a discriminative model, in a way that facilitates knowledge to flow across tasks in complementary directions. To this end, we propose bowtie networks that jointly learn 3D geometric and semantic representations with a feedback loop. Experimental evaluation on challenging fine-grained recognition datasets demonstrates that our synthesized images are realistic from multiple viewpoints and significantly improve recognition performance as ways of data augmentation, especially in the low-data regime.
****
Asymmetric self-play for automatic goal discovery in robotic manipulation
OpenAI OpenAI,Matthias Plappert,Raul Sampedro,Tao Xu,Ilge Akkaya,Vineet Kosaraju,Peter Welinder,Ruben D'Sa,Arthur Petron,Henrique Ponde de Oliveira Pinto,Alex Paino,Hyeonwoo Noh,Lilian Weng,Qiming Yuan,Casey Chu,Wojciech Zaremba
We train a single, goal-conditioned policy that can solve many robotic manipulation tasks, including tasks with previously unseen goals and objects. To do so, we rely on asymmetric self-play for goal discovery, where two agents, Alice and Bob, play a game. Alice is asked to propose challenging goals and Bob aims to solve them. We show that this method is able to discover highly diverse and complex goals without any human priors. We further show that Bob can be trained with only sparse rewards, because the interaction between Alice and Bob results in a natural curriculum and Bob can learn from Alice's trajectory when relabeled as a goal-conditioned demonstration. Finally, we show that our method scales, resulting in a single policy that can transfer to many unseen hold-out tasks such as setting a table, stacking blocks, and solving simple puzzles. Videos of a learned policy is available at https://robotics-self-play.github.io.
****
SLAPS: Self-Supervision Improves Structure Learning for Graph Neural Networks
Bahare Fatemi,Seyed Mehran Kazemi,Layla El Asri
Graph neural networks (GNNs) work well when the graph structure is provided. However, this structure may not always be available in real-world applications. One solution to this problem is to infer the latent structure and then apply a GNN to the inferred graph. Unfortunately, the space of possible graph structures grows super-exponentially with the number of nodes and so the available node labels may be insufficient for learning both the structure and the GNN parameters. In this work, we propose the Simultaneous Learning of Adjacency and GNN Parameters with Self-supervision, or SLAPS, a method that provides more supervision for inferring a graph structure. This approach consists of training a denoising autoencoder GNN in parallel with the task-specific GNN. The autoencoder is trained to reconstruct the initial node features given noisy node features as well as a structure provided by a learnable graph generator. We explore the design space of SLAPS by comparing different graph generation and symmetrization approaches. A comprehensive experimental study demonstrates that SLAPS scales to large graphs with hundreds of thousands of nodes and outperforms several models that have been proposed to learn a task-specific graph structure on established benchmarks.
****
Constraining Latent Space to Improve Deep Self-Supervised e-Commerce Products Embeddings for Downstream Tasks
Cristian Cardellino,Rafael Carrascosa
The representation of products in a e-commerce marketplace is a key aspect to be exploited when trying to improve the user experience on the site. A well known example of the importance of a good product representation are tasks such as product search or product recommendation. There is however a multitude of lesser known tasks relevant to the business, examples are the detection of counterfeit i

tems, the estimation of package sizes or the categorization of products, among o
thers. It is in this setting that good vector representations of products that c
an be reused on different tasks are very valuable. Past years have seen a major
increase in research in the area of latent representations for products in e-Com
merce. Examples of this are models like Prod2Vec or Meta-Prod2Vec which leverage
 from the information of a user session in order to generate vectors of the prod
ucts that can be used in product recommendations. This work proposes a novel dee
p encoder model for learning product embeddings to be applied in several downstr
eam tasks. The model uses pairs of products that appear together in a browsing s
ession of the users and adds a proximity constraint to the final latent space in
 order to project the embeddings of similar products close to each other. This h
as a regularization effect which gives better features representations to use ac
ross multiple downstream tasks, we explore such effect in our experimentation by
 assessing its impact on the performance of the tasks. Our experiments show effe
ctiveness in transfer learning scenarios comparable to several industrial baseli
nes.
****

A Unified Paths Perspective for Pruning at Initialization
Thomas Gebhart,Udit Saxena,Paul R. Schrater
A number of recent approaches have been proposed for pruning neural network para
meters at initialization with the goal of reducing the size and computational bu
rden of models while minimally affecting their training dynamics and generalizat
ion performance. While each of these approaches have some amount of well-founded
 motivation, a rigorous analysis of the effect of these pruning methods on netwo
rk training dynamics and their formal relationship to each other has thus far re
ceived little attention. Leveraging recent theoretical approximations provided b
y the Neural Tangent Kernel, we unify a number of popular approaches for pruning
 at initialization under a single path-centric framework. We introduce the Path
Kernel as the data-independent factor in a decomposition of the Neural Tangent K
ernel and show the global structure of the Path Kernel can be computed efficient
ly. This Path Kernel decomposition separates the architectural effects from the
data-dependent effects within the Neural Tangent Kernel, providing a means to pr
edict the convergence dynamics of a network from its architecture alone. We anal
yze the use of this structure in approximating training and generalization perfo
rmance of networks in the absence of data across a number of initialization prun
ing approaches. Observing the relationship between input data and paths and the
relationship between the Path Kernel and its natural norm, we additionally propo
se two augmentations of the SynFlow algorithm for pruning at initialization.
****

Formalizing Generalization and Robustness of Neural Networks to Weight Perturbat
ions
Yu-Lin Tsai,Chia-Yi Hsu,Chia-Mu Yu,Pin-Yu Chen
Studying the sensitivity of weight perturbation in neural networks and its impac
ts on model performance, including generalization and robustness, is an active r
esearch topic due to its implications on a wide range of machine learning tasks
such as model compression, generalization gap assessment, and adversarial attack
s. In this paper, we provide the first formal analysis for feed-forward neural n
etworks with non-negative monotone activation functions against norm-bounded wei
ght perturbations, in terms of the robustness in pairwise class margin functions
 and the Rademacher complexity for generalization. We further design a new theor
y-driven loss function for training generalizable and robust neural networks aga
inst weight perturbations. Empirical experiments are conducted to validate our t
heoretical analysis. Our results offer fundamental insights for characterizing t
he generalization and robustness of neural networks against weight perturbations
.
****

Saliency is a Possible Red Herring When Diagnosing Poor Generalization
Joseph D Viviano,Becks Simpson,Francis Dutil,Yoshua Bengio,Joseph Paul Cohen
Poor generalization is one symptom of models that learn to predict target variab
les using spuriously-correlated image features present only in the training dist

ribution instead of the true image features that denote a class. It is often tho
ught that this can be diagnosed visually using attribution (aka saliency) maps.
We study if this assumption is correct. In some prediction tasks, such as for me
dical images, one may have some images with masks drawn by a human expert, indic
ating a region of the image containing relevant information to make the predicti
on. We study multiple methods that take advantage of such auxiliary labels, by t
raining networks to ignore distracting features which may be found outside of th
e region of interest. This mask information is only used during training and has
 an impact on generalization accuracy depending on the severity of the shift bet
ween the training and test distributions. Surprisingly, while these methods impr
ove generalization performance in the presence of a covariate shift, there is no
 strong correspondence between the correction of attribution towards the feature
s a human expert have labelled as important and generalization performance. Thes
e results suggest that the root cause of poor generalization may not always be s
patially defined, and raise questions about the utility of masks as 'attribution
 priors' as well as saliency maps for explainable predictions.
****

(Updated submission 11/20/2020) MISIM: A Novel Code Similarity System
Fangke Ye,Shengtian Zhou,Anand Venkat,Ryan Marcus,Nesime Tatbul,Jesmin Jahan Tit
hi,Niranjan Hasabnis,Paul Petersen,Timothy G Mattson,Tim Kraska,Pradeep Dubey,Vi
vek Sarkar,Justin Gottschlich
Semantic code similarity systems are integral to a range of applications from co
de recommendation to automated software defect correction. Yet, these systems st
ill lack the maturity in accuracy for general and reliable wide-scale usage. To
help address this, we present Machine Inferred Code Similarity (MISIM), a novel
end-to-end code similarity system that consists of two core components. First, M
ISIM uses a novel context-aware semantic structure (CASS), which is designed to
aid in lifting semantic meaning from code syntax. We compare CASS with the abstr
act syntax tree (AST) and show CASS is more accurate than AST by up to 1.67x. Se
cond, MISIM provides a neural-based code similarity scoring algorithm, which can
 be implemented with various neural network architectures with learned parameter
s. We compare MISIM to four state-of-the-art systems: (i) Aroma, (ii) code2seq,
(iii) code2vec, and (iv) Neural Code Comprehension. In our experimental evaluati
on across 328,155 programs (over 18 million lines of code), MISIM has 1.5x to 43
.4x better accuracy across all four systems.

****

Disentangled cyclic reconstruction for domain adaptation
David Bertoin,Emmanuel Rachelson
The domain adaptation problem involves learning a unique classification or regre
ssion model capable of performing on both a source and a target domain. Although
 the labels for the source data are available during training, the labels in the
 target domain are unknown. An effective way to tackle this problem lies in extr
acting insightful features invariant to the source and target domains. In this w
ork, we propose splitting the information for each domain into a task-related re
presentation and its complimentary context representation. We propose an origina
l method to disentangle these two representations in the single-domain supervise
d case. We then adapt this method to the unsupervised domain adaptation problem.
 In particular, our method allows disentanglement in the target domain, despite
the absence of training labels. This enables the isolation of task-specific info
rmation from both domains and a projection into a common representation. The tas
k-specific representation allows efficient transfer of knowledge acquired from t
he source domain to the target domain. We validate the proposed method on severa
l classical domain adaptation benchmarks and illustrate the benefits of disentan
glement for domain adaptation.
****

Adversarial Data Generation of Multi-category Marked Temporal Point Processes wi
th Sparse, Incomplete, and Small Training Samples
Shashika Ranga Muramudalige,Anura Jayasumana,Haonan Wang
Asynchronous stochastic discrete event based processes are commonplace in applic

ation domains such as social science, homeland security, and health informatics.
  Modeling complex interactions of such event data via marked temporal point pro
cesses (MTPPs) provides the ability of detection and prediction of specific inte
rests or profiles. We present a novel multi-category MTPP generation technique f
or applications where training datasets are inherently sparse, incomplete, and s
mall. The proposed adversarial architecture augments adversarial autoencoder (AA
E) with feature mapping techniques, which includes a transformation between the
categories and timestamps of marked points and the percentile distribution of th
e particular category. The transformation of training data to the distribution
facilitates the accurate capture of underlying process characteristics despite t
he sparseness and incompleteness of data. The proposed method is validated using
 several benchmark datasets. The similarity between actual and generated MTPPs i
s evaluated and compared with a Markov process based baseline. Results demonstra
te the effectiveness and robustness of the proposed technique.
****

What Can You Learn From Your Muscles? Learning Visual Representation from Human Interactions

Kiana Ehsani,Daniel Gordon,Thomas Hai Dang Nguyen,Roozbeh Mottaghi,Ali Farhadi

Learning effective representations of visual data that generalize to a variety o
f downstream tasks has been a long quest for computer vision. Most representatio
n learning approaches rely solely on visual data such as images or videos. In th
is paper, we explore a novel approach, where we use human interaction and attent
ion cues to investigate whether we can learn better representations compared to
visual-only representations. For this study, we collect a dataset of human inter
actions capturing body part movements and gaze in their daily lives. Our experim
ents show that our ``"muscly-supervised" representation that encodes interaction
 and attention cues outperforms a visual-only state-of-the-art method MoCo (He e
t al.,2020), on a variety of target tasks: scene classification (semantic), acti
on recognition (temporal), depth estimation (geometric), dynamics prediction (ph
ysics) and walkable surface estimation (affordance). Our code and dataset are av
ailable at: https://github.com/ehsanik/muscleTorch.
****

Maximum Reward Formulation In Reinforcement Learning

SaiKrishna Gottipati,Yashaswi Pathak,Rohan Nuttall,. Sahir,Raviteja Chunduru,Ahm
ed Touati,Sriram Ganapathi Subramanian,Matthew E. Taylor,Sarath Chandar

Reinforcement learning (RL) algorithms typically deal with maximizing the expect
ed cumulative return (discounted or undiscounted, finite or infinite horizon). H
owever, several crucial applications in the real world, such as drug discovery,
do not fit within this framework because an RL agent only needs to identify stat
es (molecules) that achieve the highest reward within a trajectory and does not
need to optimize for the expected cumulative return. In this work, we formulate
an objective function to maximize the expected maximum reward along a trajectory
, derive a novel functional form of the Bellman equation, introduce the correspo
nding Bellman operators, and provide a proof of convergence. Using this formulat
ion, we achieve state-of-the-art results on the task of molecule generation that
 mimics a real-world drug discovery pipeline.
****

Fuzzy c-Means Clustering for Persistence Diagrams

Thomas Davies,Jack Aspinall,Bryan Wilder,Long Tran-Thanh

Persistence diagrams concisely represent the topology of a point cloud whilst ha
ving strong theoretical guarantees. Most current approaches to integrating topol
ogical information into machine learning implicitly map persistence diagrams to
a Hilbert space, resulting in deformation of the underlying metric structure whi
lst also generally requiring prior knowledge about the true topology of the spac
e. In this paper we give an algorithm for Fuzzy c-Means (FCM) clustering directl
y on the space of persistence diagrams, enabling unsupervised learning that auto
matically captures the topological structure of data, with no prior knowledge or
 additional processing of persistence diagrams. We prove the same convergence gu
arantees as traditional FCM clustering: every convergent subsequence of iterates
 tends to a local minimum or saddle point. We end by presenting experiments wher

e our fuzzy topological clustering algorithm allows for unsupervised top-k can didate selection in settings where (i) the properties of persistence diagrams ma ke them the natural choice over geometric equivalents, and (ii) the probabilisti c membership values let us rank candidates in settings where verifying candidate suitability is expensive: lattice structure classification in materials science and pre-trained model selection in machine learning.
****

D2RL: Deep Dense Architectures in Reinforcement Learning
Samarth Sinha,Homanga Bharadhwaj,Aravind Srinivas,Animesh Garg
While improvements in deep learning architectures have played a crucial role in improving the state of supervised and unsupervised learning in computer vision a nd natural language processing, neural network architecture choices for reinforc ement learning remain relatively under-explored. We take inspiration from succes sful architectural choices in computer vision and generative modeling, and inves tigate the use of deeper networks and dense connections for reinforcement learni ng on a variety of simulated robotic learning benchmark environments. Our findin gs reveal that current methods benefit significantly from dense connections and deeper networks, across a suite of manipulation and locomotion tasks, for both p roprioceptive and image-based observations. We hope that our results can serve a s a strong baseline and further motivate future research into neural network arc hitectures for reinforcement learning. The project website is at this link https ://sites.google.com/view/d2rl-anonymous/home
****

Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning
Rishabh Agarwal,Marlos C. Machado,Pablo Samuel Castro,Marc G Bellemare
Reinforcement learning methods trained on few environments rarely learn policies that generalize to unseen environments. To improve generalization, we incorpora te the inherent sequential structure in reinforcement learning into the represen tation learning process. This approach is orthogonal to recent approaches, which rarely exploit this structure explicitly. Specifically, we introduce a theoreti cally motivated policy similarity metric (PSM) for measuring behavioral similari ty between states. PSM assigns high similarity to states for which the optimal p olicies in those states as well as in future states are similar. We also present a contrastive representation learning procedure to embed any state similarity m etric, which we instantiate with PSM to obtain policy similarity embeddings (PSE s). We demonstrate that PSEs improve generalization on diverse benchmarks, inclu ding LQR with spurious correlations, a jumping task from pixels, and Distracting DM Control Suite.
****

Prior-guided Bayesian Optimization
Artur Souza,Luigi Nardi,Leonardo Oliveira,Kunle Olukotun,Marius Lindauer,Frank H utter
While Bayesian Optimization (BO) is a very popular method for optimizing expensi ve black-box functions, it fails to leverage the experience of domain experts. T his causes BO to waste function evaluations on bad design choices (e.g., machine learning hyperparameters) that the expert already knows to work poorly. To add ress this issue, we introduce Prior-guided Bayesian Optimization (PrBO). PrBO al lows users to inject their knowledge into the optimization process in the form o f priors about which parts of the input space will yield the best performance, r ather than BO’s standard priors over functions (which are much less intuitive fo r users). PrBO then combines these priors with BO’s standard probabilistic model to form a pseudo-posterior used to select which points to evaluate next. We sho w that PrBO is around 12x faster than state-of-the-art methods without user prio rs and 10,000x faster than random search on a common suite of benchmarks, and ac hieves a new state-of-the-art performance on a real-world hardware design applic ation. We also show that PrBO converges faster even if the user priors are not e ntirely accurate and that it robustly recovers from misleading priors.
****

Why Lottery Ticket Wins? A Theoretical Perspective of Sample Complexity on Spars

e Neural Networks
Shuai Zhang,Meng Wang,Sijia Liu,Pin-Yu Chen,Jinjun Xiong
The $\textit{lottery ticket hypothesis}$ (LTH) states that learning on a properly pruned network (the $\textit{winning ticket}$) has improved test accuracy over the originally unpruned network. Although LTH has been justified empirically in a broad range of deep neural network (DNN) involved applications like computer vision and natural language processing, the theoretical validation of the improved generalization of a winning ticket remains elusive. To the best of our knowledge, our work, for the first time, characterizes the performance of training a sparse neural network by analyzing the geometric structure of the objective function and the sample complexity to achieve zero generalization error. We show that the convex region near a desirable model with guaranteed generalization enlarges as the neural network model is pruned, indicating the structural importance of a winning ticket. Moreover, as the algorithm for training a sparse neural network is specified as (accelerated) stochastic gradient descent algorithm, we theoretically show that the number of samples required for achieving zero generalization error is proportional to the number of the non-pruned model weights in the hidden layer. With a fixed number of samples, training a pruned neural network enjoys a faster convergence rate to the desirable model than training the original unpruned one, providing a formal justification of the improved generalization of the winning ticket. Our theoretical results are acquired from learning a sparse neural network of one hidden layer, while experimental results are further provided to justify the implications in pruning multi-layer neural networks.
****
Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images
Rewon Child
We present a hierarchical VAE that, for the first time, generates samples quickly $\textit{and}$ outperforms the PixelCNN in log-likelihood on all natural image benchmarks. We begin by observing that, in theory, VAEs can actually represent autoregressive models, as well as faster, better models if they exist, when made sufficiently deep. Despite this, autoregressive models have historically outperformed VAEs in log-likelihood. We test if insufficient depth explains why by scaling a VAE to greater stochastic depth than previously explored and evaluating it CIFAR-10, ImageNet, and FFHQ. In comparison to the PixelCNN, these very deep VAEs achieve higher likelihoods, use fewer parameters, generate samples thousands of times faster, and are more easily applied to high-resolution images. Qualitative studies suggest this is because the VAE learns efficient hierarchical visual representations. We release our source code and models at https://github.com/openai/vdvae.
****
Conservative Safety Critics for Exploration
Homanga Bharadhwaj,Aviral Kumar,Nicholas Rhinehart,Sergey Levine,Florian Shkurti,Animesh Garg
Safe exploration presents a major challenge in reinforcement learning (RL): when active data collection requires deploying partially trained policies, we must ensure that these policies avoid catastrophically unsafe regions, while still enabling trial and error learning. In this paper, we target the problem of safe exploration in RL, by learning a conservative safety estimate of environment states through a critic, and provably upper bound the likelihood of catastrophic failures at every training iteration. We theoretically characterize the tradeoff between safety and policy improvement, show that the safety constraints are satisfied with high probability during training, derive provable convergence guarantees for our approach which is no worse asymptotically then standard RL, and empirically demonstrate the efficacy of the proposed approach on a suite of challenging navigation, manipulation, and locomotion tasks. Our results demonstrate that the proposed approach can achieve competitive task performance, while incurring significantly lower catastrophic failure rates during training as compared to prior methods. Videos are at this URL https://sites.google.com/view/conservative-safety-critics/

****
Hyperrealistic neural decoding: Reconstruction of face stimuli from fMRI measurements via the GAN latent space
Thirza Dado,Ya■mur Güçlütürk,Luca Ambrogioni,Gabrielle Ras,Sander E. Bosch,Marcel van Gerven,Umut Güçlü
We introduce a new framework for hyperrealistic reconstruction of perceived naturalistic stimuli from brain recordings. To this end, we embrace the use of generative adversarial networks (GANs) at the earliest step of our neural decoding pipeline by acquiring functional magnetic resonance imaging data as subjects perceived face images created by the generator network of a GAN. Subsequently, we used a decoding approach to predict the latent state of the GAN from brain data. Hence, latent representations for stimulus (re-)generation are obtained, leading to state-of-the-art image reconstructions. Altogether, we have developed a highly promising approach for decoding sensory perception from brain activity and systematically analyzing neural information processing in the human brain.
****
Multi-Time Attention Networks for Irregularly Sampled Time Series
Satya Narayan Shukla,Benjamin Marlin
Irregular sampling occurs in many time series modeling applications where it presents a significant challenge to standard deep learning models. This work is motivated by the analysis of physiological time series data in electronic health records, which are sparse, irregularly sampled, and multivariate. In this paper, we propose a new deep learning framework for this setting that we call Multi-Time Attention Networks. Multi-Time Attention Networks learn an embedding of continuous time values and use an attention mechanism to produce a fixed-length representation of a time series containing a variable number of observations. We investigate the performance of this framework on interpolation and classification tasks using multiple datasets. Our results show that the proposed approach performs as well or better than a range of baseline and recently proposed models while offering significantly faster training times than current state-of-the-art methods.
****
Graph Traversal with Tensor Functionals: A Meta-Algorithm for Scalable Learning
Elan Sopher Markowitz,Keshav Balasubramanian,Mehrnoosh Mirtaheri,Sami Abu-El-Haija,Bryan Perozzi,Greg Ver Steeg,Aram Galstyan
Graph Representation Learning (GRL) methods have impacted fields from chemistry to social science. However, their algorithmic implementations are specialized to specific use-cases e.g. "message passing" methods are run differently from "node embedding" ones. Despite their apparent differences, all these methods utilize the graph structure, and therefore, their learning can be approximated with stochastic graph traversals. We propose Graph Traversal via Tensor Functionals (GTTF), a unifying meta-algorithm framework for easing the implementation of diverse graph algorithms and enabling transparent and efficient scaling to large graphs. GTTF is founded upon a data structure (stored as a sparse tensor) and a stochastic graph traversal algorithm (described using tensor operations). The algorithm is a functional that accept two functions, and can be specialized to obtain a variety of GRL models and objectives, simply by changing those two functions. We show for a wide class of methods, our algorithm learns in an unbiased fashion and, in expectation, approximates the learning as if the specialized implementations were run directly.
With these capabilities, we scale otherwise non-scalable methods to set state-of-the-art on large graph datasets while being more efficient than existing GRL libraries -- with only a handful of lines of code for each method specialization.
****
Self-Supervised Variational Auto-Encoders
Ioannis Gatopoulos,Jakub Mikolaj Tomczak
Density estimation, compression, and data generation are crucial tasks in artificial intelligence. Variational Auto-Encoders (VAEs) constitute a single framework to achieve these goals. Here, we present a novel class of generative models, called self-supervised Variational Auto-Encoder (selfVAE), that utilizes determin

istic and discrete transformations of data. This class of models allows performi ng both conditional and unconditional sampling while simplifying the objective f unction. First, we use a single self-supervised transformation as a latent varia ble, where a transformation is either downscaling or edge detection. Next, we co nsider a hierarchical architecture, i.e., multiple transformations, and we show its benefits compared to the VAE. The flexibility of selfVAE in data reconstruct ion finds a particularly interesting use case in data compression tasks, where w e can trade-off memory for better data quality, and vice-versa. We present the p erformance of our approach on three benchmark image data (Cifar10, Imagenette64, and CelebA).

****

What they do when in doubt: a study of inductive biases in seq2seq learners

Eugene Kharitonov,Rahma Chaabouni

Sequence-to-sequence (seq2seq) learners are widely used, but we still have only limited knowledge about what inductive biases shape the way they generalize. We address that by investigating how popular seq2seq learners generalize in tasks t hat have high ambiguity in the training data. We use four new tasks to study le arners' preferences for memorization, arithmetic, hierarchical, and compositiona l reasoning. Further, we connect to Solomonoff's theory of induction and propose to use description length as a principled and sensitive measure of inductive bi ases. In our experimental study, we find that LSTM-based learners can learn to p erform counting, addition, and multiplication by a constant from a single traini ng example. Furthermore, Transformer and LSTM-based learners show a bias toward the hierarchical induction over the linear one, while CNN-based learners prefer the opposite. The latter also show a bias toward a compositional generalization over memorization. Finally, across all our experiments, description length prove d to be a sensitive measure of inductive biases.

****

Approximate Birkhoff-von-Neumann decomposition: a differentiable approach

Andrés Hoyos-Idrobo

The Birkhoff-von-Neumann (BvN) decomposition is a standard tool used to draw per mutation matrices from a doubly stochastic (DS) matrix. The BvN decomposition re presents such a DS matrix as a convex combination of several permutation matrice s. Currently, most algorithms to compute the BvN decomposition employ either gre edy strategies or custom-made heuristics. In this paper, we present a novel diff erentiable approach to approximate the BvN decomposition. Our algorithm builds u pon recent advances in Riemannian optimization on Birkhoff polytopes. We offer a n empirical evaluation of this approach in the fairness of exposure in rankings, where we show that the outcome of our method behaves similarly to greedy algori thms. Our approach is an excellent addition to existing methods for sampling fro m DS matrices, such as sampling from a Gumbel-Sinkhorn distribution. However, ou r approach is better suited for applications where the latency in prediction tim e is a constraint. Indeed, we can generally precompute an approximated BvN decom position offline. Then, we select a permutation matrix at random with probabilit y proportional to its coefficient. Finally, we provide an implementation of our method.

****

Adaptive N-step Bootstrapping with Off-policy Data

Guan Wang,Dong Yan,Hang Su,Jun Zhu

The definition of the update target is a crucial design choice in re inforcement learning. Due to the low computation cost and empirical high perfor mance, n-step returns with off-policy data is a widely used update target to bo otstrap from scratch. A critical issue of applying n-step returns is to identify the optimal value of n. In practice, n is often set to a fixed value, which is either determined by an empirical guess or by some hyper-parameter search. In this work, we point out that the optimal value of n actually differs on each da ta point, while the fixed value n is a rough average of them. The estimation err or can be decomposed into two sources, off-policy bias and approximation error, and the fixed value of n is a trade-off between them. Based on that observation , we introduce a new metric, policy age, to quantify the off-policyness of each

data point. We propose the Adaptive N-step Bootstrapping, which calculates the value of n for each data point by its policy age instead of the empirical guess. We conduct experiments on both MuJoCo and Atari games. The results show that adaptive n-step bootstrap-ping achieves state-of-the-art performance in terms of both final reward and data efficiency.
****
Async-RED: A Provably Convergent Asynchronous Block Parallel Stochastic Method using Deep Denoising Priors
Yu Sun,Jiaming Liu,Yiran Sun,Brendt Wohlberg,Ulugbek Kamilov
Regularization by denoising (RED) is a recently developed framework for solving inverse problems by integrating advanced denoisers as image priors. Recent work has shown its state-of-the-art performance when combined with pre-trained deep denoisers. However, current RED algorithms are inadequate for parallel processing on multicore systems. We address this issue by proposing a new{asynchronous RED (Async-RED) algorithm that enables asynchronous parallel processing of data, making it significantly faster than its serial counterparts for large-scale inverse problems. The computational complexity of Async-RED is further reduced by using a random subset of measurements at every iteration. We present a complete theoretical analysis of the algorithm by establishing its convergence under explicit assumptions on the data-fidelity and the denoiser. We validate Async-RED on image recovery using pre-trained deep denoisers as priors.
****
Estimation of Number of Communities in Assortative Sparse Networks
Neil Hwang,Jiarui Xu,Shirshendu Chatterjee,Sharmodeep Bhattacharyya
Most community detection algorithms assume the number of communities, K, to be known \textit{a priori}. Among various approaches that have been proposed to estimate K, the non-parametric methods based on the spectral properties of the Bethe Hessian matrices have garnered much popularity for their simplicity, computational efficiency, and robust performance irrespective of the sparsity of the input data. Recently, one such method has been shown to estimate K consistently if the input network is generated from the (semi-dense) stochastic block model, when the average of the expected degrees ($\tilde{d}$) of all the nodes in the network satisfies $\tilde{d} \gg \log(N)$ (N being the number of nodes in the network). In this paper, we prove some finite sample results that hold for $\tilde{d} = o(\log(N))$, which in turn show that the estimation of K based on the spectra of the Bethe Hessian matrices is consistent not only for the semi-dense regime, but also for the sub-logarithmic sparse regime when $1 \ll \tilde{d} \ll \log(N)$. Thus, our estimation procedure is a robust method for a wide range of problem settings, regardless of the sparsity of the network input.
****
Learning Consistent Deep Generative Models from Sparse Data via Prediction Constraints
Gabriel Hope,Madina Abdrakhmanova,Xiaoyin Chen,Michael C Hughes,Erik B Sudderth
We develop a new framework for learning variational autoencoders and other deep generative models that balances generative and discriminative goals. Our framework optimizes model parameters to maximize a variational lower bound on the likelihood of observed data, subject to a task-specific prediction constraint that prevents model misspecification from leading to inaccurate predictions. We further enforce a consistency constraint, derived naturally from the generative model, that requires predictions on reconstructed data to match those on the original data. We show that these two contributions -- prediction constraints and consistency constraints -- lead to promising image classification performance, especially in the semi-supervised scenario where category labels are sparse but unlabeled data is plentiful. Our approach enables advances in generative modeling to directly boost semi-supervised classification performance, an ability we demonstrate by augmenting deep generative models with latent variables capturing spatial transformations.
****
Tomographic Auto-Encoder: Unsupervised Bayesian Recovery of Corrupted Data
Francesco Tonolini,Pablo Garcia Moreno,Andreas Damianou,Roderick Murray-Smith

We propose a new probabilistic method for unsupervised recovery of corrupted dat
a. Given a large ensemble of degraded samples, our method recovers accurate post
eriors of clean values, allowing the exploration of the manifold of possible rec
onstructed data and hence characterising the underlying uncertainty. In this set
-ting, direct application of classical variational methods often gives rise to c
ollapsed densities that do not adequately explore the solution space. Instead,
we derive our novel reduced entropy condition approximate inference method that
results in rich posteriors. We test our model in a data recovery task under the
 common setting of missing values and noise, demonstrating superior performance
to existing variational methods for imputation and de-noising with different rea
l data sets. We further show higher classification accuracy after imputation, pr
oving the advantage of propagating uncertainty to downstream tasks with our mode
l.
****

OPAL: Offline Primitive Discovery for Accelerating Offline Reinforcement Learnin
g

Anurag Ajay,Aviral Kumar,Pulkit Agrawal,Sergey Levine,Ofir Nachum

Reinforcement learning (RL) has achieved impressive performance in a variety of
online settings in which an agent’s ability to query the environment for transit
ions and rewards is effectively unlimited. However, in many practical applicatio
ns, the situation is reversed: an agent may have access to large amounts of und
irected offline experience data, while access to the online environment is sever
ely limited. In this work, we focus on this offline setting. Our main insight is
 that, when presented with offline data composed of a variety of behaviors, an e
ffective way to leverage this data is to extract a continuous space of recurring
 and temporally extended primitive behaviors before using these primitives for d
ownstream task learning. Primitives extracted in this way serve two purposes:
they delineate the behaviors that are supported by the data from those that are
not, making them useful for avoiding distributional shift in offline RL; and the
y provide a degree of temporal abstraction, which reduces the effective horizon
yielding better learning in theory, and improved offline RL in practice. In addi
tion to benefiting offline policy optimization, we show that performing offline
primitive learning in this way can also be leveraged for improving few-shot imit
ation learning as well as exploration and transfer in online RL on a variety of
benchmark domains. Visualizations and code are available at https://sites.google
.com/view/opal-iclr
****

Knowledge distillation via softmax regression representation learning

Jing Yang,Brais Martinez,Adrian Bulat,Georgios Tzimiropoulos

This paper addresses the problem of model compression via knowledge distillation
. We advocate for a method that optimizes the output feature of the penultimate
layer of the student network and hence is directly related to representation lea
rning. Previous distillation methods which typically impose direct feature match
ing between the student and the teacher do not take into account the classificat
ion problem at hand. On the contrary, our distillation method decouples represen
tation learning and classification and utilizes the teacher's pre-trained classi
fier to train the student's penultimate layer feature. In particular, for the sa
me input image, we wish the teacher's and student's feature to produce the same
output when passed through the teacher's classifier which is achieved with a sim
ple L_2 loss. Our method is extremely simple to implement and straightforward
to train and is shown to consistently outperform previous state-of-the-art metho
ds over a large set of experimental settings including different (a) network arc
hitectures, (b) teacher-student capacities, (c) datasets, and (d) domains. The c
ode will be available at \url{https://github.com/jingyang2017/KD_SRRL}.
****

Large Associative Memory Problem in Neurobiology and Machine Learning

Dmitry Krotov,John J. Hopfield

Dense Associative Memories or modern Hopfield networks permit storage and reliab
le retrieval of an exponentially large (in the dimension of feature space) numb
er of memories. At the same time, their naive implementation is non-biological,

since it seemingly requires the existence of many-body synaptic junctions betwee
n the neurons. We show that these models are effective descriptions of a more m
icroscopic (written in terms of biological degrees of freedom) theory that has a
dditional (hidden) neurons and only requires two-body interactions between them.
 For this reason our proposed microscopic theory is a valid model of large assoc
iative memory with a degree of biological plausibility. The dynamics of our netw
ork and its reduced dimensional equivalent both minimize energy (Lyapunov) funct
ions. When certain dynamical variables (hidden neurons) are integrated out from
our microscopic theory, one can recover many of the models that were previously
discussed in the literature, e.g. the model presented in "Hopfield Networks is A
ll You Need" paper. We also provide an alternative derivation of the energy func
tion and the update rule proposed in the aforementioned paper and clarify the re
lationships between various models of this class.
****

FTBNN: Rethinking Non-linearity for 1-bit CNNs and Going Beyond
Zhuo Su,Linpu Fang,Deke Guo,Dewen Hu,Matti Pietikäinen,Li Liu
Binary neural networks (BNNs), where both weights and activations are binarized
into 1 bit, have been widely studied in recent years due to its great benefit of
 highly accelerated computation and substantially reduced memory footprint that
appeal to the development of resource constrained devices. In contrast to previo
us methods tending to reduce the quantization error for training BNN structures,
 we argue that the binarized convolution process owns an increasing linearity to
wards the target of minimizing such error, which in turn hampers BNN's discrimin
ative ability. In this paper, we re-investigate and tune proper non-linear modul
es to fix that contradiction, leading to a strong baseline which achieves state-
of-the-art performance on the large-scale ImageNet dataset in terms of accuracy
and training efficiency. To go further, we find that the proposed BNN model stil
l has much potential to be compressed by making a better use of the efficient bi
nary operations, without losing accuracy. In addition, the limited capacity of t
he BNN model can also be increased with the help of group execution. Based on th
ese insights, we are able to improve the baseline with an additional 4$\sim$5%
top-1 accuracy gain even with less computational cost. Our code and all trained
models will be made public.
****

Remembering for the Right Reasons: Explanations Reduce Catastrophic Forgetting
Sayna Ebrahimi,Suzanne Petryk,Akash Gokul,William Gan,Joseph E. Gonzalez,Marcus
Rohrbach,trevor darrell
The goal of continual learning (CL) is to learn a sequence of tasks without suff
ering from the phenomenon of catastrophic forgetting. Previous work has shown th
at leveraging memory in the form of a replay buffer can reduce performance degra
dation on prior tasks. We hypothesize that forgetting can be further reduced whe
n the model is encouraged to remember the \textit{evidence} for previously made
decisions. As a first step towards exploring this hypothesis, we propose a simpl
e novel training paradigm, called Remembering for the Right Reasons (RRR), that
additionally stores visual model explanations for each example in the buffer and
 ensures the model has ``the right reasons'' for its predictions by encouraging
its explanations to remain consistent with those used to make decisions at train
ing time. Without this constraint, there is a drift in explanations and increase
 in forgetting as conventional continual learning algorithms learn new tasks. We
 demonstrate how RRR can be easily added to any memory or regularization-based a
pproach and results in reduced forgetting, and more importantly, improved model
explanations. We have evaluated our approach in the standard and few-shot settin
gs and observed a consistent improvement across various CL approaches using diff
erent architectures and techniques to generate model explanations and demonstrat
ed our approach showing a promising connection between explainability and contin
ual learning. Our code is available at \url{https://github.com/SaynaEbrahimi/Rem
embering-for-the-Right-Reasons}.
****

Deep Networks and the Multiple Manifold Problem
Sam Buchanan,Dar Gilboa,John Wright

We study the multiple manifold problem, a binary classification task modeled on applications in machine vision, in which a deep fully-connected neural network is trained to separate two low-dimensional submanifolds of the unit sphere. We provide an analysis of the one-dimensional case, proving for a simple manifold configuration that when the network depth L is large relative to certain geometric and statistical properties of the data, the network width n grows as a sufficiently large polynomial in L, and the number of i.i.d. samples from the manifolds is polynomial in L, randomly-initialized gradient descent rapidly learns to classify the two manifolds perfectly with high probability. Our analysis demonstrates concrete benefits of depth and width in the context of a practically-motivated model problem: the depth acts as a fitting resource, with larger depths corresponding to smoother networks that can more readily separate the class manifolds, and the width acts as a statistical resource, enabling concentration of the randomly-initialized network and its gradients. The argument centers around the "neural tangent kernel" of Jacot et al. and its role in the nonasymptotic analysis of training overparameterized neural networks; to this literature, we contribute essentially optimal rates of concentration for the neural tangent kernel of deep fully-connected ReLU networks, requiring width $n \geq L\,\mathrm{poly}(d_0)$ to achieve uniform concentration of the initial kernel over a d_0-dimensional submanifold of the unit sphere $\mathbb{S}^{n_0-1}$, and a nonasymptotic framework for establishing generalization of networks trained in the "NTK regime" with structured data. The proof makes heavy use of martingale concentration to optimally treat statistical dependencies across layers of the initial random network. This approach should be of use in establishing similar results for other network architectures.

****

Transfer Learning of Graph Neural Networks with Ego-graph Information Maximization

Qi Zhu,Yidan Xu,Haonan Wang,Chao Zhang,Jiawei Han,Carl Yang

Graph neural networks (GNNs) have been shown with superior performance in various applications, but training dedicated GNNs can be costly for large-scale graphs. Some recent work started to study the pre-training of GNNs. However, none of them provide theoretical insights into the design of their frameworks, or clear requirements and guarantees towards the transferability of GNNs. In this work, we establish a theoretically grounded and practically useful framework for the transfer learning of GNNs. Firstly, we propose a novel view towards the essential graph information and advocate the capturing of it as the goal of transferable GNN training, which motivates the design of EGI (ego-graph information maximization) to analytically achieve this goal. Secondly, we specify the requirement of structure-respecting node features as the GNN input, and conduct a rigorous analysis of GNN transferability based on the difference between the local graph Laplacians of the source and target graphs. Finally, we conduct controlled synthetic experiments to directly justify our theoretical conclusions. Extensive experiments on real-world networks towards role identification show consistent results in the rigorously analyzed setting of direct-transfering (freezing parameters), while those towards large-scale relation prediction show promising results in the more generalized and practical setting of transfering with fine-tuning.

****

Chameleon: Learning Model Initializations Across Tasks With Different Schemas

Lukas Brinkmeyer,Rafael Rego Drumond,Randolf Scholz,Josif Grabocka,Lars Schmidt-Thieme

Parametric models, and particularly neural networks, require weight initialization as a starting point for gradient-based optimization. Recent work shows that an initial parameter set can be learned from a population of supervised learning tasks that enables a fast convergence for unseen tasks even when only a handful of instances is available (model-agnostic meta-learning).

Currently, methods for learning model initializations are limited to a population of tasks sharing the same schema, i.e., the same number, order, type, and semantics of predictor and target variables.

In this paper, we address the problem of meta-learning weight initialization acr

oss tasks with different schemas, for example, if the number of predictors varie s across tasks, while they still share some variables. We propose Chameleon, a m odel that learns to align different predictor schemas to a common representation .
In experiments on 23 datasets of the OpenML-CC18 benchmark, we show that Chamele on can successfully learn parameter initializations across tasks with different schemas, presenting, to the best of our knowledge, the first cross-dataset few-s hot classification approach for unstructured data.
****
Interpreting Knowledge Graph Relation Representation from Word Embeddings
Carl Allen,Ivana Balazevic,Timothy Hospedales
Many models learn representations of knowledge graph data by exploiting its low- rank latent structure, encoding known relations between entities and enabling un known facts to be inferred. To predict whether a relation holds between entities , embeddings are typically compared in the latent space following a relation-spe cific mapping. Whilst their predictive performance has steadily improved, how su ch models capture the underlying latent structure of semantic information remain s unexplained. Building on recent theoretical understanding of word embeddings, we categorise knowledge graph relations into three types and for each derive exp licit requirements of their representations. We show that empirical properties o f relation representations and the relative performance of leading knowledge gra ph representation methods are justified by our analysis.
****
Learning perturbation sets for robust machine learning
Eric Wong,J Zico Kolter
Although much progress has been made towards robust deep learning, a significant gap in robustness remains between real-world perturbations and more narrowly de fined sets typically studied in adversarial defenses. In this paper, we aim to b ridge this gap by learning perturbation sets from data, in order to characterize real-world effects for robust training and evaluation. Specifically, we use a c onditional generator that defines the perturbation set over a constrained region of the latent space. We formulate desirable properties that measure the quality of a learned perturbation set, and theoretically prove that a conditional varia tional autoencoder naturally satisfies these criteria. Using this framework, our approach can generate a variety of perturbations at different complexities and scales, ranging from baseline spatial transformations, through common image corr uptions, to lighting variations. We measure the quality of our learned perturbat ion sets both quantitatively and qualitatively, finding that our models are capa ble of producing a diverse set of meaningful perturbations beyond the limited da ta seen during training. Finally, we leverage our learned perturbation sets to t rain models which are empirically and certifiably robust to adversarial image co rruptions and adversarial lighting variations, while improving generalization on non-adversarial data. All code and configuration files for reproducing the expe riments as well as pretrained model weights can be found at https://github.com/l ocuslab/perturbation_learning.
****
Efficient Architecture Search for Continual Learning
Qiang Gao,Zhipeng Luo,Diego Klabjan,Fengli Zhang
Continual learning with neural networks is an important learning framework in AI that aims to learn a sequence of tasks well. However, it is often confronted wi th three challenges: (1) overcome the catastrophic forgetting problem, (2) adapt the current network to new tasks, and meanwhile (3) control its model complexit y. To reach these goals, we propose a novel approach named as Continual Learning with Efficient Architecture Search, or CLEAS in short. CLEAS works closely with neural architecture search (NAS) which leverages reinforcement learning techniq ues to search for the best neural architecture that fits a new task. In particul ar, we design a neuron-level NAS controller that decides which old neurons from previous tasks should be reused (knowledge transfer), and which new neurons shou ld be added (to learn new knowledge). Such a fine-grained controller allows find ing a very concise architecture that can fit each new task well. Meanwhile, sinc

e we do not alter the weights of the reused neurons, we perfectly memorize the k nowledge learned from previous tasks. We evaluate CLEAS on numerous sequential c lassification tasks, and the results demonstrate that CLEAS outperforms other st ate-of-the-art alternative methods, achieving higher classification accuracy whi le using simpler neural architectures.

****
Gradient Projection Memory for Continual Learning
Gobinda Saha,Isha Garg,Kaushik Roy
The ability to learn continually without forgetting the past tasks is a desired attribute for artificial learning systems. Existing approaches to enable such le arning in artificial neural networks usually rely on network growth, importance based weight update or replay of old data from the memory. In contrast, we propo se a novel approach where a neural network learns new tasks by taking gradient s teps in the orthogonal direction to the gradient subspaces deemed important for the past tasks. We find the bases of these subspaces by analyzing network repres entations (activations) after learning each task with Singular Value Decompositi on (SVD) in a single shot manner and store them in the memory as Gradient Projec tion Memory (GPM). With qualitative and quantitative analyses, we show that such orthogonal gradient descent induces minimum to no interference with the past ta sks, thereby mitigates forgetting. We evaluate our algorithm on diverse image cl assification datasets with short and long sequences of tasks and report better o r on-par performance compared to the state-of-the-art approaches.
****
Evaluating Online Continual Learning with CALM
Germán Kruszewski,Ionut Teodor Sorodoc,Tomas Mikolov
Online Continual Learning (OCL) studies learning over a continuous data stream w ithout observing any single example more than once, a setting that is closer to the experience of humans and systems that must learn “on-the-wild”. Yet, commonl y available benchmarks are far from these real world conditions, because they ex plicitly signal different tasks, lack latent similarity structure or assume temp oral independence between different examples. Here, we propose a new benchmark f or OCL based on language modelling in which input alternates between different l anguages and domains without any explicit delimitation. Additionally, we propose new metrics to study catastrophic forgetting in this setting and evaluate multi ple baseline models based on compositions of experts. Finally, we introduce a si mple gating technique that learns the latent similarities between different inpu ts, improving the performance of a Products of Experts model.
****
A Trainable Optimal Transport Embedding for Feature Aggregation and its Relation ship to Attention
Grégoire Mialon,Dexiong Chen,Alexandre d'Aspremont,Julien Mairal
We address the problem of learning on sets of features, motivated by the need of performing pooling operations in long biological sequences of varying sizes, wi th long-range dependencies, and possibly few labeled data. To address this chall enging task, we introduce a parametrized representation of fixed size, which em beds and then aggregates elements from a given input set according to the optima l transport plan between the set and a trainable reference. Our approach scales to large datasets and allows end-to-end training of the reference, while also pr oviding a simple unsupervised learning mechanism with small computational cost. Our aggregation technique admits two useful interpretations: it may be seen as a mechanism related to attention layers in neural networks, or it may be seen as a scalable surrogate of a classical optimal transport-based kernel. We experimen tally demonstrate the effectiveness of our approach on biological sequences, ach ieving state-of-the-art results for protein fold recognition and detection of ch romatin profiles tasks, and, as a proof of concept, we show promising results fo r processing natural language sequences. We provide an open-source implementatio n of our embedding that can be used alone or as a module in larger learning mode ls at https://github.com/claying/OTK.
****

Unifying Regularisation Methods for Continual Learning
Frederik Benzing
Continual Learning addresses the challenge of learning a number of different distributions sequentially. The goal of maintaining knowledge of earlier distributions without re-accessing them starkly conflicts with standard SGD training for artificial neural networks. An influential method to tackle this are so-called regularisation approaches. They measure the importance of each parameter for modelling a given distribution and subsequently protect important parameters from large changes. In the literature, three ways to measure parameter importance have been put forward and they have inspired a large body of follow-up work. Here, we present strong theoretical and empirical evidence that these three methods, Elastic Weight Consolidation (EWC), Synaptic Intelligence (SI) and Memory Aware Synapses (MAS), all approximate the Fisher Information. Only EWC intentionally relies on the Fisher, while the other two methods stem from rather different motivations. We find that for SI the relation to the Fisher -- and in fact its performance -- is due to a previously unknown bias. Altogether, this unifies a large body of regularisation approaches. It also provides the first theoretical explanation for the effectiveness of SI- and MAS-based algorithms and offers theoretically justified versions of these algorithms. From a practical viewpoint, our insights offer computational speed-ups and uncover conditions needed for different algorithms to work.
****
Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-Series Data
Sindy Löwe,David Madras,Richard Zemel,Max Welling
Standard causal discovery methods must fit a new model whenever they encounter samples from a new underlying causal graph. However, these samples often share relevant information - for instance, the dynamics describing the effects of causal relations - which is lost when following this approach. We propose Amortized Causal Discovery, a novel framework that leverages such shared dynamics to learn to infer causal relations from time-series data. This enables us to train a single, amortized model that infers causal relations across samples with different underlying causal graphs, and thus makes use of the information that is shared. We demonstrate experimentally that this approach, implemented as a variational model, leads to significant improvements in causal discovery performance, and show how it can be extended to perform well under hidden confounding.

****
RODE: Learning Roles to Decompose Multi-Agent Tasks
Tonghan Wang,Tarun Gupta,Anuj Mahajan,Bei Peng,Shimon Whiteson,Chongjie Zhang
Role-based learning holds the promise of achieving scalable multi-agent learning by decomposing complex tasks using roles. However, it is largely unclear how to efficiently discover such a set of roles. To solve this problem, we propose to first decompose joint action spaces into restricted role action spaces by clustering actions according to their effects on the environment and other agents. Learning a role selector based on action effects makes role discovery much easier because it forms a bi-level learning hierarchy: the role selector searches in a smaller role space and at a lower temporal resolution, while role policies learn in significantly reduced primitive action-observation spaces. We further integrate information about action effects into the role policies to boost learning efficiency and policy generalization. By virtue of these advances, our method (1) outperforms the current state-of-the-art MARL algorithms on 9 of the 14 scenarios that comprise the challenging StarCraft II micromanagement benchmark and (2) achieves rapid transfer to new environments with three times the number of agents. Demonstrative videos can be viewed at https://sites.google.com/view/rode-marl.
****
Graph Representation Learning for Multi-Task Settings: a Meta-Learning Approach
Davide Buffelli,Fabio Vandin
Graph Neural Networks (GNNs) have become the state-of-the-art method for many applications on graph structured data. GNNs are a framework for graph representati

on learning, where a model learns to generate low dimensional node embeddings th
at encapsulate structural and feature-related information. GNNs are usually trai
ned in an end-to-end fashion, leading to highly specialized node embeddings. Whi
le this approach achieves great results in the single-task setting, generating n
ode embeddings that can be used to perform multiple tasks (with performance comp
arable to single-task models) is an open problem. We propose a novel representat
ion learning strategy, based on meta-learning, capable of producing multi-task n
ode embeddings. Our method avoids the difficulties arising when learning to perf
orm multiple tasks concurrently by, instead, learning to quickly (i.e. with a fe
w steps of gradient descent) adapt to multiple tasks singularly. We show that t
he embeddings produced by our method can be used to perform multiple tasks with
comparable or higher performance than both single-task and multi-task end-to-end
 models. Our method is model-agnostic and task-agnostic and can hence be applied
 to a wide variety of multi-task domains.
****
Are Graph Convolutional Networks Fully Exploiting the Graph Structure?
Davide Buffelli,Fabio Vandin
Graph Convolutional Networks (GCNs) represent the state-of-the-art for many grap
h related tasks. At every layer, GCNs rely on the graph structure to define an a
ggregation strategy where each node updates its representation by combining info
rmation from its neighbours. A known limitation of GCNs is their inability to in
fer long-range dependencies. In fact, as the number of layers increases, informa
tion gets smoothed and node embeddings become indistinguishable, negatively affe
cting performance. In this paper we formalize four levels of injection of graph
structural information, and use them to analyze the importance of long-range dep
endencies. We then propose a novel regularization technique based on random walk
s with restart, called RWRReg, which encourages the network to encode long-range
 information into node embeddings. RWRReg does not require additional operations
 at inference time, is model-agnostic, and is further supported by our theoretic
al analysis connecting it to the Weisfeiler-Leman algorithm. Our experimental an
alysis, on both transductive and inductive tasks, shows that the lack of long-ra
nge structural information greatly affects the performance of state-of-the-art m
odels, and that the long-range information exploited by RWRReg leads to an avera
ge accuracy improvement of more than 5% on all considered tasks.
****
Bractivate: Dendritic Branching in Medical Image Segmentation Neural Architectur
e Search
Leila Abdelrahman
Researchers manually compose most neural networks through painstaking experiment
ation.
This process is taxing and explores only a limited subset of possible
architecture. Researchers design architectures to address objectives ranging fro
m low
space complexity to high accuracy through hours of experimentation. Neural archi
tecture
search (NAS) is a thriving field for automatically discovering architectures
achieving these same objectives. Addressing these ever-increasing challenges in
computing, we take inspiration from the brain because it has the most efficient
neuronal wiring of any complex structure; its physiology inspires us to propose
Bractivate, a NAS algorithm inspired by neural dendritic branching. An evolution
ary algorithm that adds new skip connection combinations to the most active bloc
ks in the network, propagating salient
information through the network. We apply our methods to lung x-ray, cell nuclei
microscopy, and electron microscopy segmentation tasks to highlight Bractivate's
 robustness.
Moreover, our ablation studies emphasize dendritic branching's necessity: ablati
ng
these connections leads to significantly lower model performance. We finally
compare our discovered architecture with other state-of-the-art UNet models,
highlighting how efficient skip connections allow Bractivate to achieve comparab

le
results with substantially lower space and time complexity, proving how
Bractivate balances efficiency with performance. We invite you to work with our
code here: \href{https://tinyurl.com/bractivate}{\textcolor{violet}{https://tiny
url.com/bractivate}}.
****
Scalable Bayesian Inverse Reinforcement Learning
Alex James Chan,Mihaela van der Schaar
Bayesian inference over the reward presents an ideal solution to the ill-posed n
ature of the inverse reinforcement learning problem. Unfortunately current metho
ds generally do not scale well beyond the small tabular setting due to the need
for an inner-loop MDP solver, and even non-Bayesian methods that do themselves s
cale often require extensive interaction with the environment to perform well, b
eing inappropriate for high stakes or costly applications such as healthcare. In
 this paper we introduce our method, Approximate Variational Reward Imitation Le
arning (AVRIL), that addresses both of these issues by jointly learning an appro
ximate posterior distribution over the reward that scales to arbitrarily complic
ated state spaces alongside an appropriate policy in a completely offline manner
 through a variational approach to said latent reward. Applying our method to re
al medical data alongside classic control simulations, we demonstrate Bayesian r
eward inference in environments beyond the scope of current methods, as well as
task performance competitive with focused offline imitation learning algorithms.
****
Multi-Task Multicriteria Hyperparameter Optimization
Kirill Akhmetzyanov,Alexander Yuzhakov
We present a new method for searching optimal hyperparameters among several task
s and several criteria. Multi-Task Multi Criteria method (MTMC) provides several
 Pareto-optimal solutions, among which one solution is selected with given crite
ria significance coefficients. The article begins with a mathematical formulatio
n of the problem of choosing optimal hyperparameters. Then, the steps of the MTM
C method that solves this problem are described. The proposed method is evaluate
d on the image classification problem using a convolutional neural network. The
article presents optimal hyperparameters for various criteria significance coeff
icients.
****
Learning "What-if" Explanations for Sequential Decision-Making
Ioana Bica,Daniel Jarrett,Alihan Hüyük,Mihaela van der Schaar
Building interpretable parameterizations of real-world decision-making on the ba
sis of demonstrated behavior--i.e. trajectories of observations and actions made
 by an expert maximizing some unknown reward function--is essential for introspe
cting and auditing policies in different institutions. In this paper, we propose
 learning explanations of expert decisions by modeling their reward function in
terms of preferences with respect to ``"what if'' outcomes: Given the current hi
story of observations, what would happen if we took a particular action? To lear
n these cost-benefit tradeoffs associated with the expert's actions, we integrat
e counterfactual reasoning into batch inverse reinforcement learning. This offer
s a principled way of defining reward functions and explaining expert behavior,
and also satisfies the constraints of real-world decision-making---where active
experimentation is often impossible (e.g. in healthcare). Additionally, by estim
ating the effects of different actions, counterfactuals readily tackle the off-p
olicy nature of policy evaluation in the batch setting, and can naturally accomm
odate settings where the expert policies depend on histories of observations rat
her than just current states. Through illustrative experiments in both real and
simulated medical environments, we highlight the effectiveness of our batch, cou
nterfactual inverse reinforcement learning approach in recovering accurate and i
nterpretable descriptions of behavior.
****
Learning advanced mathematical computations from examples
Francois Charton,Amaury Hayat,Guillaume Lample
Using transformers over large generated datasets, we train models to learn mathe

matical properties of differential systems, such as local stability, behavior at infinity and controllability. We achieve near perfect prediction of qualitative characteristics, and good approximations of numerical features of the system. This demonstrates that neural networks can learn to perform complex computations, grounded in advanced theory, from examples, without built-in mathematical knowledge.

****

Federated learning using mixture of experts

Edvin Listo Zec,John Martinsson,Olof Mogren,Leon René Sütfeld,Daniel Gillblad

Federated learning has received attention for its efficiency and privacy benefits,in settings where data is distributed among devices. Although federated learning shows significant promise as a key approach when data cannot be shared or centralized, current incarnations show limited privacy properties and have shortcomings when applied to common real-world scenarios. One such scenario is heterogeneous data among devices, where data may come from different generating distributions. In this paper, we propose a federated learning framework using a mixture of experts to balance the specialist nature of a locally trained model with the generalist knowledge of a global model in a federated learning setting. Our results show that the mixture of experts model is better suited as a personalized model for devices when data is heterogeneous, outperforming both global and lo-cal models. Furthermore, our framework gives strict privacy guarantees, which allows clients to select parts of their data that may be excluded from the federation. The evaluation shows that the proposed solution is robust to the setting where some users require a strict privacy setting and do not disclose their models to a central server at all, opting out from the federation partially or entirely. The proposed framework is general enough to include any kind of machine learning models, and can even use combinations of different kinds.

****

Guiding Neural Network Initialization via Marginal Likelihood Maximization

Anthony Tai,Chunfeng Huang

We propose a simple, data-driven approach to help guide hyperparameter selection for neural network initialization. We leverage the relationship between neural network and Gaussian process models having corresponding activation and covariance functions to infer the hyperparameter values desirable for model initialization. Our experiment shows that marginal likelihood maximization provides recommendations that yield near-optimal prediction performance on MNIST classification task under experiment constraints. Furthermore, our empirical results indicate consistency in the proposed technique, suggesting that computation cost for the procedure could be significantly reduced with smaller training sets.

****

Just How Toxic is Data Poisoning? A Benchmark for Backdoor and Data Poisoning Attacks

Avi Schwarzschild,Micah Goldblum,Arjun Gupta,John P Dickerson,Tom Goldstein

Data poisoning and backdoor attacks manipulate training data in order to cause models to fail during inference. A recent survey of industry practitioners found that data poisoning is the number one concern among threats ranging from model stealing to adversarial attacks. However, we find that the impressive performance evaluations from data poisoning attacks are, in large part, artifacts of inconsistent experimental design. Moreover, we find that existing poisoning methods have been tested in contrived scenarios, and many fail in more realistic settings. In order to promote fair comparison in future work, we develop standardized benchmarks for data poisoning and backdoor attacks.

****

Diffeomorphic Template Transformers

Tycho F.A. van der Ouderaa,Ivana Isgum,Wouter B. Veldhuis,Bob D. De Vos,Pim Moeskops

In this paper we propose a spatial transformer network where the spatial transformations are limited to the group of diffeomorphisms. Diffeomorphic transformations are a kind of homeomorphism, which by definition preserve topology, a compelling property in certain applications.

We apply this diffemorphic spatial transformer to model the output of a neural n
etwork as a topology preserving mapping of a prior shape. By carefully choosing
the prior shape we can enforce properties on the output of the network without r
equiring any changes to the loss function, such as smooth boundaries and a hard
constraint on the number of connected components.
The diffeomorphic transformer networks outperform their non-diffeomorphic precur
sors when applied to learn data invariances in classification tasks. On a breast
 tissue segmentation task, we show that the approach is robust and flexible enou
gh to deform simple artificial priors, such as Gaussian-shaped prior energies, i
nto high-quality predictive probability densities. In addition to desirable topo
logical properties, the segmentation maps have competitive quantitative fidelity
 compared to those obtained by direct estimation (i.e. plain U-Net).
****

Neural CDEs for Long Time Series via the Log-ODE Method
James Morrill,Patrick Kidger,Cristopher Salvi,James Foster,Terry Lyons
Neural Controlled Differential Equations (Neural CDEs) are the continuous-time a
nalogue of an RNN, just as Neural ODEs are analogous to ResNets. However just li
ke RNNs, training Neural CDEs can be difficult for long time series. Here, we pr
opose to apply a technique drawn from stochastic analysis, namely the log-ODE me
thod. Instead of using the original input sequence, our procedure summarises the
 information over local time intervals via the log-signature map, and uses the r
esulting shorter stream of log-signatures as the new input. This represents a le
ngth/channel trade-off. In doing so we demonstrate efficacy on problems of lengt
h up to 17k observations and observe significant training speed-ups, improvement
s in model performance, and reduced memory requirements compared to the existing
 algorithm.
****

Leveraging Class Hierarchies with Metric-Guided Prototype Learning
Vivien Sainte Fare Garnot,Loic Landrieu
In many classification tasks, the set of classes can be organized according to a
 meaningful hierarchy. This structure can be used to assess the severity of conf
using each pair of classes, and summarized under the form of a cost matrix whic
h also defines a finite metric. We propose to integrate this metric in the super
vision of a prototypical network in order to model the hierarchical class struct
ure. Our method relies on jointly learning a feature-extracting network and a se
t of class representations, or prototypes, which incorporate the error metric in
to their relative arrangement in the embedding space. We show that this simultan
eous training allows for consistent improvement of the severity of the network's
 errors with regard to the class hierarchy when compared to traditional methods
and other prototype-based strategies. Furthermore, when the induced metric conta
ins insight on the data structure, our approach improves the overall precision a
s well. Experiments on four different public datasets—from agricultural time ser
ies classification to depth image semantic segmentation—validate our approach.
****

Repurposing Pretrained Models for Robust Out-of-domain Few-Shot Learning
Namyeong Kwon,Hwidong Na,Gabriel Huang,Simon Lacoste-Julien
Model-agnostic meta-learning (MAML) is a popular method for few-shot learning bu
t assumes that we have access to the meta-training set. In practice, training on
 the meta-training set may not always be an option due to data privacy concerns,
 intellectual property issues, or merely lack of computing resources. In this pa
per, we consider the novel problem of repurposing pretrained MAML checkpoints to
 solve new few-shot classification tasks. Because of the potential distribution
mismatch, the original MAML steps may no longer be optimal. Therefore we propose
 an alternative meta-testing procedure and combine MAML gradient steps with adve
rsarial training and uncertainty-based stepsize adaptation. Our method outperfor
ms "vanilla" MAML on same-domain and cross-domains benchmarks using both SGD and
 Adam optimizers and shows improved robustness to the choice of base stepsize.
****

"Hey, that's not an ODE'": Faster ODE Adjoints with 12 Lines of Code
Patrick Kidger,Ricky T. Q. Chen,Terry Lyons

Neural differential equations may be trained by backpropagating gradients via th
e adjoint method, which is another differential equation typically solved using
an adaptive-step-size numerical differential equation solver. A proposed step is
 accepted if its error, \emph{relative to some norm}, is sufficiently small; els
e it is rejected, the step is shrunk, and the process is repeated. Here, we demo
nstrate that the particular structure of the adjoint equations makes the usual c
hoices of norm (such as L^2) unnecessarily stringent. By replacing it with a m
ore appropriate (semi)norm, fewer steps are unnecessarily rejected and the backp
ropagation is made faster. This requires only minor code modifications. Experime
nts on a wide range of tasks---including time series, generative modeling, and p
hysical control---demonstrate a median improvement of 40% fewer function evaluat
ions. On some problems we see as much as 62% fewer function evaluations, so that
 the overall training time is roughly halved.
****

Reviving Autoencoder Pretraining
You Xie,Nils Thuerey
The pressing need for pretraining algorithms has been diminished by numerous adv
ances in terms of regularization, architectures, and optimizers. Despite this tr
end, we re-visit the classic idea of unsupervised autoencoder pretraining and pr
opose a modified variant that relies on a full reverse pass trained in conjuncti
on with a given training task. We establish links between SVD and pretraining an
d show how it can be leveraged for gaining insights about the learned structures
. Most importantly, we demonstrate that our approach yields an improved performa
nce for a wide variety of relevant learning and transfer tasks ranging from full
y connected networks over ResNets to GANs. Our results demonstrate that unsuperv
ised pretraining has not lost its practical relevance in today's deep learning e
nvironment.
****

Regularized Mutual Information Neural Estimation
Kwanghee Choi,Siyeong Lee
With the variational lower bound of mutual information (MI), the estimation of M
I can be understood as an optimization task via stochastic gradient descent. In
this work, we start by showing how Mutual Information Neural Estimator (MINE) se
arches for the optimal function T that maximizes the Donsker-Varadhan represen
tation. With our synthetic dataset, we directly observe the neural network outpu
ts during the optimization to investigate why MINE succeeds or fails: We discove
r the drifting phenomenon, where the constant term of T is shifting through th
e optimization process, and analyze the instability caused by the interaction be
tween the $logsumexp$ and the insufficient batch size. Next, through theoretical
 and experimental evidence, we propose a novel lower bound that effectively regu
larizes the neural network to alleviate the problems of MINE. We also introduce
an averaging strategy that produces an unbiased estimate by utilizing multiple b
atches to mitigate the batch size limitation. Finally, we show that L^2 regula
rization achieves significant improvements in both discrete and continuous setti
ngs.
****

Class-Weighted Evaluation Metrics for Imbalanced Data Classification
Akhilesh Gupta,Nesime Tatbul,Ryan Marcus,Shengtian Zhou,Insup Lee,Justin Gottsch
lich
Class distribution skews in imbalanced datasets may lead to models with predicti
on bias towards majority classes, making fair assessment of classifiers a challe
nging task. Balanced Accuracy is a popular metric used to evaluate a classifier’
s prediction performance under such scenarios. However, this metric falls short
when classes vary in importance, especially when class importance is skewed diff
erently from class cardinality distributions. In this paper, we propose a simple
 and general-purpose evaluation framework for imbalanced data classification tha
t is sensitive to arbitrary skews in class cardinalities and importances. Experi
ments with several state-of-the-art classifiers tested on real-world datasets an
d benchmarks from two different domains show that our new framework is more effe
ctive than Balanced Accuracy –- not only in evaluating and ranking model predict

ions, but also in training the models themselves.
****
Learning Continuous-Time Dynamics by Stochastic Differential Networks
Yingru Liu,Yucheng Xing,Xuewen Yang,Xin Wang,Jing Shi,Di Jin,Zhaoyue Chen
Learning continuous-time stochastic dynamics is a fundamental and essential problem in modeling sporadic time series, whose observations are irregular and sparse in both time and dimension. For a given system whose latent states and observed data are high-dimensional, it is generally impossible to derive a precise continuous-time stochastic process to describe the system behaviors. To solve the above problem, we apply Variational Bayesian method and propose a flexible continuous-time stochastic recurrent neural network named Variational Stochastic Differential Networks (VSDN), which embed the complicated dynamics of thesporadic time series by neural Stochastic Differential Equations (SDE). VSDNs capture the stochastic dependency among latent states and observations by deep neural networks. We also incorporate two differential Evidence Lower Bounds to efficiently train the models. Through comprehensive experiments, we show that VSDNs outperform state-of-the-art continuous-time deep learning models and achieve remarkable performance on prediction and interpolation tasks for sporadic data.
****
Empirical Analysis of Unlabeled Entity Problem in Named Entity Recognition
Yangming Li,lemao liu,Shuming Shi
In many scenarios, named entity recognition (NER) models severely suffer from unlabeled entity problem, where the entities of a sentence may not be fully annotated. Through empirical studies performed on synthetic datasets, we find two causes of performance degradation. One is the reduction of annotated entities and the other is treating unlabeled entities as negative instances. The first cause has less impact than the second one and can be mitigated by adopting pretraining language models. The second cause seriously misguides a model in training and greatly affects its performances. Based on the above observations, we propose a general approach, which can almost eliminate the misguidance brought by unlabeled entities. The key idea is to use negative sampling that, to a large extent, avoids training NER models with unlabeled entities. Experiments on synthetic datasets and real-world datasets show that our model is robust to unlabeled entity problem and surpasses prior baselines. On well-annotated datasets, our model is competitive with the state-of-the-art method.
****
Luring of transferable adversarial perturbations in the black-box paradigm
Rémi Bernhard,Pierre-Alain Moëllic,Jean-Max Dutertre
The growing interest for adversarial examples, i.e. maliciously modified examples which fool a classifier, has resulted in many defenses intended to detect them, render them inoffensive or make the model more robust against them. In this paper, we pave the way towards a new approach to improve the robustness of a model against black-box transfer attacks. A removable additional neural network is included in the target model and is designed to induce the "luring effect", which tricks the adversary into choosing false directions to fool the target model. Training the additional model is achieved thanks to a loss function acting on the logits sequence order. Our deception-based method only needs to have access to the predictions of the target model and does not require a labeled data set. We explain the luring effect thanks to the notion of robust and non-robust useful features and perform experiments on MNIST, SVHN and CIFAR10 to characterize and evaluate this phenomenon. Additionally, we discuss two simple prediction schemes, and verify experimentally that our approach can be used as a defense to efficiently thwart an adversary using state-of-the-art attacks and allowed to perform large perturbations.
****
Neurally Guided Genetic Programming for Turing Complete Programming by Example
Alexander Newton Wild,Barry Porter
The ability to synthesise source code from input/output examples allows nonexperts to generate programs, and experts to abstract away a wide range of simple pro

gramming tasks. Current research in this area has explored neural synthesis, SMT solvers, and genetic programming; each of these approaches is limited, however, often using highly specialised target languages for synthesis. In this paper we present a novel hybrid approach using neural networks to guide genetic programming (GP), which allows us to successfully synthesise code from just ten I/O examples in a generalised Turing complete target language, up to and including a sorting algorithm. We show that GP by itself is able to synthesise a set of simple programs, and show which hints (suggested lines of code for inclusion) are of most utility to GP in solving harder problems. Using a form of unstructured curriculum learning, we then demonstrate that neural networks can be used to determine when to make use of these high-utility hints for specific I/O problems and thus enable complex functions to be successfully synthesised. We apply our approach to two different problem sets: common array-to-array programs (including sorting), and a canvas drawing problem set inspired by So & Oh (2018).
****

Attention-Based Clustering: Learning a Kernel from Context
Samuel Coward,Erik Visse-Martindale,Chithrupa Ramesh
In machine learning, no data point stands alone. We believe that context is an underappreciated concept in many machine learning methods. We propose Attention-Based Clustering (ABC), a neural architecture based on the attention mechanism, which is designed to learn latent representations that adapt to context within an input set, and which is inherently agnostic to input sizes and number of clusters. By learning a similarity kernel, our method directly combines with any out-of-the-box kernel-based clustering approach. We present competitive results for clustering Omniglot characters and include analytical evidence of the effectiveness of an attention-based approach for clustering.
****

Deep Curvature Suite
Diego Granziol,Xingchen Wan,Timur Garipov
The curvature of the loss, provides rich information on the geometry underlying neural networks, with applications in second order optimisation and Bayesian deep learning. However, accessing curvature information is still a daunting engineering challenge, inaccessible to most practitioners. We hence provide a software package the \textbf{Deep Curvature Suite}, which allows easy curvature evaluation for large modern neural networks. Beyond the calculation of a highly accurate moment matched approximation of the Hessian spectrum using Lanczos, our package provides: extensive \emph{loss surface visualisation}, the calculation of the \emph{Hessian variance} and \emph{stochastic second order optimisers}. We further address and disprove many common misconceptions in the literature about the Lanczos algorithm, namely that it learns eigenvalues from the top down. We prove using high dimensional concentration inequalities that for specific matrices a single random vector is sufficient for accurate spectral estimation, informing our spectral visualisation method. We showcase our package practical utility on a series of examples based on realistic modern neural networks such as the VGG-16 and Preactivated ResNets on the CIFAR-10/100 datasets. We further detail 3 specific potential use cases enabled by our software: research in stochastic second order optimisation for deep learning, learning rate scheduling using known optimality formulae for convex surfaces and empirical verification of deep learning theory based on comparing empirical and theoretically implied spectra.
****

Keep the Gradients Flowing: Using Gradient Flow to study Sparse Network Optimization
Kale-ab Tessera,Sara Hooker,Benjamin Rosman
Training sparse networks to converge to the same performance as dense neural architectures has proven to be elusive. Recent work suggests that initialization is the key. However, while this direction of research has had some success, focusing on initialization alone appears to be inadequate. In this paper, we take a broader view of training sparse networks and consider various choices made during training that might disadvantage sparse networks. We measure the gradient flow across different networks and datasets, and show that the default choices of opti

mizers, activation functions and regularizers used for dense networks can disadv
antage sparse networks. Based upon these findings, we show that gradient flow in
 sparse networks can be improved by reconsidering aspects of the architecture de
sign and the training regime. Our work suggests that initialization is only one
 piece of the puzzle and a wider view of tailoring optimization to sparse network
s yields promising results.
****

DHOG: Deep Hierarchical Object Grouping
Luke Nicholas Darlow,Amos Storkey
Unsupervised learning of categorical representations using data augmentations ap
pears to be a promising approach and has proven useful for finding suitable repr
esentations for downstream tasks. However current state-of-the-art methods requi
re preprocessing (e.g. Sobel edge detection) to work. We introduce a mutual info
rmation minimization strategy for unsupervised learning from augmentations, that
 prevents learning from locking on to easy to find, yet unimportant, representat
ions at the expense of more informative ones requiring more complex processing.
We demonstrate specifically that this process learns representations which captu
re higher mutual information between augmentations, and demonstrate that these r
epresentations are better suited to the downstream exemplar task of clustering.
We obtain substantial accuracy improvements on CIFAR-10, CIFAR-100-20, and SVHN.
****

Flatness is a False Friend
Diego Granziol
Hessian based measures of flatness, such as the trace, Frobenius and spectral no
rms, have been argued, used and shown to relate to generalisation. In this paper
 we demonstrate that, for feed-forward neural networks under the cross-entropy l
oss, low-loss solutions with large neural network weights have small Hessian bas
ed measures of flatness. This implies that solutions obtained without L2 regular
isation should be less sharp than those with despite generalising worse. We show
 this to be true for logistic regression, multi-layer perceptrons, simple convol
utional, pre-activated and wide residual networks on the MNIST and CIFAR-100 d
atasets. Furthermore, we show that adaptive optimisation algorithms using iterat
e averaging, on the VGG-16 network and CIFAR-100 dataset, achieve superior g
eneralisation to SGD but are $30 \times$ sharper. These theoretical and experime
ntal results further advocate the need to use flatness in conjunction with the w
eights scale to measure generalisation \citep{neyshabur2017exploring,dziugaite20
17computing}.
****

Exploring single-path Architecture Search ranking correlations
Kevin Alexander Laube,Andreas Zell
Recently presented benchmarks for Neural Architecture Search (NAS) provide the r
esults of training thousands of different architectures in a specific search spa
ce, thus enabling the fair and rapid comparison of different methods.
Based on these results, we quantify the ranking correlations of single-path arch
itecture search methods
in different search space subsets and under several training variations;
studying their impact on the expected search results.
The experiments support the few-shot approach and Linear Transformers,
provide evidence against disabling cell topology sharing during the training pha
se or using strong regularization in the NAS-Bench-201 search space,
and show the necessity of further research regarding super-network size and path
 sampling strategies.
****

Improved Gradient based Adversarial Attacks for Quantized Networks
Kartik Gupta,Thalaiyasingam Ajanthan
Neural network quantization has become increasingly popular due to efficient mem
ory consumption and faster computation resulting from bitwise operations on the
quantized networks. Even though they exhibit excellent generalization capabiliti
es, their robustness properties are not well-understood. In this work, we system
atically study the robustness of quantized networks against gradient based adver

sarial attacks and demonstrate that these quantized models suffer from gradient vanishing issues and show a fake sense of robustness. By attributing gradient vanishing to poor forward-backward signal propagation in the trained network, we introduce a simple temperature scaling approach to mitigate this issue while preserving the decision boundary. Despite being a simple modification to existing gradient based adversarial attacks, experiments on CIFAR-10/100 datasets with multiple network architectures demonstrate that our temperature scaled attacks obtain near-perfect success rate on quantized networks while outperforming original attacks on adversarially trained models as well as floating point networks.

****

Calibration of Neural Networks using Splines
Kartik Gupta,Amir Rahimi,Thalaiyasingam Ajanthan,Thomas Mensink,Cristian Sminchisescu,Richard Hartley
Calibrating neural networks is of utmost importance when employing them in safety-critical applications where the downstream decision making depends on the predicted probabilities. Measuring calibration error amounts to comparing two empirical distributions. In this work, we introduce a binning-free calibration measure inspired by the classical Kolmogorov-Smirnov (KS) statistical test in which the main idea is to compare the respective cumulative probability distributions. From this, by approximating the empirical cumulative distribution using a differentiable function via splines, we obtain a recalibration function, which maps the network outputs to actual (calibrated) class assignment probabilities. The spline-fitting is performed using a held-out calibration set and the obtained recalibration function is evaluated on an unseen test set. We tested our method against existing calibration approaches on various image classification datasets and our spline-based recalibration approach consistently outperforms existing methods on KS error as well as other commonly used calibration measures. Code is available online at https://github.com/kartikgupta-at-anu/spline-calibration.
****

Graph Joint Attention Networks
Tiantian He,Lu Bai,Yew-Soon Ong
Graph attention networks (GATs) have been recognized as powerful tools for learning in graph structured data. However, how to enable the attention mechanisms in GATs to smoothly consider both structural and feature information is still very challenging. In this paper, we propose Graph Joint Attention Networks (JATs) to address the aforementioned challenge. Different from previous attention-based graph neural networks (GNNs), JATs adopt novel joint attention mechanisms which can automatically determine the relative significance between node features and structural coefficients learned from graph subspace, when computing the attention scores. Therefore, representations concerning more structural properties can be inferred by JATs.
Besides, we theoretically analyze the expressive power of JATs and further propose an improved strategy for the joint attention mechanisms that enables JATs to reach the upper bound of expressive power which every message-passing GNN can ultimately achieve, i.e., 1-WL test. JATs can thereby be seen as most powerful message-passing GNNs. The proposed neural architecture has been extensively tested on widely used benchmarking datasets, including Cora, Cite, and Pubmed and has been compared with state-of-the-art GNNs for node classification tasks. Experimental results show that JATs achieve state-of-the-art performance on all the testing datasets.
****

Do Transformers Understand Polynomial Simplification?
Vishesh Agarwal,Somak Aditya,Navin Goyal
Recently researchers have demonstrated that Transformers can be trained to learn symbolic tasks such as solving integration and differential equations in an end-to-end fashion. In these setups, for an input symbolic expression, the Transformer predicts the final solution in a single step. Since such tasks may consist of a sequence of logical steps, question remains whether such networks have understood and learnt individual steps to reach the solution. To take a deeper look,

we consider the task of polynomial simplification. Polynomials can be written in a simple normal form as a sum of monomials which are ordered in a lexicographic order. For a polynomial which is not necessarily in this normal form, a sequence of simplification steps is applied to reach the fully simplified (i.e., in the normal form) polynomial. For this task, we describe a synthetic Polynomial data set generation algorithm which generates polynomials with unique proof steps. Then, we conduct an extensive analysis of the Transformer’s abilities to learn the polynomial simplification task along different dimensions.
****
Probing BERT in Hyperbolic Spaces
Boli Chen,Yao Fu,Guangwei Xu,Pengjun Xie,Chuanqi Tan,Mosha Chen,Liping Jing
Recently, a variety of probing tasks are proposed to discover linguistic properties learned in contextualized word embeddings. Many of these works implicitly assume these embeddings lay in certain metric spaces, typically the Euclidean space. This work considers a family of geometrically special spaces, the hyperbolic spaces, that exhibit better inductive biases for hierarchical structures and may better reveal linguistic hierarchies encoded in contextualized representations. We introduce a $\textit{Poincaré probe}$, a structural probe projecting these embeddings into a Poincaré subspace with explicitly defined hierarchies. We focus on two probing objectives: (a) dependency trees where the hierarchy is defined as head-dependent structures; (b) lexical sentiments where the hierarchy is defined as the polarity of words (positivity and negativity). We argue that a key desideratum of a probe is its sensitivity to the existence of linguistic structures. We apply our probes on BERT, a typical contextualized embedding model. In a syntactic subspace, our probe better recovers tree structures than Euclidean probes, revealing the possibility that the geometry of BERT syntax may not necessarily be Euclidean. In a sentiment subspace, we reveal two possible meta-embeddings for positive and negative sentiments and show how lexically-controlled contextualization would change the geometric localization of embeddings. We demonstrate the findings with our Poincaré probe via extensive experiments and visualization. Our results can be reproduced at https://github.com/FranxYao/PoincareProbe
****
Source-free Domain Adaptation via Distributional Alignment by Matching Batch Normalization Statistics
Masato Ishii,Masashi Sugiyama
In this paper, we propose a novel domain adaptation method for the source-free setting. In this setting, we cannot access source data during adaptation, while unlabeled target data and a model pretrained with source data are given. Due to lack of source data, we cannot directly match the data distributions between domains unlike typical domain adaptation algorithms. To cope with this problem, we propose utilizing batch normalization statistics stored in the pretrained model to approximate the distribution of unobserved source data. Specifically, we fix the classifier part of the model during adaptation and only fine-tune the remaining feature encoder part so that batch normalization statistics of the features extracted by the encoder match those stored in the fixed classifier. Additionally, we also maximize the mutual information between the features and the classifier's outputs to further boost the classification performance. Experimental results with several benchmark datasets show that our method achieves competitive performance with state-of-the-art domain adaptation methods even though it does not require access to source data.
****
Refining Deep Generative Models via Discriminator Gradient Flow
Abdul Fatir Ansari,Ming Liang Ang,Harold Soh
Deep generative modeling has seen impressive advances in recent years, to the point where it is now commonplace to see simulated samples (e.g., images) that closely resemble real-world data. However, generation quality is generally inconsistent for any given model and can vary dramatically between samples. We introduce Discriminator Gradient $f$low (DG$f$low), a new technique that improves generated samples via the gradient flow of entropy-regularized f-divergences between the real and the generated data distributions. The gradient flow takes the form

of a non-linear Fokker-Plank equation, which can be easily simulated by sampling from the equivalent McKean-Vlasov process. By refining inferior samples, our technique avoids wasteful sample rejection used by previous methods (DRS & MH-GAN). Compared to existing works that focus on specific GAN variants, we show our refinement approach can be applied to GANs with vector-valued critics and even other deep generative models such as VAEs and Normalizing Flows. Empirical results on multiple synthetic, image, and text datasets demonstrate that DG$f$low leads to significant improvement in the quality of generated samples for a variety of generative models, outperforming the state-of-the-art Discriminator Optimal Transport (DOT) and Discriminator Driven Latent Sampling (DDLS) methods.
****

Coping with Label Shift via Distributionally Robust Optimisation
Jingzhao Zhang,Aditya Krishna Menon,Andreas Veit,Srinadh Bhojanapalli,Sanjiv Kumar,Suvrit Sra
The label shift problem refers to the supervised learning setting where the train and test label distributions do not match. Existing work addressing label shift usually assumes access to an unlabelled test sample. This sample may be used to estimate the test label distribution, and to then train a suitably re-weighted classifier. While approaches using this idea have proven effective, their scope is limited as it is not always feasible to access the target domain; further, they require repeated retraining if the model is to be deployed in multiple test environments. Can one instead learn a single classifier that is robust to arbitrary label shifts from a broad family? In this paper, we answer this question by proposing a model that minimises an objective based on distributionally robust optimisation (DRO). We then design and analyse a gradient descent-proximal mirror ascent algorithm tailored for large-scale problems to optimise the proposed objective. Finally, through experiments on CIFAR-100 and ImageNet, we show that our technique can significantly improve performance over a number of baselines in settings where label shift is present.
****

Targeted VAE: Structured Inference and Targeted Learning for Causal Parameter Estimation
Matthew James Vowels,Necati Cihan Camgoz,Richard Bowden
Undertaking causal inference with observational data is extremely useful across a wide range of domains including the development of medical treatments, advertisements and marketing, and policy making. There are two main challenges associated with undertaking causal inference using observational data: treatment assignment heterogeneity (i.e., differences between the treated and untreated groups), and an absence of counterfactual data (i.e., not knowing what would have happened if an individual who did get treatment, were instead to have not been treated). We address these two challenges by combining structured inference and targeted learning. To our knowledge, Targeted Variational AutoEncoder (TVAE) is the first method to incorporate targeted learning into deep latent variable models. Results demonstrate competitive and state of the art performance.

****

Stochastic Optimization with Non-stationary Noise: The Power of Moment Estimation
Jingzhao Zhang,Hongzhou Lin,Subhro Das,Suvrit Sra,Ali Jadbabaie
We investigate stochastic optimization under weaker assumptions on the distribution of noise than those used in usual analysis. Our assumptions are motivated by empirical observations in training neural networks. In particular, standard results on optimal convergence rates for stochastic optimization assume either there exists a uniform bound on the moments of the gradient noise, or that the noise decays as the algorithm progresses. These assumptions do not match the empirical behavior of optimization algorithms used in neural network training where the noise level in stochastic gradients could even increase with time. We address this nonstationary behavior of noise by analyzing convergence rates of stochastic gradient methods subject to changing second moment (or variance) of the stochastic oracle. When the noise variation is known, we show that it is always benefic

ial to adapt the step-size and exploit the noise variability. When the noise sta
tistics are unknown, we obtain similar improvements by developing an online esti
mator of the noise level, thereby recovering close variants of RMSProp~\citep{ti
eleman2012lecture}. Consequently, our results reveal why adaptive step size meth
ods can outperform SGD, while still enjoying theoretical guarantees.
****
Learning Monotonic Alignments with Source-Aware GMM Attention
Tae Gyoon Kang,Ho-Gyeong Kim,Min-Joong Lee,Jihyun Lee,Seongmin Ok,Hoshik Lee,You
ng Sang Choi
Transformers with soft attention have been widely adopted in various sequence-to
-sequence (Seq2Seq) tasks. Whereas soft attention is effective for learning sema
ntic similarities between queries and keys based on their contents, it does not
explicitly model the order of elements in sequences which is crucial for monoton
ic Seq2Seq tasks. Learning monotonic alignments between input and output sequenc
es may be beneficial for long-form and online inference applications that are st
ill challenging for the conventional soft attention algorithm. Herein, we focus
on monotonic Seq2Seq tasks and propose a source-aware Gaussian mixture model att
ention in which the attention scores are monotonically calculated considering bo
th the content and order of the source sequence. We experimentally demonstrate t
hat the proposed attention mechanism improved the performance on the online and
long-form speech recognition problems without performance degradation in offline
 in-distribution speech recognition.
****
Learning to Generate Questions by Recovering Answer-containing Sentences
Seohyun Back,Akhil Kedia,Sai Chetan Chinthakindi,Haejun Lee,Jaegul Choo
To train a question answering model based on machine reading comprehension (MRC)
, significant effort is required to prepare annotated training data composed of
questions and their answers from contexts. To mitigate this issue, recent resear
ch has focused on synthetically generating a question from a given context and a
n annotated (or generated) answer by training an additional generative model, wh
ich can be utilized to augment the training data. In light of this research dire
ction, we propose a novel pre-training approach that learns to generate contextu
ally rich questions, by recovering answer-containing sentences. Our approach is
composed of two novel components, (1) dynamically determining K answers from a g
iven document and (2) pre-training the question generator on the task of generat
ing the answer-containing sentence. We evaluate our method against existing ones
 in terms of the quality of generated questions as well as the fine-tuned MRC mo
del accuracy after training on the data synthetically generated by our method. E
xperimental results demonstrate that our approach consistently improves the ques
tion generation capability of existing models such as T5 and UniLM, and shows st
ate-of-the-art results on MS MARCO and NewsQA, and comparable results to the sta
te-of-the-art on SQuAD. Additionally, we demonstrate that the data synthetically
 generated by our approach is beneficial for boosting up the downstream MRC accu
racy across a wide range of datasets, such as SQuAD-v1.1, v2.0, and KorQuAD, wit
hout any modification to the existing MRC models. Furthermore, our experiments h
ighlight that our method shines especially when a limited amount of training dat
a is given, in terms of both pre-training and downstream MRC data.
****
Semi-Supervised Learning of Multi-Object 3D Scene Representations
Cathrin Elich,Martin R. Oswald,Marc Pollefeys,Joerg Stueckler
Representing scenes at the granularity of objects is a prerequisite for scene un
derstanding and decision making. We propose a novel approach for learning multi-
object 3D scene representations from images. A recurrent encoder regresses a lat
ent representation of 3D shapes, poses and texture of each object from an input
RGB image. The 3D shapes are represented continuously in function-space as signe
d distance functions (SDF) which we efficiently pre-train from example shapes. B
y differentiable rendering, we train our model to decompose scenes self-supervis
ed from RGB-D images. Our approach learns to decompose images into the constitue
nt objects of the scene and to infer their shape, pose and texture properties fr
om a single view. In experiments, we evaluate the accuracy of our model in infer

ring the 3D scene layout and demonstrate the capabilities of the generative 3D s
cene model.
****
Variational State-Space Models for Localisation and Dense 3D Mapping in 6 DoF
Atanas Mirchev,Baris Kayalibay,Patrick van der Smagt,Justin Bayer
We solve the problem of 6-DoF localisation and 3D dense reconstruction in spatia
l environments as approximate Bayesian inference in a deep state-space model. Ou
r approach leverages both learning and domain knowledge from multiple-view geome
try and rigid-body dynamics. This results in an expressive predictive model of t
he world, often missing in current state-of-the-art visual SLAM solutions. The c
ombination of variational inference, neural networks and a differentiable raycas
ter ensures that our model is amenable to end-to-end gradient-based optimisation
. We evaluate our approach on realistic unmanned aerial vehicle flight data, nea
ring the performance of state-of-the-art visual-inertial odometry systems. We de
monstrate the applicability of the model to generative prediction and planning.
****
Non-greedy Gradient-based Hyperparameter Optimization Over Long Horizons
Paul Micaelli,Amos Storkey
Gradient-based meta-learning has earned a widespread popularity in few-shot lear
ning, but remains broadly impractical for tasks with long horizons (many gradien
t steps), due to memory scaling and gradient degradation issues. A common workar
ound is to learn meta-parameters online, but this introduces greediness which co
mes with a significant performance drop. In this work, we enable non-greedy meta
-learning of hyperparameters over long horizons by sharing hyperparameters that
are contiguous in time, and using the sign of hypergradients rather than their m
agnitude to indicate convergence. We implement this with forward-mode differenti
ation, which we extend to the popular momentum-based SGD optimizer. We demonstra
te that the hyperparameters of this optimizer can be learned non-greedily withou
t gradient degradation over $\sim 10^4$ inner gradient steps, by only requiring
~ 10 outer gradient steps. On CIFAR-10, we outperform greedy and random sea
rch methods for the same computational budget by nearly 10%. Code will be ava
ilable upon publication.
****
A Distributional Perspective on Actor-Critic Framework
Daniel Wontae Nam,Younghoon Kim,Chan Youn Park
Recent distributional reinforcement learning methods, despite their successes, s
till contain fundamental problems that can lead to inaccurate representations of
value distributions, such as distributional instability, action type restrictio
n, and conflation between samples and statistics. In this paper, we present a no
vel distributional actor-critic framework, GMAC, to address such problems. Adopt
ing a stochastic policy removes the first two problems, and the conflation in th
e approximation is alleviated by minimizing the Crame ■r distance between the va
lue distribution and its Bellman target distribution. In addition, GMAC improves
data efficiency by generating the Bellman target distribution through the Sampl
e-Replacement algorithm, denoted by SR(λ), which provides a distributional gener
alization of multi-step policy evaluation algorithms. We empirically show that o
ur method captures the multimodality of value distributions and improves the per
formance of a conventional actor-critic method with low computational cost in bo
th discrete and continuous action spaces, using Arcade Learning Environment (ALE
) and PyBullet environment.
****
Few-Shot Bayesian Optimization with Deep Kernel Surrogates
Martin Wistuba,Josif Grabocka
Hyperparameter optimization (HPO) is a central pillar in the automation of machi
ne learning solutions and is mainly performed via Bayesian optimization, where a
parametric surrogate is learned to approximate the black box response function
(e.g. validation error). Unfortunately, evaluating the response function is comp
utationally intensive. As a remedy, earlier work emphasizes the need for transfe
r learning surrogates which learn to optimize hyperparameters for an algorithm f
rom other tasks. In contrast to previous work, we propose to rethink HPO as a fe

w-shot learning problem in which we train a shared deep surrogate model to quick ly adapt (with few response evaluations) to the response function of a new task. We propose the use of a deep kernel network for a Gaussian process surrogate th at is meta-learned in an end-to-end fashion in order to jointly approximate the response functions of a collection of training data sets. As a result, the novel few-shot optimization of our deep kernel surrogate leads to new state-of-the-ar t results at HPO compared to several recent methods on diverse metadata sets.
****
Fully Convolutional Approach for Simulating Wave Dynamics
Mario Lino Valencia,Chris D Cantwell,Eduardo Pignatelli,Stathi Fotiadis,Anil Ant hony Bharath
We investigate the performance of fully convolutional networks to predict the mo tion and interaction of surface waves in open and closed complex geometries. We focus on a U-Net type architecture and assess its ability to capture and extrapo late wave propagation in time as well as the reflection, interference and diffra ction of waves. We investigate how well the network generalises both to long-ti me predictions and to geometric configurations not seen during training. We demo nstrate that this neural network is capable of accurately predicting the height distribution of waves on a liquid surface within curved and multi-faceted open a nd closed geometries, when only simple box and right-angled corner geometries we re seen during training. We found that the RMSE of the predictions remained of o rder 1×10^{-4} times the characteristic length of the domain for at least 20 time-steps.
****
Robust Learning via Golden Symmetric Loss of (un)Trusted Labels
Amirmasoud Ghiassi,Robert Birke,Lydia Y. Chen
Learning robust deep models against noisy labels becomes ever critical when toda y's data is commonly collected from open platforms and subject to adversarial co rruption. The information on the label corruption process, i.e., corruption matr ix, can greatly enhance the robustness of deep models but still fall behind in c ombating hard classes. In this paper, we propose to construct a golden symmetric loss (GSL) based on the estimated confusion matrix as to avoid overfitting to n oisy labels and learn effectively from hard classes. GSL is the weighted sum of the corrected regular cross entropy and reverse cross entropy. By leveraging a s mall fraction of trusted clean data, we estimate the corruption matrix and use i t to correct the loss as well as to determine the weights of GSL. We theoretical ly prove the robustness of the proposed loss function in the presence of dirty l abels. We provide a heuristics to adaptively tune the loss weights of GSL accord ing to the noise rate and diversity measured from the dataset. We evaluate our p roposed golden symmetric loss on both vision and natural language deep models su bject to different types of label noise patterns. Empirical results show that GS L can significantly outperform the existing robust training methods on different noise patterns, showing accuracy improvement up to 18% on CIFAR-100 and 1% on r eal world noisy dataset of Clothing1M.
****
Adaptive norms for deep learning with regularized Newton methods
Jonas K Kohler,Leonard Adolphs,Aurelien Lucchi
We investigate the use of regularized Newton methods with adaptive norms for opt imizing neural networks. This approach can be seen as a second-order counterpart of adaptive gradient methods, which we here show to be interpretable as first-o rder trust region methods with ellipsoidal constraints. In particular, we prove that the preconditioning matrix used in RMSProp and Adam satisfies the necessary conditions for provable convergence of second-order trust region methods with s tandard worst-case complexities on general non-convex objectives. Furthermore, w e run experiments across different neural architectures and datasets to find tha t the ellipsoidal constraints constantly outperform their spherical counterpart both in terms of number of backpropagations and asymptotic loss value. Finally, we find comparable performance to state-of-the-art first-order methods in terms of backpropagations, but further advances in hardware are needed to render Newto n methods competitive in terms of computational time.

****
Domain-Free Adversarial Splitting for Domain Generalization
Xiang Gu,Jiasun Feng,Jian Sun,Zongben Xu
Domain generalization is an approach that utilizes several source domains to tra in the learner to be generalizable to unseen target domain to tackle domain shif t issue. It has drawn much attention in machine learning community. This paper a ims to learn to generalize well to unseen target domain without relying on the k nowledge of the number of source domains and domain labels. To achieve that goal , we unify adversarial training and meta-learning in a novel proposed Domain-Fre e Adversarial Splitting (DFAS) framework. In this framework, we model the domain generalization as a learning problem that enforces the learner to be able to ge neralize well for any train/val subsets splitting of the training dataset. This model can be further transformed to be a min-max optimization problem which can be solved by an iterative adversarial training process. In each iteration, it ad versarially splits the training dataset into train/val subsets to maximize domai n shift between them using current learner, and then updates the learner on this splitting to be able to generalize well from train-subset to val-subset using m eta-learning approach. Extensive experiments on three benchmark datasets under t hree different settings on the source and target domains show that our method ac hieves state-of-the-art results and confirm the effectiveness of our method by a blation study. We also derive a generalization error bound for theoretical under standing of our method.
****
Neural Network Surgery: Combining Training with Topology Optimization
Elisabeth Schiessler,Roland Aydin,Kevin Linka,Christian Cyron
With ever increasing computational capacities, neural networks become more and m ore proficient at solving complex tasks. However, picking a sufficiently good ne twork topology usually relies on expert human knowledge. Neural architecture sea rch aims to reduce the extent of expertise that is needed. Modern architecture s earch techniques often rely on immense computational power, or apply trained met a controllers for decision making. We develop a framework for a genetic algorith m that is both computationally cheap and makes decisions based on mathematical c riteria rather than trained parameters. It is a hybrid approach that fuses train ing and topology optimization together into one process. Structural modification s that are performed include adding or removing layers of neurons, with some re-training applied to make up for incurred change in input-output behaviour. Our a nsatz is tested on both the SVHN and (augmented) CIFAR-10 datasets with limited computational overhead compared to training only the baseline. This algorithm ca n achieve a significant increase in accuracy (as compared to a fully trained bas eline), rescue insufficient topologies that in their current state are only able to learn to a limited extent, and dynamically reduce network size without loss in achieved accuracy.
****
i-Mix: A Domain-Agnostic Strategy for Contrastive Representation Learning
Kibok Lee,Yian Zhu,Kihyuk Sohn,Chun-Liang Li,Jinwoo Shin,Honglak Lee
Contrastive representation learning has shown to be effective to learn represent ations from unlabeled data. However, much progress has been made in vision domai ns relying on data augmentations carefully designed using domain knowledge. In t his work, we propose i-Mix, a simple yet effective domain-agnostic regularizatio n strategy for improving contrastive representation learning. We cast contrastiv e learning as training a non-parametric classifier by assigning a unique virtual class to each data in a batch. Then, data instances are mixed in both the input and virtual label spaces, providing more augmented data during training. In exp eriments, we demonstrate that i-Mix consistently improves the quality of learned representations across domains, including image, speech, and tabular data. Furt hermore, we confirm its regularization effect via extensive ablation studies acr oss model and dataset sizes. The code is available at https://github.com/kibok90 /imix.
****
ColdExpand: Semi-Supervised Graph Learning in Cold Start

Il-Jae Kwon,Kyoung-Woon On,Dong-Geon Lee,Byoung-Tak Zhang
Most real-world graphs are dynamic and eventually face the cold start problem. A fundamental question is how the new cold nodes acquire initial information in order to be adapted into the existing graph. Here we postulates the cold start problem as a fundamental issue in graph learning and propose a new learning setting, "``Expanded Semi-supervised Learning." In expanded semi-supervised learning we extend the original semi-supervised learning setting even to new cold nodes that are disconnected from the graph. To this end, we propose ColdExpand model that classifies the cold nodes based on link prediction with multiple goals to tackle. We experimentally prove that by adding additional goal to existing link prediction method, our method outperforms the baseline in both expanded semi-supervised link prediction (at most 24\%) and node classification tasks (at most 15%). To the best of our knowledge this is the first study to address expansion of semi-supervised learning to unseen nodes.
****
Information distance for neural network functions
Xiao Zhang,Dejing Dou,Ji Wu
We provide a practical distance measure in the space of functions parameterized by neural networks. It is based on the classical information distance, and we propose to replace the uncomputable Kolmogorov complexity with information measured by codelength of prequential coding. We also provide a method for directly estimating the expectation of such codelength with limited examples. Empirically, we show that information distance is invariant with respect to different parameterization of the neural networks. We also verify that information distance can faithfully reflect similarities of neural network functions. Finally, we applied information distance to investigate the relationship between neural network models, and demonstrate the connection between information distance and multiple characteristics and behaviors of neural networks.
****
Graph Information Bottleneck for Subgraph Recognition
Junchi Yu,Tingyang Xu,Yu Rong,Yatao Bian,Junzhou Huang,Ran He
Given the input graph and its label/property, several key problems of graph learning, such as finding interpretable subgraphs, graph denoising and graph compression, can be attributed to the fundamental problem of recognizing a subgraph of the original one. This subgraph shall be as informative as possible, yet contains less redundant and noisy structure. This problem setting is closely related to the well-known information bottleneck (IB) principle, which, however, has less been studied for the irregular graph data and graph neural networks (GNNs). In this paper, we propose a framework of Graph Information Bottleneck (GIB) for the subgraph recognition problem in deep graph learning. Under this framework, one can recognize the maximally informative yet compressive subgraph, named IB-subgraph. However, the GIB objective is notoriously hard to optimize, mostly due to the intractability of the mutual information of irregular graph data and the unstable optimization process. In order to tackle these challenges, we propose: i) a GIB objective based-on a mutual information estimator for the irregular graph data; ii) a bi-level optimization scheme to maximize the GIB objective; iii) a connectivity loss to stabilize the optimization process. We evaluate the properties of the IB-subgraph in three application scenarios: improvement of graph classification, graph interpretation and graph denoising. Extensive experiments demonstrate that the information-theoretic IB-subgraph enjoys superior graph properties.
****
Perturbation Type Categorization for Multiple ℓ_p Bounded Adversarial Robustness
Pratyush Maini,Xinyun Chen,Bo Li,Dawn Song
Despite the recent advances in $\textit{adversarial training}$ based defenses, deep neural networks are still vulnerable to adversarial attacks outside the perturbation type they are trained to be robust against. Recent works have proposed defenses to improve the robustness of a single model against the union of multiple perturbation types. However, when evaluating the model against each individua

l attack, these methods still suffer significant trade-offs compared to the ones specifically trained to be robust against that perturbation type. In this work, we introduce the problem of categorizing adversarial examples based on their ℓ_p perturbation types. Based on our analysis, we propose $\textit{PROTECTOR}$, a two-stage pipeline to improve the robustness against multiple perturbation types. Instead of training a single predictor, $\textit{PROTECTOR}$ first categorizes the perturbation type of the input, and then utilizes a predictor specifically trained against the predicted perturbation type to make the final prediction. We first theoretically show that adversarial examples created by different perturbation types constitute different distributions, which makes it possible to distinguish them. Further, we show that at test time the adversary faces a natural trade-off between fooling the perturbation type classifier and the succeeding predictor optimized with perturbation specific adversarial training. This makes it challenging for an adversary to plant strong attacks against the whole pipeline. In addition, we demonstrate the realization of this trade-off in deep networks by adding random noise to the model input at test time, enabling enhanced robustness against strong adaptive attacks. Extensive experiments on MNIST and CIFAR-10 show that $\textit{PROTECTOR}$ outperforms prior adversarial training based defenses by over 5%, when tested against the union of $\ell_1, \ell_2, \ell_\infty$ attacks.
****

Model information as an analysis tool in deep learning
Xiao Zhang,Di Hu,Xingjian Li,Dejing Dou,Ji Wu
Information-theoretic perspectives can provide an alternative dimension of analyzing the learning process and complements usual performance metrics. Recently several works proposed methods for quantifying information content in a model (which we refer to as "model information"). We demonstrate using model information as a general analysis tool to gain insight into problems that arise in deep learning. By utilizing model information in different scenarios with different control variables, we are able to adapt model information to analyze fundamental elements of learning, i.e., task, data, model, and algorithm. We provide an example in each domain that model information is used as a tool to provide new solutions to problems or to gain insight into the nature of the particular learning setting. These examples help to illustrate the versatility and potential utility of model information as an analysis tool in deep learning.
****

Cross-Probe BERT for Efficient and Effective Cross-Modal Search
TAN YU,Hongliang Fei,Ping Li
Inspired by the great success of BERT in NLP tasks, many text-vision BERT models emerged recently. Benefited from cross-modal attention, text-vision BERT models have achieved excellent performance in many language-vision tasks including text-image retrieval. Nevertheless, cross-modal attentions used in text-vision BERT models require too expensive computation cost when solving text-vision retrieval, which is impractical for large-scale search. In this work, we develop a novel architecture, Cross-Probe BERT. It relies on devised text and vision probes, and cross-modal attentions are conducted on text and vision probes. It takes lightweight computation cost, and meanwhile effectively exploits cross-modal attention. Systematic experiments conducted on two public benchmarks demonstrate state-of-the-art effectiveness and efficiency of the proposed method.
****

Learning Safe Policies with Cost-sensitive Advantage Estimation
Bingyi Kang,Shie Mannor,Jiashi Feng
Reinforcement Learning (RL) with safety guarantee is critical for agents performing tasks in risky environments. Recent safe RL algorithms, developed based on Constrained Markov Decision Process (CMDP), mostly take the safety requirement as additional constraints when learning to maximize the return. However, they usually make unnecessary compromises in return for safety and only learn sub-optimal policies, due to the inability of differentiating safe and unsafe state-actions with high rewards. To address this, we propose Cost-sensitive Advantage Estimation (CSAE), which is simple to deploy for policy optimization and effective for

guiding the agents to avoid unsafe state-actions by penalizing their advantage v
alue properly. Moreover, for stronger safety guarantees, we develop a Worst-case
 Constrained Markov Decision Process (WCMDP) method to augment CMDP by constrain
ing the worst-case safety cost instead of the average one. With CSAE and WCMDP,
we develop new safe RL algorithms with theoretical justifications on their benef
its for safety and performance of the obtained policies. Extensive experiments c
learly demonstrate the superiority of our algorithms in learning safer and bette
r agents under multiple settings.
****

Towards Adversarial Robustness of Bayesian Neural Network through Hierarchical V
ariational Inference

Byung-Kwan Lee,Youngjoon Yu,Yong Man Ro

Recent works have applied Bayesian Neural Network (BNN) to adversarial training,
 and shown the improvement of adversarial robustness via the BNN's strength of s
tochastic gradient defense. However, we have found that in general, the BNN lose
s its stochasticity after its training with the BNN's posterior. As a result, th
e lack of the stochasticity leads to weak regularization effect to the BNN, whic
h increases KL divergence in ELBO from variational inference. In this paper, we
propose an enhanced Bayesian regularizer through hierarchical variational infere
nce in order to boost adversarial robustness against gradient-based attack. Furt
hermore, we also prove that the proposed method allows the BNN's stochasticity t
o be elevated with the reduced KL divergence. Exhaustive experiment results demo
nstrate the effectiveness of the proposed method by showing the improvement of a
dversarial robustness, compared with adversarial training (Madry et al., 2018) a
nd adversarial-BNN (Liu et al., 2019) under PGD attack and EOT-PGD attack to the
 L_{∞} perturbation on CIFAR-10/100, STL-10, and Tiny-ImageNet.
****

Rethinking Positional Encoding in Language Pre-training

Guolin Ke,Di He,Tie-Yan Liu

In this work, we investigate the positional encoding methods used in language pr
e-training (e.g., BERT) and identify several problems in the existing formulatio
ns. First, we show that in the absolute positional encoding, the addition operat
ion applied on positional embeddings and word embeddings brings mixed correlatio
ns between the two heterogeneous information resources. It may bring unnecessary
 randomness in the attention and further limit the expressiveness of the model.
 Second, we question whether treating the position of the symbol \texttt{[CLS]}
the same as other words is a reasonable design, considering its special role (th
e representation of the entire sentence) in the downstream tasks. Motivated from
 above analysis, we propose a new positional encoding method called \textbf{T}ra
nsformer with \textbf{U}ntied \textbf{P}ositional \textbf{E}ncoding (TUPE). In t
he self-attention module, TUPE computes the word contextual correlation and posi
tional correlation separately with different parameterizations and then adds the
m together. This design removes the mixed and noisy correlations over heterogene
ous embeddings and offers more expressiveness by using different projection matr
ices. Furthermore, TUPE unties the \texttt{[CLS]} symbol from other positions, m
aking it easier to capture information from all positions. Extensive experiments
 and ablation studies on GLUE benchmark demonstrate the effectiveness of the pro
posed method. Codes and models are released at \url{https://github.com/guolinke/
TUPE}.
****

Semi-supervised regression with skewed data via adversarially forcing the distri
bution of predicted values

Dae-Woong Jeong,Kiyoung Kim,Changyoung Park,Sehui Han,Woohyung Lim

Advances in scientific fields including drug discovery or material design are ac
companied by numerous trials and errors. However, generally only representative
experimental results are reported. Because of this reporting bias, the distribut
ion of labeled result data can deviate from their true distribution. A regressio
n model can be erroneous if it is built on these skewed data. In this work, we p
ropose a new approach to improve the accuracy of regression models that are trai
ned using a skewed dataset. The method forces the regression outputs to follow t

he true distribution; the forcing algorithm regularizes the regression results w hile keeping the information of the training data. We assume the existence of en ough unlabeled data that follow the true distribution, and that the true distrib ution can be roughly estimated from domain knowledge or a few samples. During tr aining neural networks to generate a regression model, an adversarial network is used to force the distribution of predicted values to follow the estimated ‘tru e’ distribution. We evaluated the proposed approach on four real-world datasets (pLogP, Diamond, House, Elevators). In all four datasets, the proposed approach reduced the root mean squared error of the regression by around 55 percent to 75 percent compared to regression models without adjustment of the distribution.
****
TEAC: Intergrating Trust Region and Max Entropy Actor Critic for Continuous Cont rol
Hongyu Zang,Xin Li,Li Zhang,Peiyao Zhao,Mingzhong Wang
Trust region methods and maximum entropy methods are two state-of-the-art branch es used in reinforcement learning (RL) for the benefits of stability and explora tion in continuous environments, respectively. This paper proposes to integrate both branches in a unified framework, thus benefiting from both sides. We first transform the original RL objective to a constraint optimization problem and the n proposes trust entropy actor-critic (TEAC), an off-policy algorithm to learn s table and sufficiently explored policies for continuous states and actions. TEAC trains the critic by minimizing the refined Bellman error and updates the actor by minimizing KL-divergence loss derived from the closed-form solution to the L agrangian.
We prove that the policy evaluation and policy improvement in TEAC is guaranteed to converge.
We compare TEAC with 4 state-of-the-art solutions on 6 tasks in the MuJoCo envir onment. The results show that TEAC outperforms state-of-the-art solutions in ter ms of efficiency and effectiveness.

****
Don't stack layers in graph neural networks, wire them randomly
Diego Valsesia,Giulia Fracastoro,Enrico Magli
Graph neural networks have become a staple in problems addressing learning and a nalysis of data defined over graphs. However, several results suggest an inheren t difficulty in extracting better performance by increasing the number of layers . Besides the classic vanishing gradient issues, recent works attribute this to a phenomenon peculiar to the extraction of node features in graph-based tasks, i .e., the need to consider multiple neighborhood sizes at the same time and adapt ively tune them. In this paper, we investigate the recently proposed randomly wi red architectures in the context of graph neural networks. Instead of building d eeper networks by stacking many layers, we prove that employing a randomly-wired architecture can be a more effective way to increase the capacity of the networ k and obtain richer representations. We show that such architectures behave like an ensemble of paths, which are able to merge contributions from receptive fiel ds of varied size. Moreover, these receptive fields can also be modulated to be wider or narrower through the trainable weights over the paths. We also provide extensive experimental evidence of the superior performance of randomly wired ar chitectures over three tasks and five graph convolution definitions, using a rec ent benchmarking framework that addresses the reliability of previous testing me thodologies.
****
Practical Massively Parallel Monte-Carlo Tree Search Applied to Molecular Design
Xiufeng Yang,Tanuj Aasawat,Kazuki Yoshizoe
It is common practice to use large computational resources to train neural netwo rks, known from many examples, such as reinforcement learning applications. Howe ver, while massively parallel computing is often used for training models, it is rarely used to search solutions for combinatorial optimization problems. This p aper proposes a novel massively parallel Monte-Carlo Tree Search (MP-MCTS) algor ithm that works efficiently for a 1,000 worker scale on a distributed memory env

ironment using multiple compute nodes and applies it to molecular design. This p aper is the first work that applies distributed MCTS to a real-world and non-gam e problem. Existing works on large-scale parallel MCTS show efficient scalabilit y in terms of the number of rollouts up to 100 workers. Still, they suffer from the degradation in the quality of the solutions. MP-MCTS maintains the search qu ality at a larger scale. By running MP-MCTS on 256 CPU cores for only 10 minutes , we obtained candidate molecules with similar scores to non-parallel MCTS runni ng for 42 hours. Moreover, our results based on parallel MCTS (combined with a s imple RNN model) significantly outperform existing state-of-the-art work. Our me thod is generic and is expected to speed up other applications of MCTS.
****

Causal Curiosity: RL Agents Discovering Self-supervised Experiments for Causal R epresentation Learning
Sumedh Anand Sontakke,Arash Mehrjou,Theofanis Karaletsos,Laurent Itti,Bernhard S chölkopf
Humans show an innate ability to learn the regularities of the world through int eraction. By performing experiments in our environment, we are able to discern t he causal factors of variation and infer how they affect the dynamics of our wor ld. Analogously, here we attempt to equip reinforcement learning agents with the ability to perform experiments that facilitate a categorization of the rolled-o ut trajectories, and to subsequently infer the causal factors of the environment in a hierarchical manner. We introduce a novel intrinsic reward, called causal curiosity, and show that it allows our agents to learn optimal sequences of acti ons, and to discover causal factors in the dynamics. The learned behavior allows the agent to infer a binary quantized representation for the ground-truth causa l factors in every environment. Additionally, we find that these experimental be haviors are semantically meaningful (e.g., to differentiate between heavy and li ght blocks, our agents learn to lift them), and are learnt in a self-supervised manner with approximately 2.5 times less data than conventional supervised plann ers. We show that these behaviors can be re-purposed and fine-tuned (e.g., from lifting to pushing or other downstream tasks). Finally, we show that the knowled ge of causal factor representations aids zero-shot learning for more complex tas ks.
****

Improving Neural Network Accuracy and Calibration Under Distributional Shift wit h Prior Augmented Data
Jeffrey Ryan Willette,Juho Lee,Sung Ju Hwang
Neural networks have proven successful at learning from complex data distributio ns by acting as universal function approximators. However, neural networks are o ften overconfident in their predictions, which leads to inaccurate and miscalibr ated probabilistic predictions. The problem of overconfidence becomes especially apparent in cases where the test-time data distribution differs from that which was seen during training. We propose a solution to this problem by seeking out regions in arbitrary feature space where the model is unjustifiably overconfiden t, and conditionally raising the entropy of those predictions towards that of th e Bayesian prior on the distribution of the labels. Our method results in a bett er calibrated network and is agnostic to the underlying model structure, so it c an be applied to any neural network which produces a probability density as an o utput. We demonstrate the effectiveness of our method and validate its performan ce on both classification and regression problems by applying it to the training of recent state-of-the-art neural network models.
****

Near-Black-Box Adversarial Attacks on Graph Neural Networks as An Influence Maxi mization Problem
Jiaqi Ma,Junwei Deng,Qiaozhu Mei
Graph neural networks (GNNs) have attracted increasing interests. With broad dep loyments of GNNs in real-world applications, there is an urgent need for underst anding the robustness of GNNs under adversarial attacks, especially in realistic setups. In this work, we study the problem of attacking GNNs in a restricted ne ar-black-box setup, by perturbing the features of a small set of nodes, with no

access to model parameters and model predictions. Our formal analysis draws a connection between this type of attacks and an influence maximization problem on the graph. This connection not only enhances our understanding on the problem of adversarial attack on GNNs, but also allows us to propose a group of effective near-black-box attack strategies. Our experiments verify that the proposed strategies significantly degrade the performance of three popular GNN models and outperform baseline adversarial attack strategies.
****

Augmenting Physical Models with Deep Networks for Complex Dynamics Forecasting
Yuan Yin,Vincent LE GUEN,Jérémie DONA,Emmanuel de Bezenac,Ibrahim Ayed,Nicolas THOME,patrick gallinari
Forecasting complex dynamical phenomena in settings where only partial knowledge of their dynamics is available is a prevalent problem across various scientific fields. While purely data-driven approaches are arguably insufficient in this context, standard physical modeling based approaches tend to be over-simplistic, inducing non-negligible errors. In this work, we introduce the APHYNITY framework, a principled approach for augmenting incomplete physical dynamics described by differential equations with deep data-driven models. It consists in decomposing the dynamics into two components: a physical component accounting for the dynamics for which we have some prior knowledge, and a data-driven component accounting for errors of the physical model. The learning problem is carefully formulated such that the physical model explains as much of the data as possible, while the data-driven component only describes information that cannot be captured by the physical model, no more, no less. This not only provides the existence and uniqueness for this decomposition, but also ensures interpretability and benefits generalization. Experiments made on three important use cases, each representative of a different family of phenomena, i.e. reaction-diffusion equations, wave equations and the non-linear damped pendulum, show that APHYNITY can efficiently leverage approximate physical models to accurately forecast the evolution of the system and correctly identify relevant physical parameters.
****

DyHCN: Dynamic Hypergraph Convolutional Networks
Nan Yin,zhigang luo,wenjie wang,Fuli Feng,Xiang Zhang
Hypergraph Convolutional Network (HCN) has become a default choice for capturing high-order relations among nodes, \emph{i.e., } encoding the structure of a hypergraph. However, existing HCN models ignore the dynamic evolution of hypergraphs in the real-world scenarios, \emph{i.e., } nodes and hyperedges in a hypergraph change dynamically over time. To capture the evolution of high-order relations and facilitate relevant analytic tasks, we formulate dynamic hypergraph and devise the Dynamic Hypergraph Convolution Networks (DyHCN). In general, DyHCN consists of a Hypergraph Convolution (HC) to encode the hypergraph structure at a time point and a Temporal Evolution module (TE) to capture the varying of the relations. The HC is delicately designed by equipping inner attention and outer attention, which adaptively aggregate nodes' features to hyperedge and estimate the importance of each hyperedge connected to the centroid node, respectively. Extensive experiments on the Tiigo and Stocktwits datasets show that DyHCN achieves superior performance over existing methods, which implies the effectiveness of capturing the property of dynamic hypergraphs by HC and TE modules.
****

Sobolev Training for the Neural Network Solutions of PDEs
Hwijae Son,Jin Woo Jang,Woo Jin Han,Hyung Ju Hwang
Approximating the numerical solutions of partial differential equations (PDEs) using neural networks is a promising application of deep learning. The smooth architecture of a fully connected neural network is appropriate for finding the solutions of PDEs; the corresponding loss function can also be intuitively designed and guarantees the convergence for various kinds of PDEs. However, the rate of convergence has been considered as a weakness of this approach. This paper introduces a novel loss function for the training of neural networks to find the solutions of PDEs, making the training substantially efficient. Inspired by the recent studies that incorporate derivative information for the training of neural ne

tworks, we develop a loss function that guides a neural network to reduce the er ror in the corresponding Sobolev space. Surprisingly, a simple modification of t he loss function can make the training process similar to Sobolev Training altho ugh solving PDEs with neural networks is not a fully supervised learning task. W e provide several theoretical justifications for such an approach for the viscou s Burgers equation and the kinetic Fokker--Planck equation. We also present seve ral simulation results, which show that compared with the traditional L^2 loss function, the proposed loss function guides the neural network to a significant ly faster convergence. Moreover, we provide the empirical evidence that shows th at the proposed loss function, together with the iterative sampling techniques, performs better in solving high dimensional PDEs.
****

OFFER PERSONALIZATION USING TEMPORAL CONVOLUTION NETWORK AND OPTIMIZATION
Ankur Verma
Lately, personalized marketing has become important for retail/e-retail firms du e to significant rise in online shopping and market competition. Increase in onl ine shopping and high market competition has led to an increase in promotional e xpenditure for online retailers, and hence, rolling out optimal offers has becom e imperative to maintain balance between number of transactions and profit. In t his paper, we propose our approach to solve the offer optimization problem at th e intersection of consumer, item and time in retail setting. To optimize offer, we first build a generalized non-linear model using Temporal Convolutional Netwo rk to predict the item purchase probability at consumer level for the given time period. Secondly, we establish the functional relationship between historical o ffer values and purchase probabilities obtained from the model, which is then us ed to estimate offer-elasticity of purchase probability at consumer item granula rity. Finally, using estimated elasticities, we optimize offer values using cons traint based optimization technique. This paper describes our detailed methodolo gy and presents the results of modelling and optimization across categories.
****

Manifold-aware Training: Increase Adversarial Robustness with Feature Clustering
Ting-An Yen,Chun-Shien Lu,Pau-Choo Chung
The problem of defending against adversarial attacks has attracted increasing at tention in recent years. While various types of defense methods ($\textit{e.g.}$, adversarial training, detection and rejection, and recovery) were proven empir ically to bring robustness to the network, their weakness was shown by later wor ks. Inspired by the observation from the distribution properties of the features extracted by the CNNs in the feature space and their link to robustness, this w ork designs a novel training process called Manifold-Aware Training (MAT), which forces CNNs to learn compact features to increase robustness. The effectivenes s of the proposed method is evaluated via comparisons with existing defense mech anisms, $\textit{i.e.}$, the TRADES algorithm, which has been recognized as a re presentative state-of-the-art technology, and the MMC method, which also aims to learn compact features. Further verification is also conducted using the attack adaptive to our method. Experimental results show that MAT-trained CNNs exhibit significantly higher performance than state-of-the-art robustness.
****

Sensory Resilience based on Synesthesia
Eric Platon,Tom Sonoda
Situated cognition depends on accessing environmental state through sensors. Eng ineering and cost constraints usually lead to limited “pathways” where, for exam ple, a vision sub-system only includes a camera and the software to deal with it . This traditional and rational design style entails any hardware defect on the pathway causes the system to grind to a halt until repair. We propose a “sensori plexer” as drop-in neural component architecture to address this issue, under th e common scenario of multiple sensors availability. This component architecture learns to mix and relate pathways, such that an agent facing failure in a sensor y sub-system can degrade gracefully and coherently by relying on its other sub- systems. The architecture is inspired by the concept of synesthesia, and relies on statistical coupling between sensor signals. We show the benefit and limitati

on of the architecture on a simple shape recognition and a more complex emotion recognition scenarios.
****
When does preconditioning help or hurt generalization?
Shun-ichi Amari,Jimmy Ba,Roger Baker Grosse,Xuechen Li,Atsushi Nitanda,Taiji Suzuki,Denny Wu,Ji Xu
While second order optimizers such as natural gradient descent (NGD) often speed up optimization, their effect on generalization has been called into question. This work presents a more nuanced view on how the \textit{implicit bias} of optimizers affects the comparison of generalization properties.
We provide an exact asymptotic bias-variance decomposition of the generalization error of preconditioned ridgeless regression in the overparameterized regime, and consider the inverse population Fisher information matrix (used in NGD) as a particular example. We determine the optimal preconditioner $\boldsymbol{P}$ for both the bias and variance, and find that the relative generalization performance of different optimizers depends on label noise and ``shape'' of the signal (true parameters): when the labels are noisy, the model is misspecified, or the signal is misaligned with the features, NGD can achieve lower risk; conversely, GD generalizes better under clean labels, a well-specified model, or aligned signal.
Based on this analysis, we discuss several approaches to manage the bias-variance tradeoff, and the potential benefit of interpolating between first- and second-order updates. We then extend our analysis to regression in the reproducing kernel Hilbert space and demonstrate that preconditioning can lead to more efficient decrease in the population risk. Lastly, we empirically compare the generalization error of first- and second-order optimizers in neural network experiments, and observe robust trends matching our theoretical analysis.
****
SketchEmbedNet: Learning Novel Concepts by Imitating Drawings
Alexander Wang,Mengye Ren,Richard Zemel
Sketch drawings are an intuitive visual domain that appeals to human instinct. Previous work has shown that recurrent neural networks are capable of producing sketch drawings of a single or few classes at a time. In this work we investigate representations developed by training a generative model to produce sketches from pixel images across many classes in a sketch domain. We find that the embeddings learned by this sketching model are extremely informative for visual tasks and infer a unique visual understanding. We then use them to exceed state-of-the-art performance in unsupervised few-shot classification on the Omniglot and mini-ImageNet benchmarks. We also leverage the generative capacity of our model to produce high quality sketches of novel classes based on just a single example.
****
Data augmentation as stochastic optimization
Boris Hanin,Yi Sun
We present a theoretical framework recasting data augmentation as stochastic optimization for a sequence of time-varying proxy losses. This provides a unified language for understanding techniques commonly thought of as data augmentation, including synthetic noise and label-preserving transformations, as well as more traditional ideas in stochastic optimization such as learning rate and batch size scheduling. We then specialize our framework to study arbitrary augmentations in the context of a simple model (overparameterized linear regression). We extend in this setting the classical Monro-Robbins theorem to include augmentation and obtain rates of convergence, giving conditions on the learning rate and augmentation schedule under which augmented gradient descent converges. Special cases give provably good schedules for augmentation with additive noise, minibatch SGD, and minibatch SGD with noise.
****
A Good Image Generator Is What You Need for High-Resolution Video Synthesis
Yu Tian,Jian Ren,Menglei Chai,Kyle Olszewski,Xi Peng,Dimitris N. Metaxas,Sergey Tulyakov
Image and video synthesis are closely related areas aiming at generating content

from noise. While rapid progress has been demonstrated in improving image-based models to handle large resolutions, high-quality renderings, and wide variations in image content, achieving comparable video generation results remains problematic. We present a framework that leverages contemporary image generators to render high-resolution videos. We frame the video synthesis problem as discovering a trajectory in the latent space of a pre-trained and fixed image generator. Not only does such a framework render high-resolution videos, but it also is an order of magnitude more computationally efficient. We introduce a motion generator that discovers the desired trajectory, in which content and motion are disentangled. With such a representation, our framework allows for a broad range of applications, including content and motion manipulation. Furthermore, we introduce a new task, which we call cross-domain video synthesis, in which the image and motion generators are trained on disjoint datasets belonging to different domains. This allows for generating moving objects for which the desired video data is not available. Extensive experiments on various datasets demonstrate the advantages of our methods over existing video generation techniques. Code will be released at https://github.com/snap-research/MoCoGAN-HD.
****

Out-of-Distribution Generalization Analysis via Influence Function
Haotian Ye,Chuanlong Xie,Yue Liu,Zhenguo Li
The mismatch between training dataset and target environment is one major challenge for current machine learning systems. When training data is collected from multiple environments and the the evaluation is on any new environment, we are facing an Out-of-Distribution (OOD) generalization problem that aims to find a model with the best OOD accuracy, i.e. the best worst-environment accuracy. However, with limited access to environments, the worst environment may be unseen, and test accuracy is a biased estimate of OOD accuracy. In this paper, we show that test accuracy may dramatically fail to identify OOD accuracy and mislead the tuning procedure. To this end, we introduce Influence Function, a classical tool from robust statistics, into the OOD generalization problem and suggest the variance of influence function to measure the stability of a model on training environments. We show that the proposed index and test accuracy together can help us discern whether OOD algorithms are needed and whether a model achieves good OOD generalization.
****

Learning to Solve Nonlinear Partial Differential Equation Systems To Accelerate MOSFET Simulation
Seungcheol Han,Jonghyun Choi,Sung-Min Hong
Semiconductor device simulation uses numerical analysis, where a set of coupled nonlinear partial differential equations is solved with the iterative Newton-Raphson method. Since an appropriate initial guess to start the Newton-Raphson method is not available, a solution of practical importance with desired boundary conditions cannot be trivially achieved. Instead, several solutions with intermediate boundary conditions should be calculated to address the nonlinearity and introducing intermediate boundary conditions significantly increases the computation time. In order to accelerate the semiconductor device simulation, we propose to use a neural network to learn an approximate solution for desired boundary conditions. With an initial solution sufficiently close to the final one by a trained neural network, computational cost to calculate several unnecessary solutions is significantly reduced. Specifically, a convolutional neural network for MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor), the most widely used semiconductor device, are trained in a supervised manner to compute the initial solution. Particularly, we propose to consider device grids with varying size and spacing and derive a compact expression of the solution based upon the electrostatic potential. We empirically show that the proposed method accelerates the simulation by more than 12 times. Results from the local linear regression and a fully-connected network are compared and extension to a complex two-dimensional domain is sketched.
****

A Surgery of the Neural Architecture Evaluators

Xuefei Ning,Wenshuo Li,Zixuan Zhou,Tianchen Zhao,Shuang Liang,Yin Zheng,Huazhong Yang,Yu Wang

Neural architecture search (NAS) has recently received extensive attention due to its effectiveness in automatically designing effective neural architectures. A major challenge in NAS is to conduct a fast and accurate evaluation of neural architectures. Commonly used fast architecture evaluators include parameter-sharing ones and predictor-based ones. Despite their high evaluation efficiency, the evaluation correlation (especially of the well-performing architectures) is still questionable. In this paper, we conduct an extensive assessment of both the parameter-sharing and predictor-based evaluators on the NAS-Bench-201 search space, and break up how and why different configurations and strategies influence the fitness of the evaluators. Specifically, we carefully develop a set of NAS-oriented criteria to understand the behavior of fast architecture evaluators in different training stages. And based on the findings of our experiments, we give pieces of knowledge and suggestions to guide NAS application and motivate further research.

****

Transformer-QL: A Step Towards Making Transformer Network Quadratically Large

Suvadeep Hajra

Transformer networks have shown outstanding performance on many natural language processing tasks. However the context length (the number of previous tokens on which the output states depend) of a Transformer network grows at best linearly with the memory and computational power used. This limitation prevents a transformer network to have very long context in a resource limited application. In this work, we propose a class of transformer networks, namely Transformer-QL (\bf{Q}uadratically \bf{L}arge), in which, the context length can grow at best quadratically with the memory and computational power used. We have empirically evaluated a Transformer-QL model in three long range language modeling datasets. The results show that Transformer-QL can provide significant improvements over other state of the art networks.

****

Embedding a random graph via GNN: mean-field inference theory and RL applications to NP-Hard multi-robot/machine scheduling

HYUNWOOK KANG,SEUNGWOO SCHIN,James Morrison,Jinkyoo Park

We develop a theory for embedding a random graph using graph neural networks (GNN) and illustrate its capability to solve NP-hard scheduling problems. We apply the theory to address the challenge of developing a near-optimal learning algorithm to solve the NP-hard problem of scheduling multiple robots/machines with time-varying rewards. In particular, we consider a class of reward collection problems called Multi-Robot Reward Collection (MRRC). Such MRRC problems well model ride-sharing, pickup-and-delivery, and a variety of related problems. We consider the classic identical parallel machine scheduling problem (IPMS) in the Appendix.

For the theory, we first observe that MRRC system state can be represented as an extension of probabilistic graphical models (PGMs), which we refer to as random PGMs. We then develop a mean-field inference method for random PGMs.
We prove that a simple modification of a typical GNN embedding is sufficient to embed a random graph even when the edge presence probabilities are interdependent.

Our theory enables a two-step hierarchical inference for precise and transferable Q-function estimation for MRRC and IPMS. For scalable computation, we show that the transferability of Q-function estimation enables us to design a polynomial-time algorithm with 1-1/e optimality bound.
Experimental results on solving NP-hard MRRC problems (and IMPS in the Appendix) highlight the near-optimality and transferability of the proposed methods.

****

ARMOURED: Adversarially Robust MOdels using Unlabeled data by REgularizing Diver

sity
Kangkang Lu,Cuong Manh Nguyen,Xun Xu,Kiran Krishnamachari,Yu Jing Goh,Chuan-Shen g Foo
Adversarial attacks pose a major challenge for modern deep neural networks. Rece nt advancements show that adversarially robust generalization requires a large a mount of labeled data for training. If annotation becomes a burden, can unlabele d data help bridge the gap? In this paper, we propose ARMOURED, an adversarially robust training method based on semi-supervised learning that consists of two c omponents. The first component applies multi-view learning to simultaneously opt imize multiple independent networks and utilizes unlabeled data to enforce label ing consistency. The second component reduces adversarial transferability among the networks via diversity regularizers inspired by determinantal point processe s and entropy maximization. Experimental results show that under small perturbat ion budgets, ARMOURED is robust against strong adaptive adversaries. Notably, AR MOURED does not rely on generating adversarial samples during training. When use d in combination with adversarial training, ARMOURED yields competitive performa nce with the state-of-the-art adversarially-robust benchmarks on SVHN and outper forms them on CIFAR-10, while offering higher clean accuracy.
****
Optimizing Loss Functions Through Multivariate Taylor Polynomial Parameterizatio n
Santiago Gonzalez,Risto Miikkulainen
Metalearning of deep neural network (DNN) architectures and hyperparameters has become an increasingly important area of research. Loss functions are a type of metaknowledge that is crucial to effective training of DNNs, however, their pote ntial role in metalearning has not yet been fully explored. Whereas early work f ocused on genetic programming (GP) on tree representations, this paper proposes continuous CMA-ES optimization of multivariate Taylor polynomial parameterizatio ns. This approach, TaylorGLO, makes it possible to represent and search useful l oss functions more effectively. In MNIST, CIFAR-10, and SVHN benchmark tasks, Ta ylorGLO finds new loss functions that outperform functions previously discovered through GP, as well as the standard cross-entropy loss, in fewer generations. T hese functions serve to regularize the learning task by discouraging overfitting to the labels, which is particularly useful in tasks where limited training dat a is available. The results thus demonstrate that loss function optimization is a productive new avenue for metalearning.
****
Learning Energy-Based Generative Models via Coarse-to-Fine Expanding and Samplin g
Yang Zhao,Jianwen Xie,Ping Li
Energy-based models (EBMs) parameterized by neural networks can be trained by th e Markov chain Monte Carlo (MCMC) sampling-based maximum likelihood estimation. Despite the recent significant success of EBMs in image generation, the current approaches to train EBMs are unstable and have difficulty synthesizing diverse a nd high-fidelity images. In this paper, we propose to train EBMs via a multistag e coarse-to-fine expanding and sampling strategy, which starts with learning a c oarse-level EBM from images at low resolution and then gradually transits to lea rn a finer-level EBM from images at higher resolution by expanding the energy fu nction as the learning progresses. The proposed framework is computationally eff icient with smooth learning and sampling. It achieves the best performance on im age generation amongst all EBMs and is the first successful EBM to synthesize hi gh-fidelity images at 512×512 resolution. It can also be useful for image restoration and out-of-distribution detection. Lastly, the proposed framework i s further generalized to the one-sided unsupervised image-to-image translation a nd beats baseline methods in terms of model size and training budget. We also pr esent a gradient-based generative saliency method to interpret the translation d ynamics.
****
Undistillable: Making A Nasty Teacher That CANNOT teach students
Haoyu Ma,Tianlong Chen,Ting-Kuei Hu,Chenyu You,Xiaohui Xie,Zhangyang Wang

Knowledge Distillation (KD) is a widely used technique to transfer knowledge fro m pre-trained teacher models to (usually more lightweight) student models. Howe ver, in certain situations, this technique is more of a curse than a blessing. F or instance, KD poses a potential risk of exposing intellectual properties (IPs) : even if a trained machine learning model is released in ``black boxes'' (e.g., as executable software or APIs without open-sourcing code), it can still be rep licated by KD through imitating input-output behaviors. To prevent this unwanted effect of KD, this paper introduces and investigates a concept called $\textit{ Nasty Teacher}$: a specially trained teacher network that yields nearly the same performance as a normal one, but would significantly degrade the performance of student models learned by imitating it. We propose a simple yet effective algor ithm to build the nasty teacher, called $\textit{self-undermining knowledge dist illation}$. Specifically, we aim to maximize the difference between the output o f the nasty teacher and a normal pre-trained network. Extensive experiments on s everal datasets demonstrate that our method is effective on both standard KD and data-free KD, providing the desirable KD-immunity to model owners for the first time. We hope our preliminary study can draw more awareness and interest in thi s new practical problem of both social and legal importance. Our codes and pre-t rained models can be found at: $\url{https://github.com/VITA-Group/Nasty-Teacher }$.

****

An information-theoretic framework for learning models of instance-independent l abel noise

Xia Huang,Kai Fong Ernest Chong

Given a dataset $\mathcal{D}$ with label noise, how do we learn its underlying n oise model? If we assume that the label noise is instance-independent, then the noise model can be represented by a noise transition matrix $Q_{\mathcal{D}}$. R ecent work has shown that even without further information about any instances w ith correct labels, or further assumptions on the distribution of the label nois e, it is still possible to estimate $Q_{\mathcal{D}}$ while simultaneously learn ing a classifier from $\mathcal{D}$. However, this presupposes that a good estim ate of $Q_{\mathcal{D}}$ requires an accurate classifier. In this paper, we show that high classification accuracy is actually not required for estimating $Q_{\ mathcal{D}}$ well. We shall introduce an information-theoretic-based framework f or estimating $Q_{\mathcal{D}}$ solely from $\mathcal{D}$ (without additional in formation or assumptions). At the heart of our framework is a discriminator that predicts whether an input dataset has maximum Shannon entropy, which shall be u sed on multiple new datasets $\hat{\mathcal{D}}$ synthesized from $\mathcal{D}$ via the insertion of additional label noise. We prove that our estimator for $Q_ {\mathcal{D}}$ is statistically consistent, in terms of dataset size, and the nu mber of intermediate datasets $\hat{\mathcal{D}}$ synthesized from $\mathcal{D}$. As a concrete realization of our framework, we shall incorporate local intrins ic dimensionality (LID) into the discriminator, and we show experimentally that with our LID-based discriminator, the estimation error for $Q_{\mathcal{D}}$ can be significantly reduced. We achieved average Kullback--Leibler loss reduction from 0.27 to 0.17 for 40% anchor-like samples removal when evaluated on t he CIFAR10 with symmetric noise. Although no clean subset of $\mathcal{D}$ is re quired for our framework to work, we show that our framework can also take advan tage of clean data to improve upon existing estimation methods.

****

Multi-Prize Lottery Ticket Hypothesis: Finding Accurate Binary Neural Networks b y Pruning A Randomly Weighted Network

James Diffenderfer,Bhavya Kailkhura

Recently, Frankle & Carbin (2019) demonstrated that randomly-initialized dense n etworks contain subnetworks that once found can be trained to reach test accurac y comparable to the trained dense network. However, finding these high performin g trainable subnetworks is expensive, requiring iterative process of training an d pruning weights. In this paper, we propose (and prove) a stronger Multi-Prize Lottery Ticket Hypothesis:

A sufficiently over-parameterized neural network with random weights contains several subnetworks (winning tickets) that (a) have comparable accuracy to a dense target network with learned weights (prize 1), (b) do not require any further training to achieve prize 1 (prize 2), and (c) is robust to extreme forms of quantization (i.e., binary weights and/or activation) (prize 3).

This provides a new paradigm for learning compact yet highly accurate binary neural networks simply by pruning and quantizing randomly weighted full precision neural networks. We also propose an algorithm for finding multi-prize tickets (MPTs) and test it by performing a series of experiments on CIFAR-10 and ImageNet datasets. Empirical results indicate that as models grow deeper and wider, multi-prize tickets start to reach similar (and sometimes even higher) test accuracy compared to their significantly larger and full-precision counterparts that have been weight-trained. Without ever updating the weight values, our MPTs-1/32 not only set new binary weight network state-of-the-art (SOTA) Top-1 accuracy -- 94.8% on CIFAR-10 and 74.03% on ImageNet -- but also outperform their full-precision counterparts by 1.78% and 0.76%, respectively. Further, our MPT-1/1 achieves SOTA Top-1 accuracy (91.9%) for binary neural networks on CIFAR-10. Code and pre-trained models are available at: https://github.com/chrundle/biprop.
****

DJMix: Unsupervised Task-agnostic Augmentation for Improving Robustness
Ryuichiro Hataya,Hideki Nakayama
Convolutional Neural Networks (CNNs) are vulnerable to unseen noise on input images at the test time, and thus improving the robustness is crucial. In this paper, we propose DJMix, a data augmentation method to improve the robustness by mixing each training image and its discretized one. Discretization is done in an unsupervised manner by an autoencoder, and the mixed images are nearly impossible to distinguish from the original images. Therefore, DJMix can easily be adapted to various image recognition tasks. We verify the effectiveness of our method using classification, semantic segmentation, and detection using clean and noisy test images.
****

Deep Learning with Data Privacy via Residual Perturbation
Wenqi Tao,Huaming Ling,Zuoqiang Shi,Bao Wang
Protecting data privacy in deep learning (DL) is at its urgency. Several celebrated privacy notions have been established and used for privacy-preserving DL. However, many of the existing mechanisms achieve data privacy at the cost of significant utility degradation. In this paper, we propose a stochastic differential equation principled \emph{residual perturbation} for privacy-preserving DL, which injects Gaussian noise into each residual mapping of ResNets. Theoretically, we prove that residual perturbation guarantees differential privacy (DP) and reduces the generalization gap for DL. Empirically, we show that residual perturbation outperforms the state-of-the-art DP stochastic gradient descent (DPSGD) in both membership privacy protection and maintaining the DL models' utility. For instance, in the process of training ResNet8 for the IDC dataset classification, residual perturbation obtains an accuracy of 85.7\% and protects the perfect membership privacy; in contrast, DPSGD achieves an accuracy of 82.8\% and protects worse membership privacy.
****

Semi-Supervised Learning via Clustering Representation Space
Yen-Chieh Huang,Yuh-Jye Lee,Chih-Chi Wu,Yi-Wei Chiu,Yong-Xiang Lin,■CHENG-YING LI,Po-Hung Ko
We proposed a novel loss function that combines supervised learning with clustering in deep neural networks. Taking advantage of the data distribution and the existence of some labeled data, we construct a meaningful latent space. Our loss function consists of three parts, the quality of the clustering result, the margin between clusters, and the classification error of labeled instances. Our proposed model is trained to minimize our loss function by backpropagation, avoiding the need for pre-training or additional networks. This guides our network to classify labeled samples correctly while able to find good clusters simultaneousl

y. We applied our proposed method on MNIST, USPS, ETH-80, and COIL-100; the comp
arison results confirm our model's outstanding performance over semi-supervised
learning.
****

A General Computational Framework to Measure the Expressiveness of Complex Netwo
rks using a Tight Upper Bound of Linear Regions
Yutong Xie,Gaoxiang Chen,Quanzheng Li
The expressiveness of deep neural network (DNN) is a perspective to understand t
he surprising performance of DNN. The number of linear regions, i.e. pieces that
 a piece-wise-linear function represented by a DNN, is generally used to measure
 the expressiveness. And the upper bound of regions number partitioned by a rect
ifier network, instead of the number itself, is a more practical measurement of
expressiveness of a rectifier DNN. In this work, we propose a new and tighter up
per bound of regions number. Inspired by the proof of this upper bound and the f
ramework of matrix computation in \citet{hinz2019framework}, we propose a genera
l computational approach to compute a tight upper bound of regions number for th
eoretically any network structures (e.g. DNN with all kind of skip connections a
nd residual structures). Our experiments show our upper bound is tighter than ex
isting ones, and explain why skip connections and residual structures can improv
e network performance.
****

Learning Accurate Entropy Model with Global Reference for Image Compression
Yichen Qian,Zhiyu Tan,Xiuyu Sun,Ming Lin,Dongyang Li,Zhenhong Sun,Li Hao,Rong Ji
n
In recent deep image compression neural networks, the entropy model plays a crit
ical role in estimating the prior distribution of deep image encodings. Existing
 methods combine hyperprior with local context in the entropy estimation functio
n. This greatly limits their performance due to the absence of a global vision.
In this work, we propose a novel Global Reference Model for image compression to
 effectively leverage both the local and the global context information, leading
 to an enhanced compression rate. The proposed method scans decoded latents and
then finds the most relevant latent to assist the distribution estimating of the
 current latent. A by-product of this work is the innovation of a mean-shifting
GDN module that further improves the performance. Experimental results demonstra
te that the proposed model outperforms the rate-distortion performance of most o
f the state-of-the-art methods in the industry.
****

Graph Autoencoders with Deconvolutional Networks
Jia Li,Jianwei Yu,Da-Cheng Juan,HAN Zhichao,Arjun Gopalan,Hong Cheng,Andrew Tomk
ins
Recent studies have indicated that Graph Convolutional Networks (GCNs) act as a
$\textit{low pass}$ filter in spectral domain and encode smoothed node represent
ations. In this paper, we consider their opposite, namely Graph Deconvolutional
 Networks (GDNs) that reconstruct graph signals from smoothed node representatio
ns. We motivate the design of Graph Deconvolutional Networks via a combination o
f inverse filters in spectral domain and de-noising layers in wavelet domain, as
 the inverse operation results in a $\textit{high pass}$ filter and may amplify
the noise. Based on the proposed GDN, we further propose a graph autoencoder fr
amework that first encodes smoothed graph representations with GCN and then deco
des accurate graph signals with GDN. We demonstrate the effectiveness of the pr
oposed method on several tasks including unsupervised graph-level representation
, social recommendation and graph generation.
****

Towards Principled Representation Learning for Entity Alignment
Lingbing Guo,Zequn Sun,Mingyang Chen,Wei Hu,Huajun Chen
Knowledge graph (KG) representation learning for entity alignment has recently r
eceived great attention. Compared with conventional methods, these embedding-bas
ed ones are considered to be robuster for highly-heterogeneous and cross-lingual
 entity alignment scenarios as they do not rely on the quality of machine trans
lation or feature extraction. Despite the significant improvement that has been

made, there is little understanding of how the embedding-based entity alignment methods actually work. Most existing methods rest on the foundation that a small number of pre-aligned entities can serve as anchors to connect the embedding spaces of two KGs. But no one investigates the rationality of such foundation. In this paper, we define a typical paradigm abstracted from the existing methods, and analyze how the representation discrepancy between two potentially-aligned entities is implicitly bounded by a predefined margin in the scoring function for embedding learning. However, such a margin cannot guarantee to be tight enough for alignment learning. We mitigate this problem by proposing a new approach that explicitly learns KG-invariant and principled entity representations, meanwhile preserves the original infrastructure of existing methods. In this sense, the model not only pursues the closeness of aligned entities on geometric distance, but also aligns the neural ontologies of two KGs to eliminate the discrepancy in feature distribution and underlying ontology knowledge. Our experiments demonstrate consistent and significant improvement in performance against the existing embedding-based entity alignment methods, including several state-of-the-art ones.

****

Neural Time-Dependent Partial Differential Equation

Yihao Hu,Tong Zhao,Zhiliang Xu,Lizhen Lin

Partial differential equations (PDEs) play a crucial role in studying a vast number of problems in science and engineering. Numerically solving nonlinear and/or high-dimensional PDEs is frequently a challenging task. Inspired by the traditional finite difference and finite elements methods and emerging advancements in machine learning, we propose a sequence-to-sequence learning (Seq2Seq) framework called Neural-PDE, which allows one to automatically learn governing rules of any time-dependent PDE system from existing data by using a bidirectional LSTM encoder, and predict the solutions in next n time steps. One critical feature of our proposed framework is that the Neural-PDE is able to simultaneously learn and simulate all variables of interest in a PDE system. We test the Neural-PDE by a range of examples, from one-dimensional PDEs to a multi-dimensional and nonlinear complex fluids model. The results show that the Neural-PDE is capable of learning the initial conditions, boundary conditions and differential operators defining the initial-boundary-value problem of a PDE system without the knowledge of the specific form of the PDE system. In our experiments, the Neural-PDE can efficiently extract the dynamics within 20 epochs training and produce accurate predictions. Furthermore, unlike the traditional machine learning approaches for learning PDEs, such as CNN and MLP, which require great quantity of parameters for model precision, the Neural-PDE shares parameters among all time steps, and thus considerably reduces computational complexity and leads to a fast learning algorithm.

****

Latent Causal Invariant Model

Xinwei Sun,Botong Wu,Chang Liu,Xiangyu Zheng,Wei Chen,Tao Qin,Tie-Yan Liu

Current supervised learning can learn spurious correlation during the data-fitting process, imposing issues regarding interpretability, out-of-distribution (OOD) generalization, and robustness. To avoid spurious correlation, we propose a \textbf{La}tent \textbf{C}ausal \textbf{I}nvariance \textbf{M}odel (LaCIM) which pursues \emph{causal prediction}. Specifically, we introduce latent variables that are separated into (a) output-causative factors and (b) others that are spuriously correlated to the output via confounders, to model the underlying causal factors. We further assume the generating mechanisms from latent space to observed data to be \emph{causally invariant}. We give the identifiable claim of such invariance, particularly the disentanglement of output-causative factors from others, as a theoretical guarantee for precise inference and avoiding spurious correlation. We propose a Variational-Bayesian-based method for estimation and to optimize over the latent space for prediction. The utility of our approach is verified by improved interpretability, prediction power on various OOD scenarios (including healthcare) and robustness on security.

****

Adaptive Stacked Graph Filter
Hoang NT,Takanori Maehara,Tsuyoshi Murata
We study Graph Convolutional Networks (GCN) from the graph signal processing vie wpoint by addressing a difference between learning graph filters with fully-conn ected weights versus trainable polynomial coefficients. We find that by stacking graph filters with learnable polynomial parameters, we can build a highly adapt ive and robust vertex classification model. Our treatment here relaxes the low-f requency (or equivalently, high homophily) assumptions in existing vertex classi fication models, resulting a more ubiquitous solution in terms of spectral prope rties. Empirically, by using only one hyper-parameter setting, our model achieve s strong results on most benchmark datasets across the frequency spectrum.
****
Convex Potential Flows: Universal Probability Distributions with Optimal Transpo rt and Convex Optimization
Chin-Wei Huang,Ricky T. Q. Chen,Christos Tsirigotis,Aaron Courville
Flow-based models are powerful tools for designing probabilistic models with tra ctable density. This paper introduces Convex Potential Flows (CP-Flow), a natura l and efficient parameterization of invertible models inspired by the optimal tr ansport (OT) theory. CP-Flows are the gradient map of a strongly convex neural p otential function. The convexity implies invertibility and allows us to resort t o convex optimization to solve the convex conjugate for efficient inversion. To enable maximum likelihood training, we derive a new gradient estimator of the lo g-determinant of the Jacobian, which involves solving an inverse-Hessian vector product using the conjugate gradient method. The gradient estimator has constant -memory cost, and can be made effectively unbiased by reducing the error toleran ce level of the convex optimization routine. Theoretically, we prove that CP-Flo ws are universal density approximators and are optimal in the OT sense. Our empi rical results show that CP-Flow performs competitively on standard benchmarks of density estimation and variational inference.
****
Towards Noise-resistant Object Detection with Noisy Annotations
Junnan Li,Caiming Xiong,Steven Hoi
Training deep object detectors requires large amounts of human-annotated images with accurate object labels and bounding box coordinates, which are extremely ex pensive to acquire. Noisy annotations are much more easily accessible, but they could be detrimental for learning. We address the challenging problem of traini ng object detectors with noisy annotations, where the noise contains a mixture o f label noise and bounding box noise. We propose a learning framework which join tly optimizes object labels, bounding box coordinates, and model parameters by p erforming alternating noise correction and model training. To disentangle label noise and bounding box noise, we propose a two-step noise correction method. The first step performs class-agnostic bounding box correction, and the second step performs label correction and class-specific bounding box refinement. We conduc t experiments on PASCAL VOC and MS-COCO dataset with both synthetic noise and ma chine-generated noise. Our method achieves state-of-the-art performance by effec tively cleaning both label noise and bounding box noise.
****
Greedy-GQ with Variance Reduction: Finite-time Analysis and Improved Complexity
Shaocong Ma,Ziyi Chen,Yi Zhou,Shaofeng Zou
Greedy-GQ is a value-based reinforcement learning (RL) algorithm for optimal con trol. Recently, the finite-time analysis of Greedy-GQ has been developed under l inear function approximation and Markovian sampling, and the algorithm is shown to achieve an ϵ-stationary point with a sample complexity in the order of $\mathcal{O}(\epsilon^{-3})$. Such a high sample complexity is due to the lar ge variance induced by the Markovian samples. In this paper, we propose a varian ce-reduced Greedy-GQ (VR-Greedy-GQ) algorithm for off-policy optimal control. In particular, the algorithm applies the SVRG-based variance reduction scheme to r educe the stochastic variance of the two time-scale updates. We study the finite -time convergence of VR-Greedy-GQ under linear function approximation and Markov ian sampling and show that the algorithm achieves a much smaller bias and varian

ce error than the original Greedy-GQ. In particular, we prove that VR-Greedy-GQ achieves an improved sample complexity that is in the order of $\mathcal{O}(\epsilon^{-2})$. We further compare the performance of VR-Greedy-GQ with that of Greedy-GQ in various RL experiments to corroborate our theoretical findings.
****
How much progress have we made in neural network training? A New Evaluation Protocol for Benchmarking Optimizers
Yuanhao Xiong,Xuanqing Liu,Li-Cheng Lan,Yang You,Si Si,Cho-Jui Hsieh
Many optimizers have been proposed for training deep neural networks, and they often have multiple hyperparameters, which make it tricky to benchmark their performance. In this work, we propose a new benchmarking protocol to evaluate both end-to-end efficiency (training a model from scratch without knowing the best hyperparameter) and data-addition training efficiency (the previously selected hyperparameters are used for periodically re-training the model with newly collected data). For end-to-end efficiency, unlike previous work that assumes random hyperparameter tuning, which over-emphasizes the tuning time, we propose to evaluate with a bandit hyperparameter tuning strategy. A human study is conducted to show our evaluation protocol matches human tuning behavior better than the random search. For data-addition training, we propose a new protocol for assessing the hyperparameter sensitivity to data shift. We then apply the proposed benchmarking framework to 7 optimizers and various tasks, including computer vision, natural language processing, reinforcement learning, and graph mining. Our results show that there is no clear winner across all the tasks.

****
Large Batch Simulation for Deep Reinforcement Learning
Brennan Shacklett,Erik Wijmans,Aleksei Petrenko,Manolis Savva,Dhruv Batra,Vladlen Koltun,Kayvon Fatahalian
We accelerate deep reinforcement learning-based training in visually complex 3D environments by two orders of magnitude over prior work, realizing end-to-end training speeds of over 19,000 frames of experience per second on a single GPU and up to 72,000 frames per second on a single eight-GPU machine. The key idea of our approach is to design a 3D renderer and embodied navigation simulator around the principle of “batch simulation”: accepting and executing large batches of requests simultaneously. Beyond exposing large amounts of work at once, batch simulation allows implementations to amortize in-memory storage of scene assets, rendering work, data loading, and synchronization costs across many simulation requests, dramatically improving the number of simulated agents per GPU and overall simulation throughput. To balance DNN inference and training costs with faster simulation, we also build a computationally efficient policy DNN that maintains high task performance, and modify training algorithms to maintain sample efficiency when training with large mini-batches. By combining batch simulation and DNN performance optimizations, we demonstrate that PointGoal navigation agents can be trained in complex 3D environments on a single GPU in 1.5 days to 97% of the accuracy of agents trained on a prior state-of-the-art system using a 64-GPU cluster over three days. We provide open-source reference implementations of our batch 3D renderer and simulator to facilitate incorporation of these ideas into RL systems.
****
Adversarial Training using Contrastive Divergence
Hongjun Wang,Guanbin Li,Liang Lin
To protect the security of machine learning models against adversarial examples, adversarial training becomes the most popular and powerful strategy against various adversarial attacks by injecting adversarial examples into training data. However, it is time-consuming and requires high computation complexity to generate suitable adversarial examples for ensuring the robustness of models, which impedes the spread and application of adversarial training. In this work, we reformulate adversarial training as a combination of stationary distribution exploring, sampling, and training. Each updating of parameters of DNN is based on several transitions from the data samples as the initial states in a Hamiltonian system

. Inspired by our new paradigm, we design a new generative method for adversaria
l training by using Contrastive Divergence (ATCD), which approaches the equilibr
ium distribution of adversarial examples with only few iterations by building fr
om small modifications of the standard Contrastive Divergence (CD). Our adversar
ial training algorithm achieves much higher robustness than any other state-of-t
he-art adversarial training acceleration method on the ImageNet, CIFAR-10, and M
NIST datasets and reaches a balance between performance and efficiency.
****
Hopper: Multi-hop Transformer for Spatiotemporal Reasoning
Honglu Zhou,Asim Kadav,Farley Lai,Alexandru Niculescu-Mizil,Martin Renqiang Min,
Mubbasir Kapadia,Hans Peter Graf
This paper considers the problem of spatiotemporal object-centric reasoning in v
ideos. Central to our approach is the notion of object permanence, i.e., the abi
lity to reason about the location of objects as they move through the video whil
e being occluded, contained or carried by other objects. Existing deep learning
based approaches often suffer from spatiotemporal biases when applied to video r
easoning problems. We propose Hopper, which uses a Multi-hop Transformer for rea
soning object permanence in videos. Given a video and a localization query, Hopp
er reasons over image and object tracks to automatically hop over critical frame
s in an iterative fashion to predict the final position of the object of interes
t. We demonstrate the effectiveness of using a contrastive loss to reduce spatio
temporal biases. We evaluate over CATER dataset and find that Hopper achieves 73
.2% Top-1 accuracy using just 1 FPS by hopping through just a few critical frame
s. We also demonstrate Hopper can perform long-term reasoning by building a CATE
R-h dataset that requires multi-step reasoning to localize objects of interest c
orrectly.
****
Efficient Reinforcement Learning in Factored MDPs with Application to Constraine
d RL
Xiaoyu Chen,Jiachen Hu,Lihong Li,Liwei Wang
Reinforcement learning (RL) in episodic, factored Markov decision processes (FMD
Ps) is studied. We propose an algorithm called FMDP-BF, which leverages the fact
orization structure of FMDP. The regret of FMDP-BF is shown to be exponentially
 smaller than that of optimal algorithms designed for non-factored MDPs, and imp
roves on the best previous result for FMDPs~\citep{osband2014near} by a factor o
f $\sqrt{nH|\mathcal{S}_i|}$, where $|\mathcal{S}_i|$ is the cardinality of the
factored state subspace, H is the planning horizon and n is the number of fa
ctored transition. To show the optimality of our bounds, we also provide a lower
 bound for FMDP, which indicates that our algorithm is near-optimal w.r.t. times
tep T, horizon H and factored state-action subspace cardinality. Finally, as
 an application, we study a new formulation of constrained RL, known as RL with
knapsack constraints (RLwK), and provides the first sample-efficient algorithm b
ased on FMDP-BF.
****
Finding Patient Zero: Learning Contagion Source with Graph Neural Networks
Chintan Shah,Nima Dehmamy,Nicola Perra,Matteo Chinazzi,Albert-Laszlo Barabasi,Al
essandro Vespignani,Rose Yu
Locating the source of an epidemic, or patient zero (P0), can provide critical i
nsights into the infection's transmission course and allow efficient resource a
llocation.
Existing methods use graph-theoretic centrality measures and expensive message-p
assing algorithms, requiring knowledge of the underlying dynamics and its parame
ters.
In this paper, we revisit this problem using graph neural networks (GNNs) to lea
rn P0.
We observe that GNNs can identify P0 close to the theoretical bound on accuracy,
 without explicit input of dynamics or its parameters.
In addition, GNN is over 100 times faster than classic methods for inference on
arbitrary graph topologies.
Our theoretical bound also shows that the epidemic is like a ticking clock, emph

asizing the importance of early contact-tracing.
We find a maximum time after which accurate recovery of the source becomes impos
sible, regardless of the algorithm used.
****

Factorized linear discriminant analysis for phenotype-guided representation lear
ning of neuronal gene expression data
Mu Qiao,Markus Meister
A central goal in neurobiology is to relate the expression of genes to the struc
tural and functional properties of neuronal types, collectively called their phe
notypes. Single-cell RNA sequencing can measure the expression of thousands of g
enes in thousands of neurons. How to interpret the data in the context of neuron
al phenotypes? We propose a supervised learning approach that factorizes the gen
e expression data into components corresponding to individual phenotypic charact
eristics and their interactions. This new method, which we call factorized linea
r discriminant analysis (FLDA), seeks a linear transformation of gene expression
s that varies highly with only one phenotypic factor and minimally with the othe
rs. We further leverage our approach with a sparsity-based regularization algori
thm, which selects a few genes important to a specific phenotypic feature or fea
ture combination. We applied this approach to a single-cell RNA-Seq dataset of D
rosophila T4/T5 neurons, focusing on their dendritic and axonal phenotypes. The
analysis confirms results obtained by conventional methods but also points to ne
w genes related to the phenotypes and an intriguing hierarchy in the genetic org
anization of these cells.
****

RankingMatch: Delving into Semi-Supervised Learning with Consistency Regularizat
ion and Ranking Loss
Trung Quang Tran,Mingu Kang,Daeyoung Kim
Semi-supervised learning (SSL) has played an important role in leveraging unlabe
led data when labeled data is limited. One of the most successful SSL approaches
 is based on consistency regularization, which encourages the model to produce u
nchanged with perturbed input. However, there has been less attention spent on i
nputs that have the same label. Motivated by the observation that the inputs hav
ing the same label should have the similar model outputs, we propose a novel met
hod, RankingMatch, that considers not only the perturbed inputs but also the sim
ilarity among the inputs having the same label. We especially introduce a new ob
jective function, dubbed BatchMean Triplet loss, which has the advantage of comp
utational efficiency while taking into account all input samples. Our RankingMat
ch achieves state-of-the-art performance across many standard SSL benchmarks wit
h a variety of labeled data amounts, including 95.13% accuracy on CIFAR-10 with
250 labels, 77.65% accuracy on CIFAR-100 with 10000 labels, 97.76% accuracy on S
VHN with 250 labels, and 97.77% accuracy on SVHN with 1000 labels. We also perfo
rm an ablation study to prove the efficacy of the proposed BatchMean Triplet los
s against existing versions of Triplet loss.
****

Unbiased Teacher for Semi-Supervised Object Detection
Yen-Cheng Liu,Chih-Yao Ma,Zijian He,Chia-Wen Kuo,Kan Chen,Peizhao Zhang,Bichen W
u,Zsolt Kira,Peter Vajda
Semi-supervised learning, i.e., training networks with both labeled and unlabele
d data, has made significant progress recently. However, existing works have pri
marily focused on image classification tasks and neglected object detection whic
h requires more annotation effort. In this work, we revisit the Semi-Supervised
Object Detection (SS-OD) and identify the pseudo-labeling bias issue in SS-OD. T
o address this, we introduce Unbiased Teacher, a simple yet effective approach t
hat jointly trains a student and a gradually progressing teacher in a mutually-b
eneficial manner. Together with a class-balance loss to downweight overly confid
ent pseudo-labels, Unbiased Teacher consistently improved state-of-the-art metho
ds by significant margins on COCO-standard, COCO-additional, and VOC datasets. S
pecifically, Unbiased Teacher achieves 6.8 absolute mAP improvements against sta
te-of-the-art method when using 1% of labeled data on MS-COCO, achieves around 1
0 mAP improvements against the supervised baseline when using only 0.5, 1, 2% of

labeled data on MS-COCO.
****
Can Students Outperform Teachers in Knowledge Distillation based Model Compressi
on?
Xiang Deng,Zhongfei Zhang
Knowledge distillation (KD) is an effective technique to compress a large model
(teacher) to a compact one (student) by knowledge transfer. The ideal case is th
at the teacher is compressed to the small student without any performance droppi
ng. However, even for the state-of-the-art (SOTA) distillation approaches, there
is still an obvious performance gap between the student and the teacher. The ex
isting literature usually attributes this to model capacity differences between
them. However, model capacity differences are unavoidable in model compression.
In this work, we systematically study this question. By designing exploratory ex
periments, we find that model capacity differences are not necessarily the root
reason, and the distillation data matters when the student capacity is greater t
han a threshold. In light of this, we propose to go beyond in-distribution disti
llation and accordingly develop KD+. KD+ is superior to the original KD as it ou
tperforms KD and the other SOTA approaches substantially and is more compatible
with the existing approaches to further improve their performances significantly
.
****
On interaction between augmentations and corruptions in natural corruption robus
tness
Eric Mintun,Alexander Kirillov,Saining Xie
Invariance to a broad array of image corruptions, such as warping, noise, or col
or shifts, is an important aspect of building robust models in computer vision.
Recently, several new data augmentations have been proposed that significantly i
mprove performance on ImageNet-C, a benchmark of such corruptions. However, ther
e is still a lack of basic understanding on the relationship between data augmen
tations and test-time corruptions. To this end, we develop a feature space for i
mage transforms, and then use a new measure in this space between augmentations
and corruptions called the Minimal Sample Distance to demonstrate there is a str
ong correlation between similarity and performance. We then investigate recent d
ata augmentations and observe a significant degradation in corruption robustness
when the test-time corruptions are sampled to be perceptually dissimilar from I
mageNet-C in this feature space. Our results suggest that test error can be impr
oved by training on perceptually similar augmentations, and data augmentations m
ay risk overfitting to the existing benchmark. We hope our results and tools wi
ll allow for more robust progress towards improving robustness to image corrupti
ons.
****
MetaPhys: Few-Shot Adaptation for Non-Contact Physiological Measurement
Xin Liu,Ziheng Jiang,Joshua Wolff Fromm,Xuhai Xu,Shwetak Patel,Daniel McDuff
There are large individual differences in physiological processes, making design
ing personalized health sensing algorithms challenging. Existing machine learnin
g systems struggle to generalize well to unseen subjects or contexts, especially
in video-based physiological measurement. Although fine-tuning for a user might
address this issue, it is difficult to collect large sets of training data for
specific individuals because supervised algorithms require medical-grade sensors
for generating the training target. Therefore, learning personalized or customi
zed models from a small number of unlabeled samples is very attractive as it wou
ld allow fast calibrations. In this paper, we present a novel meta-learning appr
oach called MetaPhys for learning personalized cardiac signals from 18-seconds o
f video data. MetaPhys works in both supervised and unsupervised manners. We eva
luate our proposed approach on two benchmark datasets and demonstrate superior p
erformance in cross-dataset evaluation with substantial reductions (42% to 44%)
in errors compared with state-of-the-art approaches. Visualization of attention
maps and ablation experiments reveal how the model adapts to each subject and wh
y our proposed approach leads to these improvements. We have also demonstrated o
ur proposed method significantly helps reduce the bias in skin type.

****
Human-Level Performance in No-Press Diplomacy via Equilibrium Search
Jonathan Gray,Adam Lerer,Anton Bakhtin,Noam Brown
Prior AI breakthroughs in complex games have focused on either the purely adversarial or purely cooperative settings. In contrast, Diplomacy is a game of shifting alliances that involves both cooperation and competition. For this reason, Diplomacy has proven to be a formidable research challenge. In this paper we describe an agent for the no-press variant of Diplomacy that combines supervised learning on human data with one-step lookahead search via regret minimization. Regret minimization techniques have been behind previous AI successes in adversarial games, most notably poker, but have not previously been shown to be successful in large-scale games involving cooperation. We show that our agent greatly exceeds the performance of past no-press Diplomacy bots, is unexploitable by expert humans, and ranks in the top 2% of human players when playing anonymous games on a popular Diplomacy website.
****
Uncovering the impact of hyperparameters for global magnitude pruning
Janice Lan,Rudy Chin,Alexei Baevski,Ari S. Morcos
A common paradigm in model pruning is to train a model, prune, and then either fine-tune or, in the lottery ticket framework, reinitialize and retrain. Prior work has implicitly assumed that the best training configuration for model evaluation is also the best configuration for mask discovery. However, what if a training configuration which yields worse performance actually yields a mask which trains to higher performance? To test this, we decoupled the hyperparameters for mask discovery (H_find) and mask evaluation (H_eval). Using unstructured magnitude pruning on vision classification tasks, we discovered the "decoupled find-eval phenomenon," in which certain H_find values lead to models which have lower performance, but generate masks with substantially higher eventual performance compared to using the same hyperparameters for both stages. We show that this phenomenon holds across a number of models, datasets, configurations, and also for one-shot structured pruning. Finally, we demonstrate that different H_find values yield masks with materially different layerwise pruning ratios and that the decoupled find-eval phenomenon is causally mediated by these ratios. Our results demonstrate the practical utility of decoupling hyperparameters and provide clear insights into the mechanisms underlying this counterintuitive effect.
****
How Neural Networks Extrapolate: From Feedforward to Graph Neural Networks
Keyulu Xu,Mozhi Zhang,Jingling Li,Simon Shaolei Du,Ken-Ichi Kawarabayashi,Stefanie Jegelka
We study how neural networks trained by gradient descent extrapolate, i.e., what they learn outside the support of the training distribution. Previous works report mixed empirical results when extrapolating with neural networks: while feedforward neural networks, a.k.a. multilayer perceptrons (MLPs), do not extrapolate well in certain simple tasks, Graph Neural Networks (GNNs) -- structured networks with MLP modules -- have shown some success in more complex tasks. Working towards a theoretical explanation, we identify conditions under which MLPs and GNNs extrapolate well. First, we quantify the observation that ReLU MLPs quickly converge to linear functions along any direction from the origin, which implies that ReLU MLPs do not extrapolate most nonlinear functions. But, they can provably learn a linear target function when the training distribution is sufficiently diverse. Second, in connection to analyzing the successes and limitations of GNNs, these results suggest a hypothesis for which we provide theoretical and empirical evidence: the success of GNNs in extrapolating algorithmic tasks to new data (e.g., larger graphs or edge weights) relies on encoding task-specific non-linearities in the architecture or features. Our theoretical analysis builds on a connection of over-parameterized networks to the neural tangent kernel. Empirically, our theory holds across different training settings.
****
Acceleration in Hyperbolic and Spherical Spaces

David Martínez-Rubio

We further research on the acceleration phenomenon on Riemannian manifolds by introducing the first global first-order method that achieves the same rates as accelerated gradient descent in the Euclidean space for the optimization of smooth and geodesically convex (g-convex) or strongly g-convex functions defined on the hyperbolic space or a subset of the sphere, up to constants and log factors. To the best of our knowledge, this is the first method that is proved to achieve these rates globally on functions defined on a Riemannian manifold $\mathcal{M}$ other than the Euclidean space.

Additionally, for any Riemannian manifold of bounded sectional curvature, we provide reductions from optimization methods for smooth and g-convex functions to methods for smooth and strongly g-convex functions and vice versa.

****

MELR: Meta-Learning via Modeling Episode-Level Relationships for Few-Shot Learning

Nanyi Fei,Zhiwu Lu,Tao Xiang,Songfang Huang

Most recent few-shot learning (FSL) approaches are based on episodic training whereby each episode samples few training instances (shots) per class to imitate the test condition. However, this strict adhering to test condition has a negative side effect, that is, the trained model is susceptible to the poor sampling of few shots. In this work, for the first time, this problem is addressed by exploiting inter-episode relationships. Specifically, a novel meta-learning via modeling episode-level relationships (MELR) framework is proposed. By sampling two episodes containing the same set of classes for meta-training, MELR is designed to ensure that the meta-learned model is robust against the presence of poorly-sampled shots in the meta-test stage. This is achieved through two key components: (1) a Cross-Episode Attention Module (CEAM) to improve the ability of alleviating the effects of poorly-sampled shots, and (2) a Cross-Episode Consistency Regularization (CECR) to enforce that the two classifiers learned from the two episodes are consistent even when there are unrepresentative instances. Extensive experiments for non-transductive standard FSL on two benchmarks show that our MELR achieves 1.0%-5.0% improvements over the baseline (i.e., ProtoNet) used for FSL in our model and outperforms the latest competitors under the same settings.

****

Enabling counterfactual survival analysis with balanced representations

Paidamoyo Chapfuwa,Serge Assaad,Shuxi Zeng,Michael Pencina,Lawrence Carin,Ricardo Henao

Balanced representation learning methods have been applied successfully to counterfactual
inference from observational data. However, approaches that account for
survival outcomes are relatively limited. Survival data are frequently encountered
across diverse medical applications, i.e., drug development, risk profiling, and clinical
trials, and such data are also relevant in fields like manufacturing (for equipment
monitoring). When the outcome of interest is time-to-event, special precautions
for handling censored events need to be taken, as ignoring censored outcomes may
lead to biased estimates. We propose a theoretically grounded unified framework
for counterfactual inference applicable to survival outcomes. Further, we formulate
a nonparametric hazard ratio metric for evaluating average and individualized
treatment effects. Experimental results on real-world and semi-synthetic datasets,
the latter which we introduce, demonstrate that the proposed approach significantly
outperforms competitive alternatives in both survival-outcome predictions and
treatment-effect estimation.

****

Partitioned Learned Bloom Filters

Kapil Vaidya,Eric Knorr,Michael Mitzenmacher,Tim Kraska
Bloom filters are space-efficient probabilistic data structures that are used to test whether an element is a member of a set, and may return false positives. Recently, variations referred to as learned Bloom filters were developed that can provide improved performance in terms of the rate of false positives, by using a learned model for the represented set. However, previous methods for learned Bloom filters do not take full advantage of the learned model. Here we show how to frame the problem of optimal model utilization as an optimization problem, and using our framework derive algorithms that can achieve near-optimal performance in many cases.
****

An Empirical Exploration of Open-Set Recognition via Lightweight Statistical Pipelines
Shu Kong,Deva Ramanan
Machine-learned safety-critical systems need to be self-aware and reliably know their unknowns in the open-world. This is often explored through the lens of anomaly/outlier detection or out-of-distribution modeling. One popular formulation is that of open-set classification, where an image classifier trained for 1-of-K classes should also recognize images belonging to a $(K+1)^{th}$ "other" class, not present in the training set. Recent work has shown that, somewhat surprisingly, most if not all existing open-world methods do not work well on high-dimensional open-world images (Shafaei et al. 2019). In this paper, we carry out an empirical exploration of open-set classification, and find that combining classic statistical methods with carefully computed features can dramatically outperform prior work. We extract features from off-the-shelf (OTS) state-of-the-art networks for the underlying K-way closed-world task. We leverage insights from the retrieval community for computing feature descriptors that are low-dimensional (via pooling and PCA) and normalized (via L2-normalization), enabling the modeling of training data densities via classic statistical tools such as kmeans and Gaussian Mixture Models (GMMs).
****

Recovering Geometric Information with Learned Texture Perturbations
Jane Wu,Yongxu Jin,Zhenglin Geng,Hui Zhou,Ronald Fedkiw
Regularization is used to avoid overfitting when training a neural network; unfortunately, this reduces the attainable level of detail hindering the ability to capture high-frequency information present in the training data. Even though various approaches may be used to re-introduce high-frequency detail, it typically does not match the training data and is often not time coherent. In the case of network inferred cloth, these sentiments manifest themselves via either a lack of detailed wrinkles or unnaturally appearing and/or time incoherent surrogate wrinkles. Thus, we propose a general strategy whereby high-frequency information is procedurally embedded into low-frequency data so that when the latter is smeared out by the network the former still retains its high-frequency detail. We illustrate this approach by learning texture coordinates which when smeared do not in turn smear out the high-frequency detail in the texture itself but merely smoothly distort it. Notably, we prescribe perturbed texture coordinates that are subsequently used to correct the over-smoothed appearance of inferred cloth, and correcting the appearance from multiple camera views naturally recovers lost geometric information.
****

Skinning a Parameterization of Three-Dimensional Space for Neural Network Cloth
Jane Wu,Zhenglin Geng,Hui Zhou,Ronald Fedkiw
We present a novel learning framework for cloth deformation by embedding virtual cloth into a tetrahedral mesh that parametrizes the volumetric region of air surrounding the underlying body. In order to maintain this volumetric parameterization during character animation, the tetrahedral mesh is constrained to follow the body surface as it deforms. We embed the cloth mesh vertices into this parameterization of three-dimensional space in order to automatically capture much of the nonlinear deformation due to both joint rotations and collisions. We then train a convolutional neural network to recover ground truth deformation by learni

ng cloth embedding offsets for each skeletal pose. Our experiments show signific ant improvement over learning cloth offsets from body surface parameterizations, both quantitatively and visually, with prior state of the art having a mean err or five standard deviations higher than ours. Without retraining, our neural net work generalizes to other body shapes and T-shirt sizes, giving the user some in dication of how well clothing might fit. Our results demonstrate the efficacy of a general learning paradigm where high-frequency details can be embedded into l ow-frequency parameterizations.
****
Wasserstein Embedding for Graph Learning
Soheil Kolouri,Navid Naderializadeh,Gustavo K. Rohde,Heiko Hoffmann
We present Wasserstein Embedding for Graph Learning (WEGL), a novel and fast fra mework for embedding entire graphs in a vector space, in which various machine l earning models are applicable for graph-level prediction tasks. We leverage new insights on defining similarity between graphs as a function of the similarity b etween their node embedding distributions. Specifically, we use the Wasserstein distance to measure the dissimilarity between node embeddings of different graph s. Unlike prior work, we avoid pairwise calculation of distances between graphs and reduce the computational complexity from quadratic to linear in the number o f graphs. WEGL calculates Monge maps from a reference distribution to each node embedding and, based on these maps, creates a fixed-sized vector representation of the graph. We evaluate our new graph embedding approach on various benchmark graph-property prediction tasks, showing state-of-the-art classification perform ance while having superior computational efficiency. The code is available at ht tps://github.com/navid-naderi/WEGL.
****
High-Capacity Expert Binary Networks
Adrian Bulat,Brais Martinez,Georgios Tzimiropoulos
Network binarization is a promising hardware-aware direction for creating effici ent deep models. Despite its memory and computational advantages, reducing the a ccuracy gap between binary models and their real-valued counterparts remains an unsolved challenging research problem. To this end, we make the following 3 cont ributions: (a) To increase model capacity, we propose Expert Binary Convolution, which, for the first time, tailors conditional computing to binary networks by learning to select one data-specific expert binary filter at a time conditioned on input features. (b) To increase representation capacity, we propose to addres s the inherent information bottleneck in binary networks by introducing an effic ient width expansion mechanism which keeps the binary operations within the same budget. (c) To improve network design, we propose a principled binary network g rowth mechanism that unveils a set of network topologies of favorable properties . Overall, our method improves upon prior work, with no increase in computationa l cost, by $\sim6 \%$, reaching a groundbreaking $\sim 71\%$ on ImageNet classif ication. Code will be made available $\href{https://www.adrianbulat.com/binary-n etworks}{here}$.
****
Neural Subgraph Matching
Zhitao Ying,Andrew Wang,Jiaxuan You,Chengtao Wen,Arquimedes Canedo,Jure Leskovec
Subgraph matching is the problem of determining the presence and location(s) of a given query graph in a large target graph.
Despite being an NP-complete problem, the subgraph matching problem is crucial i n domains ranging from network science and database systems to biochemistry and cognitive science.
However, existing techniques based on combinatorial matching and integer program ming cannot handle matching problems with both large target and query graphs.
Here we propose NeuroMatch, an accurate, efficient, and robust neural approach t o subgraph matching. NeuroMatch decomposes query and target graphs into small su bgraphs and embeds them using graph neural networks. Trained to capture geometri c constraints corresponding to subgraph relations, NeuroMatch then efficiently p erforms subgraph matching directly in the embedding space. Experiments demonstra te NeuroMatch is 100x faster than existing combinatorial approaches and 18% more

accurate than existing approximate subgraph matching methods.
****
SAFENet: A Secure, Accurate and Fast Neural Network Inference
Qian Lou,Yilin Shen,Hongxia Jin,Lei Jiang
The advances in neural networks have driven many companies to provide prediction services to users in a wide range of applications. However, current prediction systems raise privacy concerns regarding the user's private data. A cryptographic neural network inference service is an efficient way to allow two parties to execute neural network inference without revealing either party’s data or model. Nevertheless, existing cryptographic neural network inference services suffer from huge running latency; in particular, the latency of communication-expensive cryptographic activation function is 3 orders of magnitude higher than plaintext-domain activation function. And activations are the necessary components of the modern neural networks. Therefore, slow cryptographic activation has become the primary obstacle of efficient cryptographic inference.

In this paper, we propose a new technique, called SAFENet, to enable a Secure, Accurate and Fast nEural Network inference service. To speedup secure inference and guarantee inference accuracy, SAFENet includes channel-wise activation approximation with multiple-degree options. This is implemented by keeping the most useful activation channels and replacing the remaining, less useful, channels with various-degree polynomials. SAFENet also supports mixed-precision activation approximation by automatically assigning different replacement ratios to various layer; further increasing the approximation ratio and reducing inference latency. Our experimental results show SAFENet obtains the state-of-the-art inference latency and performance, reducing latency by $38\% \sim 61\%$ or improving accuracy by $1.8\% \sim 4\%$ over prior techniques on various encrypted datasets.
****
Learning Manifold Patch-Based Representations of Man-Made Shapes
Dmitriy Smirnov,Mikhail Bessmeltsev,Justin Solomon
Choosing the right representation for geometry is crucial for making 3D models compatible with existing applications. Focusing on piecewise-smooth man-made shapes, we propose a new representation that is usable in conventional CAD modeling pipelines and can also be learned by deep neural networks. We demonstrate its benefits by applying it to the task of sketch-based modeling. Given a raster image, our system infers a set of parametric surfaces that realize the input in 3D. To capture piecewise smooth geometry, we learn a special shape representation: a deformable parametric template composed of Coons patches. Naively training such a system, however, is hampered by non-manifold artifacts in the parametric shapes and by a lack of data. To address this, we introduce loss functions that bias the network to output non-self-intersecting shapes and implement them as part of a fully self-supervised system, automatically generating both shape templates and synthetic training data. We develop a testbed for sketch-based modeling, demonstrate shape interpolation, and provide comparison to related work.
****
Universal approximation power of deep residual neural networks via nonlinear control theory
Paulo Tabuada,Bahman Gharesifard
In this paper, we explain the universal approximation capabilities of deep residual neural networks through geometric nonlinear control. Inspired by recent work establishing links between residual networks and control systems, we provide a general sufficient condition for a residual network to have the power of universal approximation by asking the activation function, or one of its derivatives, to satisfy a quadratic differential equation. Many activation functions used in practice satisfy this assumption, exactly or approximately, and we show this property to be sufficient for an adequately deep neural network with $n+1$ neurons per
layer to approximate arbitrarily well, on a compact set and with respect to the supremum norm, any continuous function from $\mathbb{R}^n$ to $\mathbb{R}^n$. We further show this result to hold for very simple architectures for which the we

ights only need to assume two values. The first key technical contribution consi sts of relating the universal approximation problem to controllability of an ens emble of control systems corresponding to a residual network and to leverage cla ssical Lie algebraic techniques to characterize controllability. The second tech nical contribution is to identify monotonicity as the bridge between controllabi lity of finite ensembles and uniform approximability on compact sets.
****

Dataset Curation Beyond Accuracy
Johan Bjorck,Carla P Gomes
Neural networks are known to be data-hungry, and collecting large labeled datase ts is often a crucial step in deep learning deployment. Researchers have studied dataset aspects such as distributional shift and labeling cost, primarily using downstream prediction accuracy for evaluation. In sensitive real-world applicat ions such as medicine and self-driving cars, not only is the accuracy important, but also the calibration -- the extent that model uncertainty reflects the actu al correctness likelihood. It has recently been shown that modern neural network s are ill-calibrated. In this work, we take a complementary approach -- studying how dataset properties, rather than architecture, affects calibration. For the common issue of dataset imbalance, we show that calibration varies significantly among classes, even when common strategies to mitigate class imbalance are empl oyed. We also study the effects of label quality, showing how label noise dramat ically increases calibration error. Furthermore, poor calibration can come from small dataset sizes, which we motive via results on network expressivity. Our ex periments demonstrate that dataset properties can significantly affect calibrati on and suggest that calibration should be measured during dataset curation.
****

PareCO: Pareto-aware Channel Optimization for Slimmable Neural Networks
Rudy Chin,Ari S. Morcos,Diana Marculescu
Slimmable neural networks provide a flexible trade-off front between prediction error and computational cost (such as the number of floating-point operations or FLOPs) with the same storage cost as a single model. They have been proposed re cently for resource-constrained settings such as mobile devices. However, curren t slimmable neural networks use a single width-multiplier for all the layers to arrive at sub-networks with different performance profiles, which neglects that different layers affect the network's prediction accuracy differently and have d ifferent FLOP requirements. Hence, developing a principled approach for decidin g width-multipliers across different layers could potentially improve the perfor mance of slimmable networks. To allow for heterogeneous width-multipliers across different layers, we formulate the problem of optimizing slimmable networks fro m a multi-objective optimization lens, which leads to a novel algorithm for opti mizing both the shared weights and the width-multipliers for the sub-networks. W e perform extensive empirical analysis with 15 network and dataset combinations and two types of cost objectives, i.e., FLOPs and memory footprint, to demonstra te the effectiveness of the proposed method compared to existing alternatives. Q uantitatively, improvements up to 1.7% and 8% in top-1 accuracy on the ImageNet dataset can be attained for MobileNetV2 considering FLOPs and memory footprint, respectively. Our results highlight the potential of optimizing the channel coun ts for different layers jointly with the weights for slimmable networks.
****

Improving Local Effectiveness for Global Robustness Training
JINGYUE LU,M. Pawan Kumar
Despite its increasing popularity, deep neural networks are easily fooled. To al leviate this deficiency, researchers are actively developing new training strate gies, which encourage models that are robust to small input perturbations. Sever al successful robust training methods have been proposed. However, many of them rely on strong adversaries, which can be prohibitively expensive to generate whe n the input dimension is high and the model structure is complicated. We adopt a new perspective on robustness and propose a novel training algorithm that allow s a more effective use of adversaries. Our method improves the model robustness at each local ball centered around an adversary and then, by combining these loc

al balls through a global term, achieves overall robustness. We demonstrate that , by maximizing the use of adversaries via focusing on local balls, we achieve h igh robust accuracy with weak adversaries. Specifically, our method reaches a si milar robust accuracy level to the state of the art approaches trained on strong adversaries on MNIST, CIFAR-10 and CIFAR-100. As a result, the overall training time is reduced. Furthermore, when trained with strong adversaries, our method matches with the current state of the art on MNIST and outperforms them on CIFAR -10 and CIFAR-100.
****
On the Reproducibility of Neural Network Predictions
Srinadh Bhojanapalli,Kimberly Jenney Wilber,Andreas Veit,Ankit Singh Rawat,Seung yeon Kim,Aditya Krishna Menon,Sanjiv Kumar
Standard training techniques for neural networks involve multiple sources of ran domness, e.g., initialization, mini-batch ordering and in some cases data augmen tation. Given that neural networks are heavily over-parameterized in practice, s uch randomness can cause {\em churn} -- disagreements between predictions of the two models independently trained by the same algorithm, contributing to the `re producibility challenges' in modern machine learning. In this paper, we study th is problem of churn, identify factors that cause it, and propose two simple mean s of mitigating it. We first demonstrate that churn is indeed an issue, even for standard image classification tasks (CIFAR and ImageNet), and study the role of the different sources of training randomness that cause churn. By analyzing the relationship between churn and prediction confidences, we pursue an approach wi th two components for churn reduction. First, we propose using \emph{minimum ent ropy regularizers} to increase prediction confidences. Second, we present a nov el variant of co-distillation approach~\citep{anil2018large} to increase model a greement and reduce churn. We present empirical results showing the effectivenes s of both techniques in reducing churn while improving the accuracy of the under lying model.
****
Are all negatives created equal in contrastive instance discrimination?
Tiffany Cai,Jonathan Frankle,David J. Schwab,Ari S. Morcos
Self-supervised learning has recently begun to rival supervised learning on comp uter vision tasks. Many of the recent approaches have been based on contrastive instance discrimination (CID), in which the network is trained to recognize two augmented versions of the same instance (a query and positive while discriminati ng against a pool of other instances (negatives). Using MoCo v2 as our testbed, we divided negatives by their difficulty for a given query and studied which dif ficulty ranges were most important for learning useful representations. We found that a small minority of negatives--just the hardest 5%--were both necessary an d sufficient for the downstream task to reach full accuracy. Conversely, the eas iest 95% of negatives were unnecessary and insufficient. Moreover, we found that the very hardest 0.1% of negatives were not only unnecessary but also detriment al. Finally, we studied the properties of negatives that affect their hardness, and found that hard negatives were more semantically similar to the query, and t hat some negatives were more consistently easy or hard than we would expect by c hance. Together, our results indicate that negatives play heterogeneous roles an d CID may benefit from more intelligent negative treatment.
****
NAHAS: Neural Architecture and Hardware Accelerator Search
Yanqi Zhou,Xuanyi Dong,Daiyi Peng,Ethan Zhu,Amir Yazdanbakhsh,Berkin Akin,Mingxi ng Tan,James Laudon
Neural architectures and hardware accelerators have been two driving forces for the rapid progress in deep learning.
Although previous works have optimized either neural architectures given fixed h ardware, or hardware given fixed neural architectures, none has considered optim izing them jointly. In this paper, we study the importance of co-designing neura l architectures and hardware accelerators. To this end, we propose NAHAS, an aut omated hardware design paradigm that jointly searches for the best configuration for both neural architecture and accelerator. In NAHAS, accelerator hardware de

sign is conditioned on the dynamically explored neural networks for the targeted application, instead of fixed architectures, thus providing better performance opportunities. Our experiments with an industry-standard edge accelerator show that NAHAS consistently outperforms previous platform-aware neural architecture search and state-of-the-art EfficientNet on all latency targets by 0.5% - 1% ImageNet top-1 accuracy, while reducing latency by about 20%. Joint optimization reduces the search samples by 2x and reduces the latency constraint violations from 3 violations to 1 violation per 4 searches, compared to independently optimizing the two sub spaces.
****

SyncTwin: Transparent Treatment Effect Estimation under Temporal Confounding
Zhaozhi Qian,Yao Zhang,Ioana Bica,Angela Wood,Mihaela van der Schaar
Estimating causal treatment effects using observational data is a problem with few solutions when the confounder has a temporal structure, e.g. the history of disease progression might impact both treatment decisions and clinical outcomes. For such a challenging problem, it is desirable for the method to be transparent --- the ability to pinpoint a small subset of data points that contributes most to the estimate and to clearly indicate whether the estimate is reliable or not. This paper develops a new method, SyncTwin, to overcome temporal confounding in a transparent way. SyncTwin estimates the treatment effect of a target individual by comparing the outcome with its synthetic twin, which is constructed to closely match the target in the representation of the temporal confounders. SyncTwin achieves transparency by enforcing the synthetic twin to only depend on the weighted combination of few other individuals in the dataset. Moreover, the quality of the synthetic twin can be assessed by a performance metric, which also indicates the reliability of the estimated treatment effect. Experiments demonstrate that SyncTwin outperforms the benchmarks in clinical observational studies while still being transparent.
****

Linking average- and worst-case perturbation robustness via class selectivity and dimensionality
Matthew L Leavitt,Ari S. Morcos
Representational sparsity is known to affect robustness to input perturbations in deep neural networks (DNNs), but less is known about how the semantic content of representations affects robustness. Class selectivity—the variability of a unit’s responses across data classes or dimensions—is one way of quantifying the sparsity of semantic representations. Given recent evidence that class selectivity may not be necessary for, and in some cases can impair generalization, we sought to investigate whether it also confers robustness (or vulnerability) to perturbations of input data. We found that class selectivity leads to increased vulnerability to average-case (naturalistic) perturbations in ResNet18, ResNet50, and ResNet20, as measured using Tiny ImageNetC (ResNet18 and ResNet50) and CIFAR10C (ResNet20). Networks regularized to have lower levels of class selectivity are more robust to average-case perturbations, while networks with higher class selectivity are more vulnerable. In contrast, we found that class selectivity increases robustness to multiple types of worst-case (i.e. white box adversarial) perturbations, suggesting that while decreasing class selectivity is helpful for average-case perturbations, it is harmful for worst-case perturbations. To explain this difference, we studied the dimensionality of the networks' representations: we found that the dimensionality of early-layer representations is inversely proportional to a network's class selectivity, and that adversarial samples cause a larger increase in early-layer dimensionality than corrupted samples. We also found that the input-unit gradient was more variable across samples and units in high-selectivity networks compared to low-selectivity networks. These results lead to the conclusion that units participate more consistently in low-selectivity regimes compared to high-selectivity regimes, effectively creating a larger attack surface and hence vulnerability to worst-case perturbations.
****

Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning

Tim Sainburg,Leland McInnes,Timothy Q Gentner
We propose Parametric UMAP, a parametric variation of the UMAP (Uniform Manifold Approximation and Projection) algorithm. UMAP is a non-parametric graph-based dimensionality reduction algorithm using applied Riemannian geometry and algebraic topology to find low-dimensional embeddings of structured data. The UMAP algorithm consists of two steps: (1) Compute a graphical representation of a dataset (fuzzy simplicial complex), and (2) Through stochastic gradient descent, optimize a low-dimensional embedding of the graph. Here, we replace the second step of UMAP with a deep neural network that learns a parametric relationship between data and embedding. We demonstrate that our method performs similarly to its non-parametric counterpart while conferring the benefit of a learned parametric mapping (e.g. fast online embeddings for new data). We then show that UMAP loss can be extended to arbitrary deep learning applications, for example constraining the latent distribution of autoencoders, and improving classifier accuracy for semi-supervised learning by capturing structure in unlabeled data.
****
Second-Moment Loss: A Novel Regression Objective for Improved Uncertainties
Joachim Sicking,Maram Akila,Maximilian Alexander Pintz,Tim Wirtz,Asja Fischer,Stefan Wrobel
Quantification of uncertainty is one of the most promising approaches to establish safe machine learning. Despite its importance, it is far from being generally solved, especially for neural networks. One of the most commonly used approaches so far is Monte Carlo dropout, which is computationally cheap and easy to apply in practice. However, it can underestimate the uncertainty. We propose a new objective, referred to as second-moment loss (SML), to address this issue. While the full network is encouraged to model the mean, the dropout networks are explicitly used to optimize the model variance. We analyze the performance of the new objective on various toy and UCI regression datasets. Comparing to the state-of-the-art of deep ensembles, SML leads to comparable prediction accuracies and uncertainty estimates while only requiring a single model. Under distribution shift, we observe moderate improvements. From a safety perspective also the study of worst-case uncertainties is crucial. In this regard we improve considerably. Finally, we show that SML can be successfully applied to SqueezeDet, a modern object detection network. We improve on its uncertainty-related scores while not deteriorating regression quality. As a side result, we introduce an intuitive Wasserstein distance-based uncertainty measure that is non-saturating and thus allows to resolve quality differences between any two uncertainty estimates.
****
Adversarial Boot Camp: label free certified robustness in one epoch
Ryan Campbell,Chris Finlay,Adam M Oberman
Machine learning models are vulnerable to adversarial attacks. One approach to addressing this vulnerability is certification, which focuses on models that are guaranteed to be robust for a given perturbation size. A drawback of recent certified models is that they are stochastic: they require multiple computationally expensive model evaluations with random noise added to a given image. In our work, we present a deterministic certification approach which results in a certifiably robust model. This approach is based on an equivalence between training with a particular regularized loss, and the expected values of Gaussian averages. We achieve certified models on ImageNet-1k by retraining a model with this loss for one epoch without the use of label information.
****
Learning Neural Event Functions for Ordinary Differential Equations
Ricky T. Q. Chen,Brandon Amos,Maximilian Nickel
The existing Neural ODE formulation relies on an explicit knowledge of the termination time. We extend Neural ODEs to implicitly defined termination criteria modeled by neural event functions, which can be chained together and differentiated through. Neural Event ODEs are capable of modeling discrete and instantaneous changes in a continuous-time system, without prior knowledge of when these changes should occur or how many such changes should exist. We test our approach in modeling hybrid discrete- and continuous- systems such as switching dynamical sys

tems and collision in multi-body systems, and we propose simulation-based traini ng of point processes with applications in discrete control.
****

Neural Spatio-Temporal Point Processes
Ricky T. Q. Chen,Brandon Amos,Maximilian Nickel
We propose a new class of parameterizations for spatio-temporal point processes which leverage Neural ODEs as a computational method and enable flexible, high-f idelity models of discrete events that are localized in continuous time and spac e. Central to our approach is a combination of continuous-time neural networks w ith two novel neural architectures, \ie, Jump and Attentive Continuous-time Norm alizing Flows. This approach allows us to learn complex distributions for both t he spatial and temporal domain and to condition non-trivially on the observed ev ent history. We validate our models on data sets from a wide variety of contexts such as seismology, epidemiology, urban mobility, and neuroscience.
****

Proximal Gradient Descent-Ascent: Variable Convergence under K■ Geometry
Ziyi Chen,Yi Zhou,Tengyu Xu,Yingbin Liang
The gradient descent-ascent (GDA) algorithm has been widely applied to solve min imax optimization problems. In order to achieve convergent policy parameters for minimax optimization, it is important that GDA generates convergent variable se quences rather than convergent sequences of function value or gradient norm. How ever, the variable convergence of GDA has been proved only under convexity geome tries, and it is lack of understanding in general nonconvex minimax optimization . This paper fills such a gap by studying the convergence of a more general prox imal-GDA for regularized nonconvex-strongly-concave minimax optimization. Specif ically, we show that proximal-GDA admits a novel Lyapunov function, which monoto nically decreases in the minimax optimization process and drives the variable se quences to a critical point. By leveraging this Lyapunov function and the KL geo metry that parameterizes the local geometries of general nonconvex functions, we formally establish the variable convergence of proximal-GDA to a certain critic al point x^*, i.e., $x_t\to x^*, y_t\to y^*(x^*)$. Furthermore, over the full spectrum of the KL-parameterized geometry, we show that proximal-GDA achieves di fferent types of convergence rates ranging from sublinear convergence up to fini te-step convergence, depending on the geometry associated with the KL parameter. This is the first theoretical result on the variable convergence for nonconvex minimax optimization.
****

Adaptive Universal Generalized PageRank Graph Neural Network
Eli Chien,Jianhao Peng,Pan Li,Olgica Milenkovic
In many important graph data processing applications the acquired information in cludes both node features and observations of the graph topology. Graph neural n etworks (GNNs) are designed to exploit both sources of evidence but they do not optimally trade-off their utility and integrate them in a manner that is also un iversal. Here, universality refers to independence on homophily or heterophily g raph assumptions. We address these issues by introducing a new Generalized PageR ank (GPR) GNN architecture that adaptively learns the GPR weights so as to joint ly optimize node feature and topological information extraction, regardless of t he extent to which the node labels are homophilic or heterophilic. Learned GPR w eights automatically adjust to the node label pattern, irrelevant on the type of initialization, and thereby guarantee excellent learning performance for label patterns that are usually hard to handle. Furthermore, they allow one to avoid f eature over-smoothing, a process which renders feature information nondiscrimina tive, without requiring the network to be shallow. Our accompanying theoretical analysis of the GPR-GNN method is facilitated by novel synthetic benchmark datas ets generated by the so-called contextual stochastic block model. We also compar e the performance of our GNN architecture with that of several state-of-the-art GNNs on the problem of node-classification, using well-known benchmark homophili c and heterophilic datasets. The results demonstrate that GPR-GNN offers signifi cant performance improvement compared to existing techniques on both synthetic a nd benchmark data.

****
Open Question Answering over Tables and Text
Wenhu Chen,Ming-Wei Chang,Eva Schlinger,William Yang Wang,William W. Cohen
In open question answering (QA), the answer to a question is produced by retriev ing and then analyzing documents that might contain answers to the question. Mo st open QA systems have considered only retrieving information from unstructured text. Here we consider for the first time open QA over {\em both} tabular and textual data and present a new large-scale dataset \emph{Open Table-and-Text Que stion Answering} (OTT-QA) to evaluate performance on this task. Most questions i n OTT-QA require multi-hop inference across tabular data and unstructured text, and the evidence required to answer a question can be distributed in different w ays over these two types of input, making evidence retrieval challenging---our b aseline model using an iterative retriever and BERT-based reader achieves an exa ct match score less than 10\%. We then propose two novel techniques to address t he challenge of retrieving and aggregating evidence for OTT-QA. The first techni que is to use ``early fusion'' to group multiple highly relevant tabular and tex tual units into a fused block, which provides more context for the retriever to search for. The second technique is to use a cross-block reader to model the cr oss-dependency between multiple retrieved evidence with global-local sparse atte ntion. Combining these two techniques improves the score significantly, to above 27\%.
****
Deep Goal-Oriented Clustering
Yifeng Shi,Christopher M Bender,Linnea Olsson,Melissa Troester,Katherine A Hoadl ey,Junier Oliva,Marc Niethammer
Clustering and prediction are two primary tasks in the fields of unsupervised an d supervised learning, respectively. Although much of the recent advances in mac hine learning have been centered around those two tasks, the interdependent, mut ually beneficial relationship between them is rarely explored. One could reasona bly expect appropriately clustering the data would aid the downstream prediction task and, conversely, a better prediction performance for the downstream task c ould potentially inform a more appropriate clustering strategy. In this work, we focus on the latter part of this mutually beneficial relationship. To this end, we introduce Deep Goal-Oriented Clustering (DGC), a probabilistic framework tha t clusters the data by jointly using supervision via side-information and unsupe rvised modeling of the inherent data structure in an end-to-end fashion. We show the effectiveness of our model on a range of datasets by achieving prediction a ccuracies comparable to the state-of-the-art, while, more importantly in our set ting, simultaneously learning congruent clustering strategies.
****
Certified Watermarks for Neural Networks
Arpit Amit Bansal,Ping-yeh Chiang,Michael Curry,Hossein Souri,Rama Chellappa,Joh n P Dickerson,Rajiv Jain,Tom Goldstein
Watermarking is a commonly used strategy to protect creators' rights to digital images, videos and audio. Recently, watermarking methods have been extended to d eep learning models -- in principle, the watermark should be preserved when an a dversary tries to copy the model. However, in practice, watermarks can often be removed by an intelligent adversary. Several papers have proposed watermarking m ethods that claim to be empirically resistant to different types of removal atta cks, but these new techniques often fail in the face of new or better-tuned adve rsaries. In this paper, we propose the first certifiable watermarking method. Us ing the randomized smoothing technique proposed in Chiang et al., we show that o ur watermark is guaranteed to be unremovable unless the model parameters are cha nged by more than a certain ℓ_2 threshold. In addition to being certifiable , our watermark is also empirically more robust compared to previous watermarkin g methods.
****
Analysing the Update step in Graph Neural Networks via Sparsification
changmin wu,Johannes F. Lutzeyer,Michalis Vazirgiannis
In recent years, Message-Passing Neural Networks (MPNNs), the most prominent Gra

ph Neural Network (GNN) framework, have celebrated much success in the analysis of graph-structured data. In MPNNs the computations are split into three steps, Aggregation, Update and Readout. In this paper a series of models to successively sparsify the linear transform in the Update step is proposed. Specifically, the ExpanderGNN model with a tuneable sparsification rate and the Activation-Only GNN, which has no linear transform in the Update step, are proposed. In agreement with a growing trend in the relevant literature the sparsification paradigm is changed by initialising sparse neural network architectures rather than expensively sparsifying already trained architectures. These novel benchmark models enable a better understanding of the influence of the Update step on model performance and outperform existing simplified benchmark models such as the Simple Graph Convolution (SGC). The ExpanderGNNs, and in some cases the Activation-Only models, achieve performance on par with their vanilla counterparts on several down-stream graph prediction tasks, while containing exponentially fewer trainable parameters. In experiments with matching parameter numbers our benchmark models outperform the state-of-the-art GNNs models. These observations enable us to conclude that in practice the Update step often makes no positive contribution to the model performance.

****

Rethinking Attention with Performers

Krzysztof Marcin Choromanski,Valerii Likhosherstov,David Dohan,Xingyou Song,Andreea Gane,Tamas Sarlos,Peter Hawkins,Jared Quincy Davis,Afroz Mohiuddin,Lukasz Kaiser,David Benjamin Belanger,Lucy J Colwell,Adrian Weller

We introduce Performers, Transformer architectures which can estimate regular (softmax) full-rank-attention Transformers with provable accuracy, but using only linear (as opposed to quadratic) space and time complexity, without relying on any priors such as sparsity or low-rankness. To approximate softmax attention-kernels, Performers use a novel Fast Attention Via positive Orthogonal Random features approach (FAVOR+), which may be of independent interest for scalable kernel methods. FAVOR+ can also be used to efficiently model kernelizable attention mechanisms beyond softmax. This representational power is crucial to accurately compare softmax with other kernels for the first time on large-scale tasks, beyond the reach of regular Transformers, and investigate optimal attention-kernels. Performers are linear architectures fully compatible with regular Transformers and with strong theoretical guarantees: unbiased or nearly-unbiased estimation of the attention matrix, uniform convergence and low estimation variance. We tested Performers on a rich set of tasks stretching from pixel-prediction through text models to protein sequence modeling. We demonstrate competitive results with other examined efficient sparse and dense attention methods, showcasing effectiveness of the novel attention-learning paradigm leveraged by Performers.

****

Text Generation by Learning from Demonstrations

Richard Yuanzhe Pang,He He

Current approaches to text generation largely rely on autoregressive models and maximum likelihood estimation. This paradigm leads to (i) diverse but low-quality samples due to mismatched learning objective and evaluation metric (likelihood vs. quality) and (ii) exposure bias due to mismatched history distributions (gold vs. model-generated). To alleviate these problems, we frame text generation as an offline reinforcement learning (RL) problem with expert demonstrations (i.e., the reference), where the goal is to maximize quality given model-generated histories. We propose GOLD (generation by off-policy learning from demonstrations): an easy-to-optimize algorithm that learns from the demonstrations by importance weighting. Intuitively, GOLD upweights confident tokens and downweights unconfident ones in the reference during training, avoiding optimization issues faced by prior RL approaches that rely on online data collection. According to both automatic and human evaluation, models trained by GOLD outperform those trained by MLE and policy gradient on summarization, question generation, and machine translation. Further, our models are less sensitive to decoding algorithms and alleviate exposure bias.

****

On the Effectiveness of Weight-Encoded Neural Implicit 3D Shapes
Thomas Ryan Davies,Derek Nowrouzezahrai,Alec Jacobson
A neural implicit outputs a number indicating whether the given query point in space is inside, outside, or on a surface. Many prior works have focused on _latent-encoded_ neural implicits, where a latent vector encoding of a specific shape is also fed as input. While affording latent-space interpolation, this comes at the cost of reconstruction accuracy for any _single_ shape. Training a specific network for each 3D shape, a _weight-encoded_ neural implicit may forgo the latent vector and focus reconstruction accuracy on the details of a single shape. While previously considered as an intermediary representation for 3D scanning tasks or as a toy-problem leading up to latent-encoding tasks, weight-encoded neural implicits have not yet been taken seriously as a 3D shape representation. In this paper, we establish that weight-encoded neural implicits meet the criteria of a first-class 3D shape representation. We introduce a suite of technical contributions to improve reconstruction accuracy, convergence, and robustness when learning the signed distance field induced by a polygonal mesh --- the _de facto_ standard representation. Viewed as a lossy compression, our conversion outperforms standard techniques from geometry processing. Compared to previous latent- and weight-encoded neural implicits we demonstrate superior robustness, scalability, and performance.
****
FORK: A FORward-looKing Actor for Model-Free Reinforcement Learning
Honghao Wei,Lei Ying
In this paper, we propose a new type of Actor, named forward-looking Actor or FORK for short, for Actor-Critic algorithms. FORK can be easily integrated into a model-free Actor-Critic algorithm. Our experiments on six Box2D and MuJoCo environments with continuous state and action spaces demonstrate significant performance improvement FORK can bring to the state-of-the-art algorithms. A variation of FORK can further solve BipedalWalkerHardcore in as few as four hours using a single GPU.
****
Offline Adaptive Policy Leaning in Real-World Sequential Recommendation Systems
Xiong-Hui Chen,Yang Yu,Qingyang Li,Zhiwei Tony Qin,Wenjie Shang,Yiping Meng,Jieping Ye
The training process of RL requires many trial-and-errors that are costly in real-world applications. To avoid the cost, a promising solution is to learn the policy from an offline dataset, e.g., to learn a simulator from the dataset, and train optimal policies in the simulator. By this approach, the quality of policies highly relies on the fidelity of the simulator. Unfortunately, due to the stochasticity and unsteadiness of the real-world and the unavailability of online sampling, the distortion of the simulator is inevitable. In this paper, based on the model learning technique, we propose a new paradigm to learn an RL policy from offline data in the real-world sequential recommendation system (SRS). Instead of increasing the fidelity of models for policy learning, we handle the distortion issue via learning to adapt to diverse simulators generated by the offline dataset. The adaptive policy is suitable to real-world environments where dynamics are changing and have stochasticity in the offline setting. Experiments are conducted in synthetic environments and a real-world ride-hailing platform. The results show that the method overcomes the distortion problem and produces robust recommendations in the unseen real-world.
****
The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation
Thibault Sejourne,François-Xavier Vialard,Gabriel Peyré
Comparing metric measure spaces (i.e. a metric space endowed with a probability distribution) is at the heart of many machine learning problems. This includes for instance predicting properties of molecules in quantum chemistry or generating graphs with varying connectivity. The most popular distance between such metric measure spaces is the Gromov-Wasserstein (GW) distance, which is the solution of a quadratic assignment problem. This distance has been successfully applied to supervised learning and generative modeling, for applications as diverse as qu

antum chemistry or natural language processing. The GW distance is however limit ed to the comparison of metric measure spaces endowed with a probability distrib ution. This strong limitation is problematic for many applications in ML where t here is no a priori natural normalization on the total mass of the data. Further more, imposing an exact conservation of mass across spaces is not robust to outl iers and often leads to irregular matching. To alleviate these issues, we introd uce two Unbalanced Gromov-Wasserstein formulations: a distance and a more tracta ble upper-bounding relaxation. They both allow the comparison of metric spaces e quipped with arbitrary positive measures up to isometries. The first formulation is a positive and definite divergence based on a relaxation of the mass conserv ation constraint using a novel type of quadratically-homogeneous divergence.This divergence works hand in hand with the entropic regularization approach which i s popular to solve large scale optimal transport problems. We show that the unde rlying non-convex optimization problem can be efficiently tackled using a highly parallelizable and GPU-friendly iterative scheme. The second formulation is a d istance between mm-spaces up to isometries based on a conic lifting. Lastly, we provide numerical simulations to highlight the salient features of the unbalance d divergence and its potential applications in ML.
****

Fast Partial Fourier Transform
Yong-chan Park,Jun-Gi Jang,U Kang
Given a time-series vector, how can we efficiently compute a specified part of F ourier coefficients? Fast Fourier transform (FFT) is a widely used algorithm tha t computes the discrete Fourier transform in many machine learning applications. Despite the pervasive use, FFT algorithms do not provide a fine-tuning option f or the user to specify one’s demand, that is, the output size (the number of Fou rier coefficients to be computed) is algorithmically determined by the input siz e. Such a lack of flexibility is often followed by just discarding the unused co efficients because many applications do not require the whole spectrum of the fr equency domain, resulting in an inefficiency due to the extra computation.
In this paper, we propose a fast Partial Fourier Transform (PFT), an efficient a lgorithm for computing only a part of Fourier coefficients. PFT approximates a p art of twiddle factors (trigonometric constants) using polynomials, thereby redu cing the computational complexity due to the mixture of many twiddle factors. We derive the asymptotic time complexity of PFT with respect to input and output s izes, as well as its numerical accuracy. Experimental results show that PFT outp erforms the current state-of-the-art algorithms, with an order of magnitude of s peedup for sufficiently small output sizes without sacrificing accuracy.
****

Support-set bottlenecks for video-text representation learning
Mandela Patrick,Po-Yao Huang,Yuki Asano,Florian Metze,Alexander G Hauptmann,Joao F. Henriques,Andrea Vedaldi
The dominant paradigm for learning video-text representations – noise contrastiv e learning – increases the similarity of the representations of pairs of samples that are known to be related, such as text and video from the same sample, and pushes away the representations of all other pairs. We posit that this last beha viour is too strict, enforcing dissimilar representations even for samples that are semantically-related – for example, visually similar videos or ones that sha re the same depicted action. In this paper, we propose a novel method that allev iates this by leveraging a generative model to naturally push these related samp les together: each sample’s caption must be reconstructed as a weighted combinat ion of a support set of visual representations. This simple idea ensures that re presentations are not overly-specialized to individual samples, are reusable acr oss the dataset, and results in representations that explicitly encode semantics shared between samples, unlike noise contrastive learning. Our proposed method outperforms others by a large margin on MSR-VTT, VATEX, ActivityNet, and MSVD fo r video-to-text and text-to-video retrieval.
****

Multi-modal Self-Supervision from Generalized Data Transformations
Mandela Patrick,Yuki Asano,Polina Kuznetsova,Ruth Fong,Joao F. Henriques,Geoffre

y Zweig,Andrea Vedaldi
In the image domain, excellent representation can be learned by inducing invaria nce to content-preserving transformations, such as image distortions. In this pa per, we show that, for videos, the answer is more complex, and that better resul ts can be obtained by accounting for the interplay between invariance, distincti veness, multiple modalities and time. We introduce Generalized Data Transformat ions (GDTs) as a way to capture this interplay. GDTs reduce most previous self-s upervised approaches to a choice of data transformations, even when this was not the case in the original formulations. They also allow to choose whether the re presentation should be invariant or distinctive w.r.t. each effect and tell whic h combinations are valid, thus allowing us to explore the space of combinations systematically. We show in this manner that being invariant to certain transform ations and distinctive to others is critical to learning effective video represe ntations, improving the state-of-the-art by a large margin, and even surpassing supervised pretraining. We demonstrate results on a variety of downstream video and audio classification and retrieval tasks, on datasets such as HMDB-51, UCF-1 01, DCASE2014, ESC-50 and VGG-Sound. In particular, we achieve new state-of-the- art accuracies of 72.8% on HMDB-51 and 95.2% on UCF-101.
****

Tilted Empirical Risk Minimization
Tian Li,Ahmad Beirami,Maziar Sanjabi,Virginia Smith
Empirical risk minimization (ERM) is typically designed to perform well on the a verage loss, which can result in estimators that are sensitive to outliers, gene ralize poorly, or treat subgroups unfairly. While many methods aim to address th ese problems individually, in this work, we explore them through a unified frame work---tilted empirical risk minimization (TERM). In particular, we show that it is possible to flexibly tune the impact of individual losses through a straight forward extension to ERM using a hyperparameter called the tilt. We provide seve ral interpretations of the resulting framework: We show that TERM can increase o r decrease the influence of outliers, respectively, to enable fairness or robust ness; has variance-reduction properties that can benefit generalization; and can be viewed as a smooth approximation to a superquantile method. We develop batch and stochastic first-order optimization methods for solving TERM, and show that the problem can be efficiently solved relative to common alternatives. Finally, we demonstrate that TERM can be used for a multitude of applications, such as e nforcing fairness between subgroups, mitigating the effect of outliers, and hand ling class imbalance. TERM is not only competitive with existing solutions tailo red to these individual problems, but can also enable entirely new applications, such as simultaneously addressing outliers and promoting fairness.
****

Uncertainty-Based Adaptive Learning for Reading Comprehension
Jing Wang,Jie Shen,Xiaofei Ma,Andrew Arnold
Recent years have witnessed a surge of successful applications of machine readin g comprehension. Of central importance to the tasks is the availability of massi ve amount of labeled data, which facilitates the training of large-scale neural networks. However, in many real-world problems, annotated data are expensive to gather not only because of time cost and budget, but also of certain domain-spec ific restrictions such as privacy for healthcare data. In this regard, we propos e an uncertainty-based adaptive learning algorithm for reading comprehension, wh ich interleaves data annotation and model updating to mitigate the demand of lab eling. Our key techniques are two-fold: 1) an unsupervised uncertainty-based sam pling scheme that queries the labels of the most informative instances with resp ect to the currently learned model; and 2) an adaptive loss minimization paradig m that simultaneously fits the data and controls the degree of model updating. W e demonstrate on the benchmark datasets that 25\% less labeled samples suffice t o guarantee similar, or even improved performance. Our results demonstrate a str ong evidence that for label-demanding scenarios, the proposed approach offers a practical guide on data collection and model training.
****

Graph Pooling by Edge Cut

Alexis Galland,marc lelarge
Graph neural networks (GNNs) are very efficient at solving several tasks in graphs such as node classification or graph classification. They come from an adaptation of convolutional neural networks on images to graph structured data. These models are very effective at finding patterns in images that can discriminate images from each others. Another aspect leading to their success is their ability to uncover hierarchical structures. This comes from the pooling operation that produces different versions of the input image at different scales. The same way, we want to identify patterns at different scales in graphs in order to improve the classification accuracy. Compared to the case of images, it is not trivial to develop a pooling layer on graphs. This is mainly due to the fact that in graphs nodes are not ordered and have irregular neighborhoods. To aleviate this issue, we propose a pooling layer based on edge cuts in graphs. This pooling layer works by computing edge scores that correspond to the importance of edges in the process of information propagation of the GNN. Moreover, we define a regularization function that aims at producing edge scores that minimize the minCUT problem. Finally, through extensive experiments we show that this architecture can compete with state-of- the-art methods.
****

Explainable Subgraph Reasoning for Forecasting on Temporal Knowledge Graphs
Zhen Han,Peng Chen,Yunpu Ma,Volker Tresp
Modeling time-evolving knowledge graphs (KGs) has recently gained increasing interest. Here, graph representation learning has become the dominant paradigm for link prediction on temporal KGs. However, the embedding-based approaches largely operate in a black-box fashion, lacking the ability to interpret their predictions. This paper provides a link forecasting framework that reasons over query-relevant subgraphs of temporal KGs and jointly models the structural dependencies and the temporal dynamics. Especially, we propose a temporal relational attention mechanism and a novel reverse representation update scheme to guide the extraction of an enclosing subgraph around the query. The subgraph is expanded by an iterative sampling of temporal neighbors and by attention propagation. Our approach provides human-understandable evidence explaining the forecast. We evaluate our model on four benchmark temporal knowledge graphs for the link forecasting task. While being more explainable, our model obtains a relative improvement of up to 20 $\%$ on Hits@1 compared to the previous best temporal KG forecasting method. We also conduct a survey with 53 respondents, and the results show that the evidence extracted by the model for link forecasting is aligned with human understanding.
****

Grounded Language Learning Fast and Slow
Felix Hill,Olivier Tieleman,Tamara von Glehn,Nathaniel Wong,Hamza Merzic,Stephen Clark
Recent work has shown that large text-based neural language models acquire a surprising propensity for one-shot learning. Here, we show that an agent situated in a simulated 3D world, and endowed with a novel dual-coding external memory, can exhibit similar one-shot word learning when trained with conventional RL algorithms. After a single introduction to a novel object via visual perception and language ("This is a dax"), the agent can manipulate the object as instructed ("Put the dax on the bed"), combining short-term, within-episode knowledge of the nonsense word with long-term lexical and motor knowledge. We find that, under certain training conditions and with a particular memory writing mechanism, the agent's one-shot word-object binding generalizes to novel exemplars within the same ShapeNet category, and is effective in settings with unfamiliar numbers of objects. We further show how dual-coding memory can be exploited as a signal for intrinsic motivation, stimulating the agent to seek names for objects that may be useful later. Together, the results demonstrate that deep neural networks can exploit meta-learning, episodic memory and an explicitly multi-modal environment to account for 'fast-mapping', a fundamental pillar of human cognitive development and a potentially transformative capacity for artificial agents.
****

Transferable Recognition-Aware Image Processing
Zhuang Liu,Tinghui Zhou,Hung-Ju Wang,Zhiqiang Shen,Bingyi Kang,Evan Shelhamer,Trevor Darrell
Recent progress in image recognition has stimulated the deployment of vision systems at an unprecedented scale. As a result, visual data are now often consumed not only by humans but also by machines. Existing image processing methods only optimize for better human perception, yet the resulting images may not be accurately recognized by machines. This can be undesirable, e.g., the images can be improperly handled by search engines or recommendation systems. In this work, we propose simple approaches to improve machine interpretability of processed images: optimizing the recognition loss directly on the image processing network or through an intermediate transforming model. Interestingly, the processing model's ability to enhance recognition quality can transfer when evaluated on models of different architectures, recognized categories, tasks and training datasets. This makes the solutions applicable even when we do not have the knowledge of future recognition models, e.g., if we upload processed images to the Internet. We conduct experiments on multiple image processing tasks, with ImageNet classification and PASCAL VOC detection as recognition tasks. With our simple methods, substantial accuracy gain can be achieved with strong transferability and minimal image quality loss. Through a user study we further show that the accuracy gain can transfer to a black-box, third-party cloud model. Finally, we try to explain this transferability phenomenon by demonstrating the similarities of different models' decision boundaries. Code is available at https://github.com/anonymous20202020/Transferable_RA .
****
Graph Structural Aggregation for Explainable Learning
Alexis Galland,marc lelarge
Graph neural networks have proven to be very efficient to solve several tasks in graphs such as node classification or link prediction. These algorithms that operate by propagating information from vertices to their neighbors allow one to build node embeddings that contain local information. In order to use graph neural networks for graph classification, node embeddings must be aggregated to obtain a graph representation able to discriminate among different graphs (of possibly various sizes). Moreover, in analogy to neural networks for image classification, there is a need for explainability regarding the features that are selected in the graph classification process. To this end, we introduce StructAgg, a simple yet effective aggregation process based on the identification of structural roles for nodes in graphs that we use to create an end-to-end model. Through extensive experiments we show that this architecture can compete with state-of-the-art methods. We show how this aggregation step allows us to cluster together nodes that have comparable structural roles and how these roles provide explainability to this neural network model.

****
LEARNED HARDWARE/SOFTWARE CO-DESIGN OF NEURAL ACCELERATORS
Zhan Shi,Chirag Sakhuja,Milad Hashemi,Kevin Swersky,Calvin Lin
The use of deep learning has grown at an exponential rate, giving rise to numerous specialized hardware and software systems for deep learning. Because the design space of deep learning software stacks and hardware accelerators is diverse and vast, prior work considers software optimizations separately from hardware architectures, effectively reducing the search space. Unfortunately, this bifurcated approach means that many profitable design points are never explored. This paper instead casts the problem as hardware/software co-design, with the goal of automatically identifying desirable points in the joint design space. The key to our solution is a new constrained Bayesian optimization framework that avoids invalid solutions by exploiting the highly constrained features of this design space, which are semi-continuous/semi-discrete. We evaluate our optimization framework by applying it to a variety of neural models, improving the energy-delay product by 18% (ResNet) and 40% (DQN) over hand-tuned state-of-the-art systems, as well as demonstrating strong results on other neural network architectures, suc

h as MLPs and Transformers.
****
Bayesian Context Aggregation for Neural Processes
Michael Volpp,Fabian Flürenbrock,Lukas Grossberger,Christian Daniel,Gerhard Neumann
Formulating scalable probabilistic regression models with reliable uncertainty estimates has been a long-standing challenge in machine learning research. Recently, casting probabilistic regression as a multi-task learning problem in terms of conditional latent variable (CLV) models such as the Neural Process (NP) has shown promising results. In this paper, we focus on context aggregation, a central component of such architectures, which fuses information from multiple context data points. So far, this aggregation operation has been treated separately from the inference of a latent representation of the target function in CLV models. Our key contribution is to combine these steps into one holistic mechanism by phrasing context aggregation as a Bayesian inference problem. The resulting Bayesian Aggregation (BA) mechanism enables principled handling of task ambiguity, which is key for efficiently processing context information. We demonstrate on a range of challenging experiments that BA consistently improves upon the performance of traditional mean aggregation while remaining computationally efficient and fully compatible with existing NP-based models.
****
Conformation-Guided Molecular Representation with Hamiltonian Neural Networks
Ziyao Li,Shuwen Yang,Guojie Song,Lingsheng Cai
Well-designed molecular representations (fingerprints) are vital to combine medical chemistry and deep learning. Whereas incorporating 3D geometry of molecules (i.e. conformations) in their representations seems beneficial, current 3D algorithms are still in infancy. In this paper, we propose a novel molecular representation algorithm which preserves 3D conformations of molecules with a Molecular Hamiltonian Network (HamNet). In HamNet, implicit positions and momentums of atoms in a molecule interact in the Hamiltonian Engine following the discretized Hamiltonian equations. These implicit coordinations are supervised with real conformations with translation- & rotation-invariant losses, and further used as inputs to the Fingerprint Generator, a message-passing neural network. Experiments show that the Hamiltonian Engine can well preserve molecular conformations, and that the fingerprints generated by HamNet achieve state-of-the-art performances on MoleculeNet, a standard molecular machine learning benchmark.
****
GAN "Steerability" without optimization
Nurit Spingarn,Ron Banner,Tomer Michaeli
Recent research has shown remarkable success in revealing "steering" directions in the latent spaces of pre-trained GANs. These directions correspond to semantically meaningful image transformations (e.g., shift, zoom, color manipulations), and have the same interpretable effect across all categories that the GAN can generate. Some methods focus on user-specified transformations, while others discover transformations in an unsupervised manner. However, all existing techniques rely on an optimization procedure to expose those directions, and offer no control over the degree of allowed interaction between different transformations. In this paper, we show that "steering" trajectories can be computed in closed form directly from the generator's weights without any form of training or optimization. This applies to user-prescribed geometric transformations, as well as to unsupervised discovery of more complex effects. Our approach allows determining both linear and nonlinear trajectories, and has many advantages over previous methods. In particular, we can control whether one transformation is allowed to come on the expense of another (e.g., zoom-in with or without allowing translation to keep the object centered). Moreover, we can determine the natural end-point of the trajectory, which corresponds to the largest extent to which a transformation can be applied without incurring degradation. Finally, we show how transferring attributes between images can be achieved without optimization, even across different categories.

****
Balancing Robustness and Sensitivity using Feature Contrastive Learning
Seungyeon Kim,Daniel Glasner,Srikumar Ramalingam,Cho-Jui Hsieh,Kishore Papineni,Sanjiv Kumar
It is generally believed that robust training of extremely large networks is critical to their success in real-world applications. However, when taken to the extreme, methods that promote robustness can hurt the model’s sensitivity to rare or underrepresented patterns. In this paper, we discuss this trade-off between robustness and sensitivity by introducing two notions: contextual feature utility and contextual feature sensitivity. We propose Feature Contrastive Learning (FCL) that encourages the model to be more sensitive to the features that have higher contextual utility. Empirical results demonstrate that models trained with FCL achieve a better balance of robustness and sensitivity, leading to improved generalization in the presence of noise.
****
Convex Regularization in Monte-Carlo Tree Search
Tuan Quang Dam,Carlo D'Eramo,Jan Peters,Joni Pajarinen
Monte-Carlo planning and Reinforcement Learning (RL) are essential to sequential decision making. The recent AlphaGo and AlphaZero algorithms have shown how to successfully combine these two paradigms to solve large scale sequential decision problems. These methodologies exploit a variant of the well-known UCT algorithm to trade off the exploitation of good actions and the exploration of unvisited states, but their empirical success comes at the cost of poor sample-efficiency and high computation time. In this paper, we overcome these limitations by studying the benefit of convex regularization in Monte-Carlo Tree Search (MCTS) to drive exploration efficiently and to improve policy updates, as already observed in RL. First, we introduce a unifying theory on the use of generic convex regularizers in MCTS, deriving the first regret analysis of regularized MCTS and showing that it guarantees an exponential convergence rate. Second, we exploit our theoretical framework to introduce novel regularized backup operators for MCTS, based on the relative entropy of the policy update and on the Tsallis entropy of the policy. We provide an intuitive demonstration of the effect of each regularizer empirically verifying the consequence of our theoretical results on a toy problem. Finally, we show how our framework can easily be incorporated in AlphaGo and AlphaZero, and we empirically show the superiority of convex regularization w.r.t. representative baselines, on well-known RL problems across several Atari games.
****
Learning with AMIGo: Adversarially Motivated Intrinsic Goals
Andres Campero,Roberta Raileanu,Heinrich Kuttler,Joshua B. Tenenbaum,Tim Rocktäschel,Edward Grefenstette
A key challenge for reinforcement learning (RL) consists of learning in environments with sparse extrinsic rewards. In contrast to current RL methods, humans are able to learn new skills with little or no reward by using various forms of intrinsic motivation. We propose AMIGo, a novel agent incorporating -- as form of meta-learning -- a goal-generating teacher that proposes Adversarially Motivated Intrinsic Goals to train a goal-conditioned "student" policy in the absence of (or alongside) environment reward. Specifically, through a simple but effective "constructively adversarial" objective, the teacher learns to propose increasingly challenging -- yet achievable -- goals that allow the student to learn general skills for acting in a new environment, independent of the task to be solved. We show that our method generates a natural curriculum of self-proposed goals which ultimately allows the agent to solve challenging procedurally-generated tasks where other forms of intrinsic motivation and state-of-the-art RL methods fail.
****
Training with Quantization Noise for Extreme Model Compression
Pierre Stock,Angela Fan,Benjamin Graham,Edouard Grave,Rémi Gribonval,Herve Jegou,Armand Joulin
We tackle the problem of producing compact models, maximizing their accuracy for

a given model size. A standard solution is to train networks with Quantization Aware Training, where the weights are quantized during training and the gradients approximated with the Straight-Through Estimator. In this paper, we extend this approach to work with extreme compression methods where the approximations introduced by STE are severe. Our proposal is to only quantize a different random subset of weights during each forward, allowing for unbiased gradients to flow through the other weights. Controlling the amount of noise and its form allows for extreme compression rates while maintaining the performance of the original model. As a result we establish new state-of-the-art compromises between accuracy and model size both in natural language processing and image classification. For example, applying our method to state-of-the-art Transformer and ConvNet architectures, we can achieve 82.5% accuracy on MNLI by compressing RoBERTa to 14 MB and 80.0% top-1 accuracy on ImageNet by compressing an EfficientNet-B3 to 3.3 MB.

****

Meta-Reinforcement Learning With Informed Policy Regularization

Pierre-Alexandre Kamienny,Matteo Pirotta,Alessandro Lazaric,Thibault Lavril,Nicolas Usunier,Ludovic Denoyer

Meta-reinforcement learning aims at finding a policy able to generalize to new environments. When facing a new environment, this policy must explore to identify its particular characteristics and then exploit this information for collecting reward. We consider the online adaptation setting where the agent needs to trade-off between the two types of behaviour within the same episode. Even though policies based on recurrent neural networks can be used in this setting by training them on multiple environments, they often fail to model this trade-off, or solve it at a very high computational cost. In this paper, we propose a new algorithm that uses privileged information in the form of a task descriptor at train time to improve the learning of recurrent policies. Our method learns an informed policy (i.e., a policy receiving as input the description of the current task) that is used to both construct task embeddings from the descriptors, and to regularize the training of the recurrent policy through parameters sharing and an auxiliary objective. This approach significantly reduces the learning sample complexity without altering the representational power of RNNs, by focusing on the relevant characteristics of the task, and by exploiting them efficiently. We evaluate our algorithm in a variety of environments that require sophisticated exploration/exploitation strategies and show that it outperforms vanilla RNNs, Thompson sampling and the task-inference approaches to meta-reinforcement learning.

****

Interpreting and Boosting Dropout from a Game-Theoretic View

Hao Zhang,Sen Li,YinChao Ma,Mingjie Li,Yichen Xie,Quanshi Zhang

This paper aims to understand and improve the utility of the dropout operation from the perspective of game-theoretical interactions. We prove that dropout can suppress the strength of interactions between input variables of deep neural networks (DNNs). The theoretical proof is also verified by various experiments. Furthermore, we find that such interactions were strongly related to the over-fitting problem in deep learning. So, the utility of dropout can be regarded as decreasing interactions to alleviating the significance of over-fitting. Based on this understanding, we propose the interaction loss to further improve the utility of dropout. Experimental results on various DNNs and datasets have shown that the interaction loss can effectively improve the utility of dropout and boost the performance of DNNs.

****

Meta-Learning Bayesian Neural Network Priors Based on PAC-Bayesian Theory

Jonas Rothfuss,Martin Josifoski,Andreas Krause

Bayesian deep learning is a promising approach towards improved uncertainty quantification and sample efficiency.

Due to their complex parameter space, choosing informative priors for Bayesian Neural Networks (BNNs) is challenging. Thus, often a naive, zero-centered Gaussian is used, resulting both in bad generalization and poor uncertainty estimates when training data is scarce. In contrast, meta-learning aims to extract such prior knowledge from a set of related learning tasks. We propose a principled and s

calable algorithm for meta-learning BNN priors based on PAC-Bayesian bounds. Whereas previous approaches require optimizing the prior and multiple variational posteriors in an interdependent manner, our method does not rely on difficult nested optimization problems and is agnostic to the variational inference method in use. Our experiments show that the proposed method is not only computationally more efficient but also yields better predictions and uncertainty estimates when compared to previous meta-learning methods and BNNs with standard priors.
****
Deep Data Flow Analysis
Chris Cummins,Zacharias Fisches,Tal Ben-Nun,Torsten Hoefler,Hugh Leather,Michael O'Boyle
Compiler architects increasingly look to machine learning when building heuristics for compiler optimization. The promise of automatic heuristic design, freeing the compiler engineer from the complex interactions of program, architecture, and other optimizations, is alluring. However, most machine learning methods cannot replicate even the simplest of the abstract interpretations of data flow analysis that are critical to making good optimization decisions. This must change for machine learning to become the dominant technology in compiler heuristics.

To this end, we propose ProGraML - Program Graphs for Machine Learning - a language-independent, portable representation of whole-program semantics for deep learning. To benchmark current and future learning techniques for compiler analyses we introduce an open dataset of 461k Intermediate Representation (IR) files for LLVM, covering five source programming languages, and 15.4M corresponding data flow results. We formulate data flow analysis as an MPNN and show that, using ProGraML, standard analyses can be learned, yielding improved performance on downstream compiler optimization tasks.
****
VTNet: Visual Transformer Network for Object Goal Navigation
Heming Du,Xin Yu,Liang Zheng
Object goal navigation aims to steer an agent towards a target object based on observations of the agent. It is of pivotal importance to design effective visual representations of the observed scene in determining navigation actions. In this paper, we introduce a Visual Transformer Network (VTNet) for learning informative visual representation in navigation. VTNet is a highly effective structure that embodies two key properties for visual representations: First, the relationships among all the object instances in a scene are exploited; Second, the spatial locations of objects and image regions are emphasized so that directional navigation signals can be learned. Furthermore, we also develop a pre-training scheme to associate the visual representations with navigation signals, and thus facilitate navigation policy learning. In a nutshell, VTNet embeds object and region features with their location cues as spatial-aware descriptors and then incorporates all the encoded descriptors through attention operations to achieve informative representation for navigation. Given such visual representations, agents are able to explore the correlations between visual observations and navigation actions. For example, an agent would prioritize ``turning right'' over ``turning left'' when the visual representation emphasizes on the right side of activation map. Experiments in the artificial environment AI2-Thor demonstrate that VTNet significantly outperforms state-of-the-art methods in unseen testing environments.
****
How to compare adversarial robustness of classifiers from a global perspective
Niklas Risse,Jan Philip Göpfert,Christina Göpfert
Adversarial robustness of machine learning models has attracted considerable attention over recent years. Adversarial attacks undermine the reliability of and trust in machine learning models, but the construction of more robust models hinges on a rigorous understanding of adversarial robustness as a property of a given model. Point-wise measures for specific threat models are currently the most popular tool for comparing the robustness of classifiers and are used in most recent publications on adversarial robustness. In this work, we use robustness curv

es to show that point-wise measures fail to capture important global properties that are essential to reliably compare the robustness of different classifiers. We introduce new ways in which robustness curves can be used to systematically uncover these properties and provide concrete recommendations for researchers and practitioners when assessing and comparing the robustness of trained models. Furthermore, we characterize scale as a way to distinguish small and large perturbations, and relate it to inherent properties of data sets, demonstrating that robustness thresholds must be chosen accordingly. We hope that our work contributes to a shift of focus away from point-wise measures of robustness and towards a discussion of the question what kind of robustness could and should reasonably be expected. We release code to reproduce all experiments presented in this paper, which includes a Python module to calculate robustness curves for arbitrary data sets and classifiers, supporting a number of frameworks, including TensorFlow, PyTorch and JAX.
****

Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization
Judy Borowski,Roland Simon Zimmermann,Judith Schepers,Robert Geirhos,Thomas S. A. Wallis,Matthias Bethge,Wieland Brendel
Feature visualizations such as synthetic maximally activating images are a widely used explanation method to better understand the information processing of convolutional neural networks (CNNs). At the same time, there are concerns that these visualizations might not accurately represent CNNs' inner workings. Here, we measure how much extremely activating images help humans to predict CNN activations.
Using a well-controlled psychophysical paradigm, we compare the informativeness of synthetic images by Olah et al. (2017) with a simple baseline visualization, namely exemplary natural images that also strongly activate a specific feature map. Given either synthetic or natural reference images, human participants choose which of two query images leads to strong positive activation. The experiment is designed to maximize participants' performance, and is the first to probe intermediate instead of final layer representations. We find that synthetic images indeed provide helpful information about feature map activations ($82\pm4\%$ accuracy; chance would be 50%). However, natural images --- originally intended to be a baseline --- outperform these synthetic images by a wide margin ($92\pm2\%$). Additionally, participants are faster and more confident for natural images, whereas subjective impressions about the interpretability of the feature visualizations by Olah et al. (2017) are mixed. The higher informativeness of natural images holds across most layers, for both expert and lay participants as well as for hand- and randomly-picked feature visualizations. Even if only a single reference image is given, synthetic images provide less information than natural images ($65\pm5\%$ vs. $73\pm4\%$). In summary, synthetic images from a popular feature visualization method are significantly less informative for assessing CNN activations than natural images. We argue that visualization methods should improve over this simple baseline.
****

Learning Representations by Contrasting Clusters While Bootstrapping Instances
Junsoo Lee,Hojoon Lee,Inkyu Shin,Jaekyoung Bae,In So Kweon,Jaegul Choo
Learning visual representations using large-scale unlabelled images is a holy grail for most of computer vision tasks. Recent contrastive learning methods have focused on encouraging the learned visual representations to be linearly separable among the individual items regardless of their semantic similarity; however, it could lead to a sub-optimal solution if a given downstream task is related to non-discriminative ones such as cluster analysis and information retrieval. In this work, we propose an advanced approach to consider the instance semantics in an unsupervised environment by both i) Contrasting batch-wise Cluster assignment features and ii) Bootstrapping an INstance representations without considering negatives simultaneously, referred to as C2BIN. Specifically, instances in a mini-batch are appropriately assigned to distinct clusters, each of which aims to capture apparent similarity among instances. Moreover, we introduce a pyramidal

multi-heads technique, showing positive effects on the representations by capturing multi-scale semantics. Empirically, our method achieves comparable or better performance than both representation learning and clustering baselines on various benchmark datasets: CIFAR-10, CIFAR-100, and STL-10.
****
Measuring Progress in Deep Reinforcement Learning Sample Efficiency
Florian E. Dorner
Sampled environment transitions are a critical input to deep reinforcement learning (DRL) algorithms. Current DRL benchmarks often allow for the cheap and easy generation of large amounts of samples such that perceived progress in DRL does not necessarily correspond to improved sample efficiency. As simulating real world processes is often prohibitively hard and collecting real world experience is costly, sample efficiency is an important indicator for economically relevant applications of DRL. We investigate progress in sample efficiency on Atari games and continuous control tasks by comparing the amount of samples that a variety of algorithms need to reach a given performance level according to training curves in the corresponding publications. We find exponential progress in sample efficiency with estimated doubling times of around 10 to 18 months on Atari, 5 to 24 months on state-based continuous control and of around 4 to 9 months on pixel-based continuous control depending on the specific task and performance level.
****
Prioritized Level Replay
Minqi Jiang,Edward Grefenstette,Tim Rocktäschel
Simulated environments with procedurally generated content have become popular benchmarks for testing systematic generalization of reinforcement learning agents. Every level in such an environment is algorithmically created, thereby exhibiting a unique configuration of underlying factors of variation, such as layout, positions of entities, asset appearances, or even the rules governing environment transitions. Fixed sets of training levels can be determined to aid comparison and reproducibility, and test levels can be held out to evaluate the generalization and robustness of agents. While prior work samples training levels in a direct way (e.g.~uniformly) for the agent to learn from, we investigate the hypothesis that different levels provide different learning progress for an agent at specific times during training. We introduce Prioritized Level Replay, a general framework for estimating the future learning potential of a level given the current state of the agent's policy. We find that temporal-difference (TD) errors, while previously used to selectively sample past transitions, also prove effective for scoring a level's future learning potential when the agent replays (that is, revisits) that level to generate entirely new episodes of experiences from it. We report significantly improved sample-efficiency and generalization on the majority of Procgen Benchmark environments as well as two challenging MiniGrid environments. Lastly, we present a qualitative analysis showing that Prioritized Level Replay induces an implicit curriculum, taking the agent gradually from easier to harder levels.
****
Fast MNAS: Uncertainty-aware Neural Architecture Search with Lifelong Learning
Jihao Liu,Yangting Sun,Ming Zhang,Boxiao Liu,Yu Liu
Sampling-based neural architecture search (NAS) always guarantees better convergence yet suffers from huge computational resources compared with gradient-based approaches, due to the rollout bottleneck -- exhaustive training for each sampled generation on proxy tasks. This work provides a general pipeline to accelerate the convergence of the rollout process as well as the RL learning process in sampling-based NAS. It is motivated by the interesting observation that both the architecture and the parameter knowledge can be transferred between different experiments and even different tasks. We first introduce an uncertainty-aware critic (value function) in PPO to utilize the architecture knowledge in previous experiments, which stabilizes the training process and reduces the searching time by 4 times. Further, a life-long knowledge pool together with a block similarity function is proposed to utilize the lifelong parameter knowledge and reduces the searching time by 2 times. It is the first to introduce block-level weight shari

ng in RL-based NAS. The block similarity function guarantees a 100% hitting rati o with strict fairness. Besides, we show a simply designed off-policy correctio n factor that enables 'replay buffer' in RL optimization and further reduces hal f of the searching time. Experiments on the MNAS search space show the proposed FNAS accelerates standard RL-based NAS process by $\sim$10x (e.g. $\sim$256 2x2 TPUv2*days / 20,000 GPU*hour $\rightarrow$ 2,000 GPU*hour for MNAS), and guarant ees better performance on various vision tasks.
****

On the Inversion of Deep Generative Models
Aviad Aberdam,Dror Simon,Michael Elad
Deep generative models (e.g. GANs and VAEs) have been developed quite extensivel y in recent years. Lately, there has been an increased interest in the inversion of such a model, i.e. given a (possibly corrupted) signal, we wish to recover t he latent vector that generated it. Building upon sparse representation theory, we define conditions that rely only on the cardinalities of the hidden layer and are applicable to any inversion algorithm (gradient descent, deep encoder, etc.), under which such generative models are invertible with a unique solution. Imp ortantly, the proposed analysis is applicable to any trained model, and does not depend on Gaussian i.i.d. weights. Furthermore, we introduce two layer-wise inv ersion pursuit algorithms for trained generative networks of arbitrary depth, wh ere one of them is accompanied by recovery guarantees. Finally, we validate our theoretical results numerically and show that our method outperforms gradient de scent when inverting such generators, both for clean and corrupted signals.
****

Filter pre-pruning for improved fine-tuning of quantized deep neural networks
Jun Nishikawa,Ryoji Ikegaya
Deep Neural Networks(DNNs) have many parameters and activation data, and these b oth are expensive to implement. One method to reduce the size of the DNN is to q uantize the pre-trained model by using a low-bit expression for weights and acti vations, using fine-tuning to recover the drop in accuracy. However, it is gener ally difficult to train neural networks which use low-bit expressions. One reaso n is that the weights in the middle layer of the DNN have a wide dynamic range a nd so when quantizing the wide dynamic range into a few bits, the step size beco mes large, which leads to a large quantization error and finally a large degrada tion in accuracy. To solve this problem, this paper makes the following three co ntributions without using any additional learning parameters and hyper-parameter s. First, we analyze how batch normalization, which causes the aforementioned pr oblem, disturbs the fine-tuning of the quantized DNN. Second, based on these res ults, we propose a new pruning method called Pruning for Quantization (PfQ) whic h removes the filters that disturb the fine-tuning of the DNN while not affectin g the inferred result as far as possible. Third, we propose a workflow of fine-t uning for quantized DNNs using the proposed pruning method(PfQ). Experiments usi ng well-known models and datasets confirmed that the proposed method achieves hi gher performance with a similar model size than conventional quantization method s including fine-tuning.
****

Information Theoretic Regularization for Learning Global Features by Sequential VAE
Kei Akuzawa,Yusuke Iwasawa,Yutaka Matsuo
Sequential variational autoencoders (VAEs) with global latent variable z have been studied for the purpose of disentangling the global features of data, which is useful in many downstream tasks. To assist the sequential VAEs further in ob taining meaningful z, an auxiliary loss that maximizes the mutual information (MI) between the observation and z is often employed. However, by analyzing th e sequential VAEs from the information theoretic perspective, we can claim that simply maximizing the MI encourages the latent variables to have redundant infor mation and prevents the disentanglement of global and local features. Based on t his analysis, we derive a novel regularization method that makes z informative while encouraging the disentanglement. Specifically, the proposed method remove s redundant information by minimizing the MI between z and the local features

by using adversarial training. In the experiments, we trained state-space and au
toregressive model variants using speech and image datasets. The results indicat
e that the proposed method improves the performance of the downstream classifica
tion and data generation tasks, thereby supporting our information theoretic per
spective in the learning of global representations.
****

BROS: A Pre-trained Language Model for Understanding Texts in Document
Teakgyu Hong,DongHyun Kim,Mingi Ji,Wonseok Hwang,Daehyun Nam,Sungrae Park
Understanding document from their visual snapshots is an emerging and challengin
g problem that requires both advanced computer vision and NLP methods. Although
the recent advance in OCR enables the accurate extraction of text segments, it i
s still challenging to extract key information from documents due to the diversi
ty of layouts. To compensate for the difficulties, this paper introduces a pre-t
rained language model, BERT Relying On Spatiality (BROS), that represents and un
derstands the semantics of spatially distributed texts. Different from previous
pre-training methods on 1D text, BROS is pre-trained on large-scale semi-structu
red documents with a novel area-masking strategy while efficiently including the
 spatial layout information of input documents. Also, to generate structured out
puts in various document understanding tasks, BROS utilizes a powerful graph-bas
ed decoder that can capture the relation between text segments. BROS achieves st
ate-of-the-art results on four benchmark tasks: FUNSD, SROIE*, CORD, and SciTSR.
 Our experimental settings and implementation codes will be publicly available.
****

Learning Spatiotemporal Features via Video and Text Pair Discrimination
Tianhao Li,Limin Wang
Current video representations heavily rely on learning from manually annotated v
ideo datasets which are time-consuming and expensive to acquire. We observe vide
os are naturally accompanied by abundant text information such as YouTube titles
 and Instagram captions. In this paper, we leverage this visual-textual connecti
on to learn spatiotemporal features in an efficient weakly-supervised manner. We
 present a general cross-modal pair discrimination (CPD) framework to capture th
is correlation between a video and its associated text. We train our CPD models
on both standard video dataset (Kinetics-210k) and uncurated web video dataset (
Instagram-300k) to demonstrate its effectiveness. Without further fine-tuning, t
he learnt models obtain competitive results for action classification on Kinetic
s under the linear classification protocol. Moreover, our visual model provides
an effective initialization to fine-tune on downstream tasks, which yields a rem
arkable performance gain for action recognition on UCF101 and HMDB51, compared w
ith the existing state-of-the-art self-supervised training methods. In addition,
 our CPD demonstrates that pre-training on a relatively small dataset is able to
 yield a comparable performance to those methods of using order magnitude more d
ata, which is meaningful and practicable for the scenarios with limited computat
ional facilities.
****

Frequency-aware Interface Dynamics with Generative Adversarial Networks
Lukas Prantl,Tassilo Kugelstadt,Jan Bender,Nils Thuerey
We present a new method for reconstructing and refining complex surfaces based o
n physical simulations. Taking a roughly approximated simulation as input, our
method infers corresponding spatial details while taking into account how they e
volve over time. We consider this problem in terms of spatial and temporal freq
uencies, and leverage generative adversarial networks to learn the desired spati
o-temporal signal for the surface dynamics. Furthermore, we investigate the pos
sibility to train our network in an unsupervised manner, i.e. without predefined
 training pairs. We highlight the capabilities of our method with a set of synth
etic wave function tests and complex 3D dynamics of elasto-plastic materials.
****

Better Optimization can Reduce Sample Complexity: Active Semi-Supervised Learnin
g via Convergence Rate Control
Seo Taek Kong,Soomin Jeon,Jaewon Lee,Hong-Seok Lee,Kyu-Hwan Jung
Reducing the sample complexity associated with deep learning (DL) remains one of

the most important problems in both theory and practice since its advent. Semi-supervised learning (SSL) tackles this task by leveraging unlabeled instances which are usually more accessible than their labeled counterparts. Active learning (AL) directly seeks to reduce the sample complexity by training a classification network and querying unlabeled instances to be annotated by a human-in-the-loop. Under relatively strict settings, it has been shown that both SSL and AL can theoretically achieve the same performance of fully-supervised learning (SL) using far less labeled samples. While empirical works have shown that SSL can attain this benefit in practice, DL-based AL algorithms have yet to show their success to the extent achieved by SSL. Given the accessible pool of unlabeled instances in pool-based AL, we argue that the annotation efficiency brought by AL algorithms that seek diversity on labeled samples can be improved upon when using SSL as the training scheme. Equipped with a few theoretical insights, we designed an AL algorithm that rather focuses on controlling the convergence rate of a classification network by actively querying instances to improve the rate of convergence upon inclusion to the labeled set. We name this AL scheme convergence rate control (CRC), and our experiments show that a deep neural network trained using a combination of CRC and a recently proposed SSL algorithm can quickly achieve high performance using far less labeled samples than SL. In contrast to a few works combining independently developed AL and SSL (ASSL) algorithms, our method is a natural fit to ASSL, and we hope our work can catalyze research combining AL and SSL as opposed to an exclusion of either.

****

Multi-Class Uncertainty Calibration via Mutual Information Maximization-based Binning

Kanil Patel,William H. Beluch,Bin Yang,Michael Pfeiffer,Dan Zhang

Post-hoc multi-class calibration is a common approach for providing high-quality confidence estimates of deep neural network predictions. Recent work has shown that widely used scaling methods underestimate their calibration error, while alternative Histogram Binning (HB) methods often fail to preserve classification accuracy. When classes have small prior probabilities, HB also faces the issue of severe sample-inefficiency after the conversion into K one-vs-rest class-wise calibration problems. The goal of this paper is to resolve the identified issues of HB in order to provide calibrated confidence estimates using only a small holdout calibration dataset for bin optimization while preserving multi-class ranking accuracy. From an information-theoretic perspective, we derive the I-Max concept for binning, which maximizes the mutual information between labels and quantized logits. This concept mitigates potential loss in ranking performance due to lossy quantization, and by disentangling the optimization of bin edges and representatives allows simultaneous improvement of ranking and calibration performance. To improve the sample efficiency and estimates from a small calibration set, we propose a shared class-wise (sCW) calibration strategy, sharing one calibrator among similar classes (e.g., with similar class priors) so that the training sets of their class-wise calibration problems can be merged to train the single calibrator. The combination of sCW and I-Max binning outperforms the state of the art calibration methods on various evaluation metrics across different benchmark datasets and models, using a small calibration set (e.g., 1k samples for ImageNet).

****

Error Controlled Actor-Critic Method to Reinforcement Learning

Xingen Gao,Fei Chao,Changle Zhou,Zhen Ge,Chih-Min Lin,Longzhi Yang,Xiang Chang,Changjing Shang

In the reinforcement learning (RL) algorithms which incorporate function approximation methods, the approximation error of value function inevitably cause overestimation phenomenon and have a negative impact on the convergence of the algorithms. To mitigate the negative effects of approximation error, we propose a new actor-critic algorithm called Error Controlled Actor-critic which ensures confining the approximation error in value function. In this paper, we firstly present an analysis of how the approximation error can hinder the optimization process of actor-critic methods. Then, we *derive an upper boundary of the approximation

error of Q function approximator, and found that the error can be lowered by pl
acing restrictions on the KL-divergence between every two consecutive policies d
uring the training phase of the policy.* The results of experiments on a range o
f continuous control tasks from OpenAI gym suite demonstrate that the proposed a
ctor-critic algorithm apparently reduces the approximation error and significant
ly outperforms other model-free RL algorithms.
****

Adversarial representation learning for synthetic replacement of private attribu
tes
John Martinsson,Edvin Listo Zec,Daniel Gillblad,Olof Mogren
Data privacy is an increasingly important aspect of many real-world big data ana
lytics tasks. Data sources that contain sensitive information may have immense p
otential which could be unlocked using privacy enhancing transformations, but cu
rrent methods often fail to produce convincing output. Furthermore, finding the
right balance between privacy and utility is often a tricky trade-off. In this w
ork, we propose a novel approach for data privatization, which involves two step
s: in the first step, it removes the sensitive information, and in the second st
ep, it replaces this information with an independent random sample. Our method b
uilds on adversarial representation learning which ensures strong privacy by tra
ining the model to fool an increasingly strong adversary. While previous methods
only aim at obfuscating the sensitive information, we find that adding new rand
om information in its place strengthens the provided privacy and provides better
utility at any given level of privacy. The result is an approach that can provi
de stronger privatization on image data, and yet be preserving both the domain a
nd the utility of the inputs, entirely independent of the downstream task.
****

Data Transfer Approaches to Improve Seq-to-Seq Retrosynthesis
Katsuhiko Ishiguro,Kazuya Ujihara,Ryohto Sawada,Hirotaka Akita,Masaaki Kotera
Retrosynthesis is a problem to infer reactant compounds to synthesize a given
product compound through chemical reactions. Recent studies on retrosynthesis
focus on proposing more sophisticated prediction models, but the dataset to feed
the models also plays an essential role in achieving the best generalizing model
s.
Generally, a dataset that is best suited for a specific task tends to be small.
In
such a case, it is the standard solution to transfer knowledge from a large or
clean dataset in the same domain. In this paper, we conduct a systematic and
intensive examination of data transfer approaches on end-to-end generative model
s,
in application to retrosynthesis. Experimental results show that typical data tr
ansfer
methods can improve test prediction scores of an off-the-shelf Transformer basel
ine
model. Especially, the pre-training plus fine-tuning approach boosts the accurac
y
scores of the baseline, achieving the new state-of-the-art. In addition, we cond
uct a
manual inspection for the erroneous prediction results. The inspection shows tha
t
the pre-training plus fine-tuning models can generate chemically appropriate or
sensible proposals in almost all cases.
****

Model-Agnostic Round-Optimal Federated Learning via Knowledge Transfer
Qinbin Li,Bingsheng He,Dawn Song
Federated learning enables multiple parties to collaboratively learn a model wit
hout exchanging their local data. Currently, federated averaging (FedAvg) is the
most widely used federated learning algorithm. However, FedAvg or its variants
have obvious shortcomings. It can only be used to learn differentiable models an
d needs many communication rounds to converge. In this paper, we propose a novel
federated learning algorithm FedKT that needs only a single communication round

(i.e., round-optimal). With applying the knowledge transfer approach, our algorithm can be applied to any classification model. Moreover, we develop the differentially private versions of FedKT and theoretically analyze the privacy loss. The experiments show that our method can achieve close or better accuracy compared with the other state-of-the-art federated learning algorithms.
****
A Discriminative Gaussian Mixture Model with Sparsity
Hideaki Hayashi,Seiichi Uchida
In probabilistic classification, a discriminative model based on the softmax function has a potential limitation in that it assumes unimodality for each class in the feature space. The mixture model can address this issue, although it leads to an increase in the number of parameters. We propose a sparse classifier based on a discriminative GMM, referred to as a sparse discriminative Gaussian mixture (SDGM). In the SDGM, a GMM-based discriminative model is trained via sparse Bayesian learning. Using this sparse learning framework, we can simultaneously remove redundant Gaussian components and reduce the number of parameters used in the remaining components during learning; this learning method reduces the model complexity, thereby improving the generalization capability. Furthermore, the SDGM can be embedded into neural networks (NNs), such as convolutional NNs, and can be trained in an end-to-end manner. Experimental results demonstrated that the proposed method outperformed the existing softmax-based discriminative models.
****
Fixing Asymptotic Uncertainty of Bayesian Neural Networks with Infinite ReLU Features
Agustinus Kristiadi,Matthias Hein,Philipp Hennig
Approximate Bayesian methods can mitigate overconfidence in ReLU networks. However, far away from the training data, even Bayesian neural networks (BNNs) can still underestimate uncertainty and thus be overconfident. We suggest to fix this by considering an infinite number of ReLU features over the input domain that are never part of the training process and thus remain at prior values. Perhaps surprisingly, we show that this model leads to a tractable Gaussian process (GP) term that can be added to a pre-trained BNN's posterior at test time with negligible cost overhead. The BNN then yields structured uncertainty in the proximity of training data, while the GP prior calibrates uncertainty far away from them. As a key contribution, we prove that the added uncertainty yields cubic predictive variance growth, and thus the ideal uniform (maximum entropy) confidence in multi-class classification far from the training data.
****
Trusted Multi-View Classification
Zongbo Han,Changqing Zhang,Huazhu Fu,Joey Tianyi Zhou
Multi-view classification (MVC) generally focuses on improving classification accuracy by using information from different views, typically integrating them into a unified comprehensive representation for downstream tasks. However, it is also crucial to dynamically assess the quality of a view for different samples in order to provide reliable uncertainty estimations, which indicate whether predictions can be trusted. To this end, we propose a novel multi-view classification method, termed trusted multi-view classification, which provides a new paradigm for multi-view learning by dynamically integrating different views at an evidence level. The algorithm jointly utilizes multiple views to promote both classification reliability (uncertainty estimation during testing) and robustness (out-of-distribution-awareness during training) by integrating evidence from each view. To achieve this, the Dirichlet distribution is used to model the distribution of the class probabilities, parameterized with evidence from different views and integrated with the Dempster-Shafer theory. The unified learning framework induces accurate uncertainty and accordingly endows the model with both reliability and robustness for out-of-distribution samples. Extensive experimental results validate the effectiveness of the proposed model in accuracy, reliability and robustness.
****
Learnable Uncertainty under Laplace Approximations

Agustinus Kristiadi,Matthias Hein,Philipp Hennig
Laplace approximations are classic, computationally lightweight means to construct Bayesian neural networks (BNNs). As in other approximate BNNs, one cannot necessarily expect the induced predictive uncertainty to be calibrated. Here we develop a formalism to explicitly "train" the uncertainty in a decoupled way to the prediction itself. To this end we introduce uncertainty units for Laplace-approximated networks: Hidden units with zero weights that can be added to any pre-trained, point-estimated network. Since these units are inactive, they do not affect the predictions. But their presence changes the geometry (in particular the Hessian) of the loss landscape around the point estimate, thereby affecting the network's uncertainty estimates under a Laplace approximation. We show that such units can be trained via an uncertainty-aware objective, making the Laplace approximation competitive with more expensive alternative uncertainty-quantification frameworks.
****
Towards Powerful Graph Neural Networks: Diversity Matters
Xu Bingbing,Huawei Shen,Qi Cao,Yuanhao Liu,Keting Cen,Xueqi Cheng
Graph neural networks (GNNs) offer us an effective framework for graph representation learning via layer-wise neighborhood aggregation. Their success is attributed to their expressive power at learning representation of nodes and graphs. To achieve GNNs with high expressive power, existing methods mainly resort to complex neighborhood aggregation functions, e.g., designing injective aggregation function or using multiple aggregation functions. Consequently, their expressive power is limited by the capability of aggregation function, which is tricky to determine in practice. To combat this problem, we propose a novel framework, namely diverse sampling, to improve the expressive power of GNNs. For a target node, diverse sampling offers it diverse neighborhoods, i.e., rooted sub-graphs, and the representation of target node is finally obtained via aggregating the representation of diverse neighborhoods obtained using any GNN model. High expressive power is guaranteed by the diversity of different neighborhoods. We use classical GNNs (i.e., GCN and GAT) as base models to evaluate the effectiveness of the proposed framework. Experiments are conducted at multi-class node classification task on three benchmark datasets and multi-label node classification task on a dataset collected in this paper. Extensive experiments demonstrate the proposed method consistently improve the performance of base GNN models. The proposed framework is applicable to any GNN models and thus is general for improving the expressive power of GNNs.
****
Stochastic Subset Selection for Efficient Training and Inference of Neural Networks
Bruno Andreis,Tuan Nguyen,Juho Lee,Eunho Yang,Sung Ju Hwang
Current machine learning algorithms are designed to work with huge volumes of high dimensional data such as images. However, these algorithms are being increasingly deployed to resource constrained systems such as mobile devices and embedded systems. Even in cases where large computing infrastructure is available, the size of each data instance, as well as datasets, can provide a huge bottleneck in data transfer across communication channels. Also, there is a huge incentive both in energy and monetary terms in reducing both the computational and memory requirements of these algorithms. For non-parametric models that require to leverage the stored training data at the inference time, the increased cost in memory and computation could be even more problematic. In this work, we aim to reduce the volume of data these algorithms must process through an end-to-end two-stage neural subset selection model, where the first stage selects a set of candidate points using a conditionally independent Bernoulli mask followed by an iterative coreset selection via a conditional Categorical distribution. The subset selection model is trained by meta-learning with a distribution of sets. We validate our method on set reconstruction and classification tasks with feature selection as well as the selection of representative samples from a given dataset, on which our method outperforms relevant baselines. We also show in our experiments that our method enhances scalability of non-parametric models such as Neural Proce

sses.
****
IEPT: Instance-Level and Episode-Level Pretext Tasks for Few-Shot Learning
Manli Zhang,Jianhong Zhang,Zhiwu Lu,Tao Xiang,Mingyu Ding,Songfang Huang
The need of collecting large quantities of labeled training data for each new ta
sk has limited the usefulness of deep neural networks. Given data from a set of
source tasks, this limitation can be overcome using two transfer learning approa
ches: few-shot learning (FSL) and self-supervised learning (SSL). The former aim
s to learn `how to learn' by designing learning episodes using source tasks to s
imulate the challenge of solving the target new task with few labeled samples. I
n contrast, the latter exploits an annotation-free pretext task across all sourc
e tasks in order to learn generalizable feature representations. In this work, w
e propose a novel Instance-level and Episode-level Pretext Task (IEPT) framework
 that seamlessly integrates SSL into FSL. Specifically, given an FSL episode, we
 first apply geometric transformations to each instance to generate extended epi
sodes. At the instance-level, transformation recognition is performed as per sta
ndard SSL. Importantly, at the episode-level, two SSL-FSL hybrid learning object
ives are devised: (1) The consistency across the predictions of an FSL classifie
r from different extended episodes is maximized as an episode-level pretext task
. (2) The features extracted from each instance across different episodes are in
tegrated to construct a single FSL classifier for meta-learning. Extensive exper
iments show that our proposed model (i.e., FSL with IEPT) achieves the new state
-of-the-art.
****
Intention Propagation for Multi-agent Reinforcement Learning
Chao Qu,Hui Li,Chang Liu,Junwu Xiong,james zhang,Wei Chu,Weiqiang Wang,Yuan Qi,L
e Song
A hallmark of an AI agent is to mimic human beings to understand and interact wi
th others. In this paper, we propose a \emph{collaborative} multi-agent reinforc
ement learning algorithm to learn a \emph{joint} policy through the interactions
 over agents. To make a joint decision over the group, each agent makes an init
ial decision and tells its policy to its neighbors. Then each agent modifies it
s own policy properly based on received messages and spreads out its plan. As th
is intention propagation procedure goes on, we prove that it converges to a mean
-field approximation of the joint policy with the framework of neural embedded p
robabilistic inference. We evaluate our algorithm on several large scale challen
ging tasks and demonstrate that it outperforms previous state-of-the-arts.
****
Membership Attacks on Conditional Generative Models Using Image Difficulty
Avital Shafran,Shmuel Peleg,Yedid Hoshen
Membership inference attacks (MIA) try to detect if data samples were used to tr
ain a Neural Network model. As training data is very valuable in machine learnin
g, MIA can be used to detect the use of unauthorized data. Unlike the traditiona
l MIA approaches, addressing classification models, we address conditional image
 generation models (e.g. image translation).
Due to overfitting, reconstruction errors are typically lower for images used in
 training. A simple but effective approach for membership attacks can therefore
use the reconstruction error.
However, we observe that some images are "universally" easy, and others are diff
icult. Reconstruction error alone is less effective at discriminating between di
fficult images used in training and easy images that were never seen before. To
overcome this, we propose to use a novel difficulty score that can be computed f
or each image, and its computation does not require a training set. Our membersh
ip error, obtained by subtracting the difficulty score from the reconstruction e
rror, is shown to achieve high MIA accuracy on an extensive number of benchmarks
.
****
Learning Disentangled Representations for Image Translation
Aviv Gabbay,Yedid Hoshen
Recent approaches for unsupervised image translation are strongly reliant on gen

erative adversarial training and architectural locality constraints. Despite the
ir appealing results, it can be easily observed that the learned class and conte
nt representations are entangled which often hurts the translation performance.
To this end, we propose OverLORD, for learning disentangled representations for
the image class and attributes, utilizing latent optimization and carefully desi
gned content and style bottlenecks. We further argue that the commonly used adve
rsarial optimization can be decoupled from representation disentanglement and be
 applied at a later stage of the training to increase the perceptual quality of
the generated images. Based on these principles, our model learns significantly
more disentangled representations and achieves higher translation quality and gr
eater output diversity than state-of-the-art methods.
****

What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study
Marcin Andrychowicz,Anton Raichuk,Piotr Sta■czyk,Manu Orsini,Sertan Girgin,Rapha
ël Marinier,Leonard Hussenot,Matthieu Geist,Olivier Pietquin,Marcin Michalski,Sy
lvain Gelly,Olivier Bachem
In recent years, reinforcement learning (RL) has been successfully applied to ma
ny different continuous control tasks. While RL algorithms are often conceptuall
y simple, their state-of-the-art implementations take numerous low- and high-lev
el design decisions that strongly affect the performance of the resulting agents
. Those choices are usually not extensively discussed in the literature, leading
 to discrepancy between published descriptions of algorithms and their implement
ations. This makes it hard to attribute progress in RL and slows down overall pr
ogress [Engstrom'20]. As a step towards filling that gap, we implement >50 such
``"choices" in a unified on-policy deep actor-critic framework, allowing us to i
nvestigate their impact in a large-scale empirical study. We train over 250'000
agents in five continuous control environments of different complexity and provi
de insights and practical recommendations for the training of on-policy deep act
or-critic RL agents.
****

Deep Single Image Manipulation
Yael Vinker,Eliahu Horwitz,Nir Zabari,Yedid Hoshen
Image manipulation has attracted much research over the years due to the popular
ity and commercial importance of the task. In recent years, deep neural network
methods have been proposed for many image manipulation tasks. A major issue with
 deep methods is the need to train on large amounts of data from the same distri
bution as the target image, whereas collecting datasets encompassing the entire
long-tail of images is impossible. In this paper, we demonstrate that simply tra
ining a conditional adversarial generator on the single target image is sufficie
nt for performing complex image manipulations. We find that the key for enabling
 single image training is extensive augmentation of the input image and provide
a novel augmentation method. Our network learns to map between a primitive repre
sentation of the image (e.g. edges and segmentation) to the image itself. At man
ipulation time, our generator allows for making general image changes by modifyi
ng the primitive input representation and mapping it through the network. We ext
ensively evaluate our method and find that it provides remarkable performance.
****

Deep Encoder, Shallow Decoder: Reevaluating Non-autoregressive Machine Translati
on
Jungo Kasai,Nikolaos Pappas,Hao Peng,James Cross,Noah Smith
Much recent effort has been invested in non-autoregressive neural machine transl
ation, which appears to be an efficient alternative to state-of-the-art autoregr
essive machine translation on modern GPUs. In contrast to the latter, where gen
eration is sequential, the former allows generation to be parallelized across ta
rget token positions. Some of the latest non-autoregressive models have achieved
 impressive translation quality-speed tradeoffs compared to autoregressive basel
ines. In this work, we reexamine this tradeoff and argue that autoregressive bas
elines can be substantially sped up without loss in accuracy. Specifically, we s
tudy autoregressive models with encoders and decoders of varied depths. Our exte
nsive experiments show that given a sufficiently deep encoder, a single-layer au

toregressive decoder can substantially outperform strong non-autoregressive mode ls with comparable inference speed. We show that the speed disadvantage for auto regressive baselines compared to non-autoregressive methods has been overestimat ed in three aspects: suboptimal layer allocation, insufficient speed measurement , and lack of knowledge distillation. Our results establish a new protocol for f uture research toward fast, accurate machine translation. Our code is available at https://github.com/jungokasai/deep-shallow.
****
Effective Abstract Reasoning with Dual-Contrast Network
Tao Zhuo,Mohan Kankanhalli
As a step towards improving the abstract reasoning capability of machines, we ai m to solve Raven’s Progressive Matrices (RPM) with neural networks, since solvin g RPM puzzles is highly correlated with human intelligence. Unlike previous meth ods that use auxiliary annotations or assume hidden rules to produce appropriate feature representation, we only use the ground truth answer of each question fo r model learning, aiming for an intelligent agent to have a strong learning cap ability with a small amount of supervision. Based on the RPM problem formulatio n, the correct answer filled into the missing entry of the third row/column has to best satisfy the same rules shared between the first two rows/co lumns.Thus we design a simple yet effective Dual-Contrast Network (DCNe t) to exploit the inherent structure of RPM puzzles. Specifically, a rule con trast module is designed to compare the latent rules between the filled row/column and the first two rows/columns; a choice contrast module is desig ned to increase the relative differences between candidate choices. Experimenta l results on the RAVEN and PGM datasets show that DCNet outperforms the state-of-the-art methods by a large margin of 5.77%. Further experiments on f ew training samples and model generalization also show the effectiveness of DCNe t. Code is available at https://github.com/visiontao/dcnet.
****
Diverse Exploration via InfoMax Options
Yuji Kanagawa,Tomoyuki Kaneko
In this paper, we study the problem of autonomously discovering temporally abstr acted actions, or options, for exploration in reinforcement learning. For learni ng diverse options suitable for exploration, we introduce the infomax terminatio n objective defined as the mutual information between options and their correspo nding state transitions. We derive a scalable optimization scheme for maximizing this objective via the termination condition of options, yielding the InfoMax O ption Critic (IMOC) algorithm. Through illustrative experiments, we empirically show that IMOC learns diverse options and utilizes them for exploration. Moreove r, we show that IMOC scales well to continuous control tasks.

****
AMBERT: A Pre-trained Language Model with Multi-Grained Tokenization
Xinsong Zhang,Hang Li
Pre-trained language models such as BERT have exhibited remarkable performances in many tasks in natural language understanding (NLU). The tokens in the models are usually fine-grained in the sense that for languages like English they are w ords or sub-words and for languages like Chinese they are characters. In English , for example, there are multi-word expressions which form natural lexical units and thus the use of coarse-grained tokenization also appears to be reasonable. In fact, both fine-grained and coarse-grained tokenizations have advantages and disadvantages for learning of pre-trained language models. In this paper, we pro pose a novel pre-trained language model, referred to as AMBERT (A Multi-grained BERT), on the basis of both fine-grained and coarse-grained tokenizations. For E nglish, AMBERT takes both the sequence of words (fine-grained tokens) and the se quence of phrases (coarse-grained tokens) as input after tokenization, employs o ne encoder for processing the sequence of words and the other encoder for proces sing the sequence of the phrases, utilizes shared parameters between the two enc oders, and finally creates a sequence of contextualized representations of the w ords and a sequence of contextualized representations of the phrases. Experiment

s have been conducted on benchmark datasets for Chinese and English, including C LUE, GLUE, SQuAD and RACE. The results show that AMBERT outperforms the existing best performing models in almost all cases, particularly the improvements are s ignificant for Chinese. We also develop a version of AMBERT which performs equal ly well as AMBERT but uses about half of its inference time.
****

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 accura cy for probabilistic demand prediction. In this work, we propose novel improveme nts to the current state of the art by incorporating changes inspired by recent advances in Transformer architectures for Natural Language Processing. We develo p 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 encodin g that allows the neural network to learn context-dependent seasonality function s 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 b y failing to leverage previous errors in the forecast to improve accuracy. We pr opose a novel decoder-self attention scheme for forecasting that produces signif icant improvements in the excess variation of the forecast.
****

Noise against noise: stochastic label noise helps combat inherent label noise
Pengfei Chen,Guangyong Chen,Junjie Ye,jingwei zhao,Pheng-Ann Heng
The noise in stochastic gradient descent (SGD) provides a crucial implicit regul arization effect, previously studied in optimization by analyzing the dynamics o f parameter updates. In this paper, we are interested in learning with noisy lab els, where we have a collection of samples with potential mislabeling. We show t hat a previously rarely discussed SGD noise, induced by stochastic label noise (SLN), mitigates the effects of inherent label noise. In contrast, the common SGD noise directly applied to model parameters does not. We formalize the differenc es and connections of SGD noise variants, showing that SLN induces SGD noise dep endent on the sharpness of output landscape and the confidence of output probabi lity, which may help escape from sharp minima and prevent overconfidence. SLN no t only improves generalization in its simplest form but also boosts popular robu st training methods, including sample selection and label correction. Specifical ly, we present an enhanced algorithm by applying SLN to label correction. Our co de is released.
****

On the Landscape of Sparse Linear Networks
Dachao Lin,Ruoyu Sun,Zhihua Zhang
Network pruning, or sparse network has a long history and practical significance in modern applications. Although the loss functions of neural networks may yiel d bad landscape due to non-convexity, we focus on linear activation which alread y owes benign landscape. With no unrealistic assumption, we conclude the followi ng statements for the squared loss objective of general sparse linear neural net works: 1) every local minimum is a global minimum for scalar output with any spa rse structure, or non-intersected sparse first layer and dense other layers with orthogonal training data; 2) sparse linear networks have sub-optimal local-min for only sparse first layer due to low rank constraint, or output larger than th ree dimensions due to the global minimum of a sub-network. Overall, sparsity bre aks the normal structure, cutting out the decreasing path in original fully-conn ected networks.
****

CIGMO: Learning categorical invariant deep generative models from grouped data
Haruo Hosoya
Images of general objects are often composed of three hidden factors: category (e.g., car or chair), shape (e.g., particular car form), and view (e.g., 3D orien tation). While there have been many disentangling models that can discover eit

her a category or shape factor separately from a view factor, such models typica lly cannot capture the structure of general objects that the diversity of shapes is much larger across categories than within a category. Here, we propose a n ovel generative model called CIGMO, which can learn to represent the category, s hape, and view factors at once only with weak supervision. Concretely, we devel op mixture of disentangling deep generative models, where the mixture components correspond to object categories and each component model represents shape and v iew in a category-specific and mutually invariant manner. We devise a learning method based on variational autoencoders that does not explicitly use label info rmation but uses only grouping information that links together different views o f the same object. Using several datasets of 3D objects including ShapeNet, we demonstrate that our model often outperforms previous relevant models including state-of-the-art methods in invariant clustering and one-shot classification tas ks, in a manner exposing the importance of categorical invariant representation.

****

VAEBM: A Symbiosis between Variational Autoencoders and Energy-based Models
Zhisheng Xiao,Karsten Kreis,Jan Kautz,Arash Vahdat
Energy-based models (EBMs) have recently been successful in representing complex distributions of small images. However, sampling from them requires expensive M arkov chain Monte Carlo (MCMC) iterations that mix slowly in high dimensional pi xel space. Unlike EBMs, variational autoencoders (VAEs) generate samples quickly and are equipped with a latent space that enables fast traversal of the data ma nifold. However, VAEs tend to assign high probability density to regions in data space outside the actual data distribution and often fail at generating sharp i mages. In this paper, we propose VAEBM, a symbiotic composition of a VAE and an EBM that offers the best of both worlds. VAEBM captures the overall mode structu re of the data distribution using a state-of-the-art VAE and it relies on its EB M component to explicitly exclude non-data-like regions from the model and refin e the image samples. Moreover, the VAE component in VAEBM allows us to speed up MCMC updates by reparameterizing them in the VAE's latent space. Our experimenta l results show that VAEBM outperforms state-of-the-art VAEs and EBMs in generati ve quality on several benchmark image datasets by a large margin. It can generat e high-quality images as large as 256$\times$256 pixels with short MCMC chains. We also demonstrate that VAEBM provides complete mode coverage and performs well in out-of-distribution detection.

****

Debiased Graph Neural Networks with Agnostic Label Selection Bias
Shaohua Fan,Xiao Wang,Chuan Shi,Kun Kuang,Nian Liu,Bai Wang
Most existing Graph Neural Networks (GNNs) are proposed without considering the selection bias in data, i.e., the inconsistent distribution between the training set with test set. In reality, the test data is not even available during the t raining process, making selection bias agnostic. Training GNNs with biased selec ted nodes leads to significant parameter estimation bias and greatly impacts the generalization ability on test nodes. In this paper, we first present an experi mental investigation, which clearly shows that the selection bias drastically hi nders the generalization ability of GNNs, and theoretically prove that the selec tion bias will cause the biased estimation on GNN parameters. Then to remove the bias in GNN estimation, we propose a novel Debiased Graph Neural Networks (DGNN) with a differentiated decorrelation regularizer. The differentiated decorrelat ion regularizer estimates a sample weight for each labeled node such that the sp urious correlation of learned embeddings could be eliminated. We analyze the reg ularizer in causal view and it motivates us to differentiate the weights of the variables based on their contribution on the confounding bias. Then, these sampl e weights are used for reweighting GNNs to eliminate the estimation bias, thus h elp to improve the stability of prediction on unknown test nodes. Comprehensive experiments are conducted on several challenging graph datasets with two kinds o f label selection bias. The results well verify that our proposed model outperfo rms the state-of-the-art methods and DGNN is a flexible framework to enhance exi sting GNNs.

****
Rethinking the Truly Unsupervised Image-to-Image Translation
Kyungjune Baek,Yunjey Choi,Youngjung Uh,Jaejun Yoo,Hyunjung Shim
Every recent image-to-image translation model uses either image-level (i.e. inpu t-output pairs) or set-level (i.e. domain labels) supervision at a minimum. Howe ver, even the set-level supervision can be a serious bottleneck for data collect ion in practice. In this paper, we tackle image-to-image translation in a fully unsupervised setting, i.e., neither paired images nor domain labels. To this end , we propose a truly unsupervised image-to-image translation model (TUNIT) that simultaneously learns to separate image domains and translate input images into the estimated domains.
Experimental results show that our model achieves comparable or even better perf ormance than the set-level supervised model trained with full labels, generalize s well on various datasets, and is robust against the choice of hyperparameters (e.g. the preset number of pseudo domains). In addition, TUNIT extends well to t he semi-supervised scenario with various amount of labels provided.
****
Sufficient and Disentangled Representation Learning
Jian Huang,Yuling Jiao,Xu Liao,Jin Liu,Zhou Yu
We propose a novel approach to representation learning called sufficient and dis entangled representation learning (SDRL). With SDRL, we seek a data representati on that maps the input data to a lower-dimensional space with two properties: su fficiency and disentanglement. First, the representation is sufficient in the se nse that the original input data is conditionally independent of the response or label given the representation. Second, the representation is maximally disenta ngled with mutually independent components and rotation invariant in distributio n. We show that such a representation always exists under mild conditions on the input data distribution based on optimal transport theory. We formulate an obje ctive function characterizing conditional independence and disentanglement. This objective function is then used to train a sufficient and disentangled represen tation with deep neural networks. We provide strong statistical guarantees for t he learned representation by establishing an upper bound on the excess error of the objective function and show that it reaches the nonparametric minimax rate u nder mild conditions. We also validate the proposed method via numerical experim ents and real data analysis.
****
On Position Embeddings in BERT
Benyou Wang,Lifeng Shang,Christina Lioma,Xin Jiang,Hao Yang,Qun Liu,Jakob Grue S imonsen
Various Position Embeddings (PEs) have been proposed in Transformer based archit ectures~(e.g. BERT) to model word order. These are empirically-driven and perfor m well, but no formal framework exists to systematically study them. To address this, we present three properties of PEs that capture word distance in vector sp ace: translation invariance, monotonicity, and symmetry. These properties form ally capture the behaviour of PEs and allow us to reinterpret sinusoidal PEs in a principled way.
Moreover, we propose a new probing test (called `identical word probing') and m athematical indicators to quantitatively detect the general attention patterns with respect to the above properties. An empirical evaluation of seven PEs (and their combinations) for classification (GLUE) and span prediction (SQuAD) shows that: (1) both classification and span prediction benefit from translation in variance and local monotonicity, while symmetry slightly decreases performance; (2) The fully-learnable absolute PE performs better in classification, while rel ative PEs perform better in span prediction. We contribute the first formal and quantitative analysis of desiderata for PEs, and a principled discussion about their correlation to the performance of typical downstream tasks.
****
On The Adversarial Robustness of 3D Point Cloud Classification
Jiachen Sun,Karl Koenig,Yulong Cao,Qi Alfred Chen,Zhuoqing Mao
3D point clouds play pivotal roles in various safety-critical fields, such as au

tonomous driving, which desires the corresponding deep neural networks to be rob ust to adversarial perturbations. Though a few defenses against adversarial poin t cloud classification have been proposed, it remains unknown whether they can p rovide real robustness. To this end, we perform the first security analysis of s tate-of-the-art defenses and design adaptive attacks on them. Our 100% adaptive attack success rates demonstrate that current defense designs are still vulnerab le. Since adversarial training (AT) is believed to be the most effective defense , we present the first in-depth study showing how AT behaves in point cloud clas sification and identify that the required symmetric function (pooling operation) is paramount to the model's robustness under AT. Through our systematic analysi s, we find that the default used fixed pooling operations (e.g., MAX pooling) ge nerally weaken AT's performance in point cloud classification. Still, sorting-ba sed parametric pooling operations can significantly improve the models' robustne ss. Based on the above insights, we further propose DeepSym, a deep symmetric po oling operation, to architecturally advance the adversarial robustness under AT to 47.01% without sacrificing nominal accuracy, outperforming the original desig n and a strong baseline by 28.5% ($\sim 2.6 \times$) and 6.5%, respectively, in PointNet.

****

Neural Pruning via Growing Regularization

Huan Wang,Can Qin,Yulun Zhang,Yun Fu

Regularization has long been utilized to learn sparsity in deep neural network p runing. However, its role is mainly explored in the small penalty strength regim e. In this work, we extend its application to a new scenario where the regulariz ation grows large gradually to tackle two central problems of pruning: pruning s chedule and weight importance scoring. (1) The former topic is newly brought up in this work, which we find critical to the pruning performance while receives l ittle research attention. Specifically, we propose an L2 regularization variant with rising penalty factors and show it can bring significant accuracy gains com pared with its one-shot counterpart, even when the same weights are removed. (2) The growing penalty scheme also brings us an approach to exploit the Hessian in formation for more accurate pruning without knowing their specific values, thus not bothered by the common Hessian approximation problems. Empirically, the prop osed algorithms are easy to implement and scalable to large datasets and network s in both structured and unstructured pruning. Their effectiveness is demonstrat ed with modern deep neural networks on the CIFAR and ImageNet datasets, achievin g competitive results compared to many state-of-the-art algorithms. Our code and trained models are publicly available at https://github.com/mingsun-tse/regular ization-pruning.

****

Iterative Image Inpainting with Structural Similarity Mask for Anomaly Detection

Hitoshi Nakanishi,Masahiro Suzuki,Yutaka Matsuo

Autoencoders have emerged as popular methods for unsupervised anomaly detection. Autoencoders trained on the normal data are expected to reconstruct only the no rmal features, allowing anomaly detection by thresholding reconstruction errors. However, in practice, autoencoders fail to model small detail and yield blurry reconstructions, which makes anomaly detection challenging. Moreover, there is o bjective mismatching that models are trained to minimize total reconstruction er rors while expecting a small deviation on normal pixels and a large deviation on anomalous pixels. To tackle these two issues, we propose the iterative image in painting method that reconstructs partial regions in an adaptive inpainting mask matrix. This method constructs inpainting masks from the anomaly score of struc tural similarity. Overlaying inpainting mask on images, each pixel is bypassed o r reconstructed based on the anomaly score, enhancing reconstruction quality. Th e iterative update of inpainted images and masks by turns purifies the anomaly s core directly and follows the expected objective at test time. We evaluated the proposed method using the MVTec Anomaly Detection dataset. Our method outperform ed previous state-of-the-art in several categories and showed remarkable improve ment in high-frequency textures.

****

Bayesian neural network parameters provide insights into the earthquake rupture physics.
Sabber Ahamed
I present a simple but informative approach to gain insight into the Bayesian neural network (BNN) trained parameters. I used 2000 dynamic rupture simulations to train a BNN model to predict if an earthquake can break through a simple 2D fault. In each simulation, fault geometry, stress conditions, and friction parameters vary. The trained BNN parameters show that the network learns the physics of earthquake rupture. Neurons with high positive weights contribute to the earthquake rupture and vice versa. The results show that the stress condition of the fault plays a critical role in determining its strength. The stress is also the top source of uncertainty, followed by the dynamic friction coefficient. When stress and friction drop of a fault have higher value and are combined with higher weighted neurons, the prediction score increases, thus fault likely to be ruptured. Fault's width and height have the least amount of uncertainty, which may not be correct in a real scenario. The study shows that the potentiality of BNN that provides data patterns about rupture physics to make an additional information source for scientists studying the earthquake rupture.
****

Mixed-Features Vectors and Subspace Splitting
Alejandro Pimentel-Alarcón,Daniel L. Pimentel-Alarcón
Motivated by metagenomics, recommender systems, dictionary learning, and related problems, this paper introduces subspace splitting(SS): the task of clustering the entries of what we call amixed-features vector, that is, a vector whose subsets of coordinates agree with a collection of subspaces. We derive precise identifiability conditions under which SS is well-posed, thus providing the first fundamental theory for this problem. We also propose the first three practical SS algorithms, each with advantages and disadvantages: a random sampling method , a projection-based greedy heuristic , and an alternating Lloyd-type algorithm ; all allow noise, outliers, and missing data. Our extensive experiments outline the performance of our algorithms, and in lack of other SS algorithms, for reference we compare against methods for tightly related problems, like robust matched subspace detection and maximum feasible subsystem, which are special simpler cases of SS.
****

Hierarchical Reinforcement Learning by Discovering Intrinsic Options
Jesse Zhang,Haonan Yu,Wei Xu
We propose a hierarchical reinforcement learning method, HIDIO, that can learn task-agnostic options in a self-supervised manner while jointly learning to utilize them to solve sparse-reward tasks. Unlike current hierarchical RL approaches that tend to formulate goal-reaching low-level tasks or pre-define ad hoc lower-level policies, HIDIO encourages lower-level option learning that is independent of the task at hand, requiring few assumptions or little knowledge about the task structure. These options are learned through an intrinsic entropy minimization objective conditioned on the option sub-trajectories. The learned options are diverse and task-agnostic. In experiments on sparse-reward robotic manipulation and navigation tasks, HIDIO achieves higher success rates with greater sample efficiency than regular RL baselines and two state-of-the-art hierarchical RL methods. Code at: https://github.com/jesbu1/hidio.
****

Zero-Shot Learning with Common Sense Knowledge Graphs
Nihal Nayak,Stephen Bach
Zero-shot learning relies on semantic class representations such as hand-engineered attributes or learned embeddings to predict classes without any labeled examples. We propose to learn class representations from common sense knowledge graphs. Common sense knowledge graphs are an untapped source of explicit high-level knowledge that requires little human effort to apply to a range of tasks. To capture the knowledge in the graph, we introduce ZSL-KG, a general-purpose framework with a novel transformer graph convolutional network (TrGCN) to generate class representations. Our proposed TrGCN architecture computes non-linear combinatio

ns of the node neighbourhood and leads to significant improvements on zero-shot learning tasks. We report new state-of-the-art accuracies on six zero-shot benchmark datasets in object classification, intent classification, and fine-grained entity typing tasks. ZSL-KG outperforms the specialized state-of-the-art method for each task by an average 1.7 accuracy points and outperforms the general-purpose method with the best average accuracy by 5.3 points. Our ablation study on ZSL-KG with alternate graph neural networks shows that our transformer-based aggregator adds up to 2.8 accuracy points improvement on these tasks.
****
Sharper Generalization Bounds for Learning with Gradient-dominated Objective Functions
Yunwen Lei,Yiming Ying
Stochastic optimization has become the workhorse behind many successful machine learning applications, which motivates a lot of theoretical analysis to understand its empirical behavior. As a comparison, there is far less work to study the generalization behavior especially in a non-convex learning setting. In this paper, we study the generalization behavior of stochastic optimization by leveraging the algorithmic stability for learning with β-gradient-dominated objective functions. We develop generalization bounds of the order $O(1/(n\beta))$ plus the convergence rate of the optimization algorithm, where n is the sample size. Our stability analysis significantly improves the existing non-convex analysis by removing the bounded gradient assumption and implying better generalization bounds. We achieve this improvement by exploiting the smoothness of loss functions instead of the Lipschitz condition in Charles & Papailiopoulos (2018). We apply our general results to various stochastic optimization algorithms, which show clearly how the variance-reduction techniques improve not only training but also generalization. Furthermore, our discussion explains how interpolation helps generalization for highly expressive models.
****
Redefining The Self-Normalization Property
Zhaodong Chen,Zhao WeiQin,Lei Deng,Guoqi Li,Yuan Xie
The approaches that prevent gradient explosion and vanishing have boosted the performance of deep neural networks in recent years. A unique one among them is the self-normalizing neural network (SNN), which is generally more stable than initialization techniques without explicit normalization. The self-normalization property of SNN in previous studies comes from the Scaled Exponential Linear Unit (SELU) activation function. %, which has achieved competitive accuracy on moderate-scale benchmarks. However, it has been shown that in deeper neural networks, SELU either leads to gradient explosion or loses its self-normalization property. Besides, its accuracy on large-scale benchmarks like ImageNet is less satisfying. In this paper, we analyze the forward and backward passes of SNN with mean-field theory and block dynamical isometry. A new definition for self-normalization property is proposed that is easier to use both analytically and numerically. A proposition is also proposed which enables us compare the strength of the self-normalization property between different activation functions. We further develop two new activation functions, leaky SELU (lSELU) and scaled SELU (sSELU), that have stronger self-normalization property. The optimal parameters in them can be easily solved with a constrained optimization program. Besides, analysis on the activation's mean in the forward pass reveals that the self-normalization property on mean gets weaker with larger fan-in, which explains the performance degradation on ImageNet. This can be solved with weight centralization, mixup data augmentation, and centralized activation function. On moderate-scale datasets CIFAR-10, CIFAR-100, and Tiny ImageNet, the direct application of lSELU and sSELU achieves up to 2.13% higher accuracy. On Conv MobileNet V1 - ImageNet, sSELU with Mixup, trainable λ, and centralized activation function reaches 71.95% accuracy that is even better than Batch Normalization.(code in Supplementary Material)
****
Representation Learning for Sequence Data with Deep Autoencoding Predictive Components

Junwen Bai,Weiran Wang,Yingbo Zhou,Caiming Xiong
We propose Deep Autoencoding Predictive Components (DAPC) -- a self-supervised r epresentation learning method for sequence data, based on the intuition that use ful representations of sequence data should exhibit a simple structure in the la tent space. We encourage this latent structure by maximizing an estimate of \emp h{predictive information} of latent feature sequences, which is the mutual infor mation between the past and future windows at each time step. In contrast to the mutual information lower bound commonly used by contrastive learning, the estim ate of predictive information we adopt is exact under a Gaussian assumption. Add itionally, it can be computed without negative sampling. To reduce the degenerac y of the latent space extracted by powerful encoders and keep useful information from the inputs, we regularize predictive information learning with a challengi ng masked reconstruction loss. We demonstrate that our method recovers the laten t space of noisy dynamical systems, extracts predictive features for forecasting tasks, and improves automatic speech recognition when used to pretrain the enco der on large amounts of unlabeled data.
****

Descending through a Crowded Valley — Benchmarking Deep Learning Optimizers
Robin Marc Schmidt,Frank Schneider,Philipp Hennig
Choosing the optimizer is considered to be among the most crucial design decisio ns in deep learning, and it is not an easy one. The growing literature now lists hundreds of optimization methods. In the absence of clear theoretical guidance and conclusive empirical evidence, the decision is often made based on anecdotes . In this work, we aim to replace these anecdotes, if not with a conclusive rank ing, then at least with evidence-backed heuristics. To do so, we perform an exte nsive, standardized benchmark of more than a dozen particularly popular deep lea rning optimizers while giving a concise overview of the wide range of possible c hoices. Analyzing almost $35,000$ individual runs, we contribute the following t hree points: (i) Optimizer performance varies greatly across tasks. (ii) We obse rve that evaluating multiple optimizers with default parameters works approximat ely as well as tuning the hyperparameters of a single, fixed optimizer. (iii) Wh ile we can not discern an optimization method clearly dominating across all test ed tasks, we identify a significantly reduced subset of specific algorithms and parameter choices that generally lead to competitive results in our experiments. This subset includes popular favorites and some lesser-known contenders. We hav e open-sourced all our experimental results, making them directly available as c hallenging and well-tuned baselines. This allows for more meaningful comparisons when evaluating novel optimization methods without requiring any further comput ational efforts.
****

Robust Reinforcement Learning using Adversarial Populations
Eugene Vinitsky,Yuqing du,Kanaad V Parvate,Kathy Jang,Pieter Abbeel,Alexandre Ba yen
Reinforcement Learning (RL) is an effective tool for controller design but can s truggle with issues of robustness, failing catastrophically when the underlying system dynamics are perturbed. The Robust RL formulation tackles this by adding worst-case adversarial noise to the dynamics and constructing the noise distribu tion as the solution to a zero-sum minimax game. However, existing work on learn ing solutions to the Robust RL formulation has primarily focused on training a s ingle RL agent against a single adversary. In this work, we demonstrate that usi ng a single adversary does not consistently yield robustness to dynamics variati ons under standard parametrizations of the adversary; the resulting policy is hi ghly exploitable by new adversaries. We propose a population-based augmentation to the Robust RL formulation in which we randomly initialize a population of adv ersaries and sample from the population uniformly during training. We empiricall y validate across a variety of benchmarks that the use of an adversarial populat ion results in a less exploitable, more robust policy. Finally, we demonstrate that this approach provides comparable robustness and generalization as domain r andomization on these benchmarks while avoiding a ubiquitous domain randomizatio n failure mode.

****
Learning Two-Time-Scale Representations For Large Scale Recommendations
Xinshi Chen,Yan Zhu,Haowen Xu,Muhan Zhang,Liang Xiong,Le Song
We propose a surprisingly simple but effective two-time-scale (2TS) model for le arning user representations for recommendation. In our approach, we will partiti on users into two sets, active users with many observed interactions and inactiv e or new users with few observed interactions, and we will use two RNNs to model them separately. Furthermore, we design a two-stage training method for our mod el, where, in the first stage, we learn transductive embeddings for users and it ems, and then, in the second stage, we learn the two RNNs leveraging the transdu ctive embeddings trained in the first stage. Through the lens of online learning and stochastic optimization, we provide theoretical analysis that motivates the design of our 2TS model. The 2TS model achieves a nice bias-variance trade-off while being computationally efficient. In large scale datasets, our 2TS model is able to achieve significantly better recommendations than previous state-of-the -art, yet being much more computationally efficient.
****
Average-case Acceleration for Bilinear Games and Normal Matrices
Carles Domingo-Enrich,Fabian Pedregosa,Damien Scieur
Advances in generative modeling and adversarial learning have given rise to rene wed interest in smooth games. However, the absence of symmetry in the matrix of second derivatives poses challenges that are not present in the classical minimi zation framework. While a rich theory of average-case analysis has been develope d for minimization problems, little is known in the context of smooth games. In this work we take a first step towards closing this gap by developing average-ca se optimal first-order methods for a subset of smooth games.
We make the following three main contributions. First, we show that for zero-sum bilinear games the average-case optimal method is the optimal method for the mi nimization of the Hamiltonian. Second, we provide an explicit expression for the optimal method corresponding to normal matrices, potentially non-symmetric. Fin ally, we specialize it to matrices with eigenvalues located in a disk and show a provable speed-up compared to worst-case optimal algorithms. We illustrate our findings through benchmarks with a varying degree of mismatch with our assumptio ns.
****
Unsupervised Cross-lingual Representation Learning for Speech Recognition
Alexis Conneau,Alexei Baevski,Ronan Collobert,Abdelrahman Mohamed,Michael Auli
This paper presents XLSR which learns cross-lingual speech representations by pr etraining a single model from the raw waveform of speech in multiple languages. We build on wav2vec 2.0 which is trained by solving a contrastive task over mask ed latent speech representations and jointly learns a quantization of the latent s shared across languages. The resulting model is fine-tuned on labeled data and experiments show that cross-lingual pretraining significantly outperforms monol ingual pretraining. On the CommonVoice benchmark, XLSR shows a relative phoneme error rate reduction of 72% compared to the best known results. On BABEL, our ap proach improves word error rate by 16% relative compared to a comparable system. Our approach enables a single multilingual speech recognition model which is co mpetitive to strong individual models. Analysis shows that the latent discrete s peech representations are shared across languages with increased sharing for rel ated languages.
****
Subspace Clustering via Robust Self-Supervised Convolutional Neural Network
Dario Sitnik,Ivica Kopriva
Subspace clustering (SC) approaches based on the self-representation model achie ved encouraging performance when compared with the clustering algorithms that re ly on proximity measures between data points. However, they still face serious l imitations in real-world applications. One limitation relates to the linearity a ssumption of the self-representation model. The reason is that, usually, samples lie in non-linear manifolds, e.g. face images acquired under non-uniform illumi nation and different poses. Another limitation relates to errors that can be ran

dom or sample-specific (outliers), whereas in real-world applications their orig in is mostly unknown. Furthermore, the majority of existing algorithms use exter nal clustering validation methods to measure clustering quality, and that requir es access to ground-truth (GT) labels. Hence, it is practically important to dev elop a deep SC method that jointly learns self-expressive (SE) feature represent ation and handles data corruptions of unknown origin, and estimates clustering e rror using the internal clustering validation method, i.e. without access to the GT. Mostly, the base of the recently developed deep SC networks is convolutiona l autoencoder. It is an end-to-end fully convolutional network that is based on the minimization of the reconstruction error. Together, the autoencoder and an a dditional SE module are forming a Deep SC network (DSCNet). Hence, the total los s function of DSCNet is composed of reconstruction loss and SE loss. That is, du ring the learning process, the quality of clustering is not taken into account. Self-supervised convolutional SC network (S^2ConvSCN) addressed this issue thr ough the addition of a fully connected layer (FC) module and a spectral clusteri ng module that, respectively, generate soft- and pseudo-labels.

While inheriting the architecture of the S^2ConvSCN, this paper proposes a rob ust SE loss and an early self-stopping criterion for training. Robustness to arb itrary (unknown) types of data corruptions is achieved by using the correntropy induced metric (CIM) of the error of the SE model. By mapping the input data spa ce to a reproducible kernel Hilbert space (RKHS), correntropy defines an $\ell_2 $ distance in RKHS and creates nonlinear distance measure in the original input data space. Hence, correntropy is the optimal measure for error with the arbitra ry non-Gaussian distribution as opposed to the MSE that implies Gaussian distrib ution of errors. Thus, because it is estimated directly from data, CIM can handl e data corruption of unknown origin. As opposed to the DSCNet and S^2ConvSCN, the self-stopping criterion of the proposed algorithm is achieved by reaching a plateau in the change of loss value.

Although the architecture of S^2ConvSCN contains the FC module, which is capab le of handling out-of-sample data by using a softmax classifier, its performance has not been tested on unseen data. The importance of the FC module is, however , emphasized through self-supervision. In a truly unsupervised setting, pseudo-l abels generated from the spectral clustering module guide the learning process t hrough the induced classification error at the end of the FC module. They also e nhance the goodness of the self-representation matrix through penalizing incorre ctly connected elements. Adding these two loss functions to the total loss makes pseudo-labels an especially important component in label-constrained applicatio ns. Such dual self-supervision, in theory, enables deep networks to learn repres entation from available unlabeled data and to use the small number of labeled da ta to validate the trained model.

In addition to the main contributions of the paper, this study has three side-co ntributions. It aimed to set up a more transparent way for the proper performanc e estimation of deep SC models and to integrate block-diagonal regularization in to the gradient descent learning process. The majority of SC studies optimize hy perparameters using external clustering validation methodology, whereas the same labels are used for tuning hyperparameters and evaluating the final clustering performance. Furthermore, models like S^2ConvSCN suffer from an early commitme nt problem as they depend on weight-transfer from pretrained models that have “s een” the GT already (e.g. DSCNet transferred to S^2ConvSCN). From the machine learning perspective, data leakage and hyperparameter tuning based on the test G T are unacceptable. Measured clustering performances of S^2ConvSCN and DSCNet on the out-of-sample (unseen) data, thus, differ significantly from the optimist ic one presented in the original papers. Also, post-processing of self-represent ation matrix is reduced to a significant extent. Robust S^2ConvSCN outperform s its baseline version by a significant amount for both seen and unseen data on four well-known datasets. To the best of our knowledge, such an ablation study o n unseen data has not been reported previously in SC studies.

****
Graph-Based Continual Learning
Binh Tang,David S. Matteson
Despite significant advances, continual learning models still suffer from catast rophic forgetting when exposed to incrementally available data from non-stationa ry distributions. Rehearsal approaches alleviate the problem by maintaining and replaying a small episodic memory of previous samples, often implemented as an a rray of independent memory slots. In this work, we propose to augment such an ar ray with a learnable random graph that captures pairwise similarities between it s samples, and use it not only to learn new tasks but also to guard against forg etting. Empirical results on several benchmark datasets show that our model cons istently outperforms recently proposed baselines for task-free continual learnin g.
****
Learning Task-General Representations with Generative Neuro-Symbolic Modeling
Reuben Feinman,Brenden M. Lake
People can learn rich, general-purpose conceptual representations from only raw perceptual inputs. Current machine learning approaches fall well short of these human standards, although different modeling traditions often have complementary strengths. Symbolic models can capture the compositional and causal knowledge t hat enables flexible generalization, but they struggle to learn from raw inputs, relying on strong abstractions and simplifying assumptions. Neural network mode ls can learn directly from raw data, but they struggle to capture compositional and causal structure and typically must retrain to tackle new tasks. We bring to gether these two traditions to learn generative models of concepts that capture rich compositional and causal structure, while learning from raw data. We develo p a generative neuro-symbolic (GNS) model of handwritten character concepts that uses the control flow of a probabilistic program, coupled with symbolic stroke primitives and a symbolic image renderer, to represent the causal and compositio nal processes by which characters are formed. The distributions of parts (stroke s), and correlations between parts, are modeled with neural network subroutines, allowing the model to learn directly from raw data and express nonparametric st atistical relationships. We apply our model to the Omniglot challenge of human-l evel concept learning, using a background set of alphabets to learn an expressiv e prior distribution over character drawings. In a subsequent evaluation, our GN S model uses probabilistic inference to learn rich conceptual representations fr om a single training image that generalize to 4 unique tasks, succeeding where p revious work has fallen short.
****
Preventing Value Function Collapse in Ensemble Q-Learning by Maximizing Represe ntation Diversity
Hassam Sheikh,Ladislau Boloni
The first deep RL algorithm, DQN, was limited by the overestimation bias of the learned Q-function. Subsequent algorithms proposed techniques to reduce this pro blem, without fully eliminating it. Recently, the Maxmin and Ensemble Q-learning algorithms used the different estimates provided by ensembles of learners to re duce the bias. Unfortunately, these learners can converge to the same point in t he parametric or representation space, falling back to the classic single neural network DQN. In this paper, we describe a regularization technique to maximize diversity in the representation space in these algorithms. We propose and compar e five regularization functions inspired from economics theory and consensus opt imization. We show that the resulting approach significantly outperforms the Max min and Ensemble Q-learning algorithms as well as non-ensemble baselines.
****
Neural spatio-temporal reasoning with object-centric self-supervised learning
David Ding,Felix Hill,Adam Santoro,Matthew Botvinick
Transformer-based language models have proved capable of rudimentary symbolic re asoning, underlining the effectiveness of applying self-attention computations to sets of discrete entities. In this work, we apply this lesson to videos of ph ysical interaction between objects. We show that self-attention-based models ope

rating on discrete, learned, object-centric representations perform well on spatio-temporal reasoning tasks which were expressly designed to trouble traditional neural network models and to require higher-level cognitive processes such as causal reasoning and understanding of intuitive physics and narrative structure. We achieve state of the art results on two datasets, CLEVRER and CATER, significantly outperforming leading hybrid neuro-symbolic models. Moreover, we find that techniques from language modelling, such as BERT-style semi-supervised predictive losses, allow our model to surpass neuro-symbolic approaches while using 40% less labelled data. Our results corroborate the idea that neural networks can reason about the causal, dynamic structure of visual data and attain understanding of intuitive physics, which counters the popular claim that they are only effective at perceptual pattern-recognition and not reasoning per se.
****

Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study
Zhiqiang Shen,Zechun Liu,Dejia Xu,Zitian Chen,Kwang-Ting Cheng,Marios Savvides
This work aims to empirically clarify a recently discovered perspective that label smoothing is incompatible with knowledge distillation. We begin by introducing the motivation behind on how this incompatibility is raised, i.e., label smoothing erases relative information between teacher logits. We provide a novel connection on how label smoothing affects distributions of semantically similar and dissimilar classes. Then we propose a metric to quantitatively measure the degree of erased information in sample's representation. After that, we study its one-sidedness and imperfection of the incompatibility view through massive analyses, visualizations and comprehensive experiments on Image Classification, Binary Networks, and Neural Machine Translation. Finally, we broadly discuss several circumstances wherein label smoothing will indeed lose its effectiveness.
****

MVP-BERT: Redesigning Vocabularies for Chinese BERT and Multi-Vocab Pretraining
Wei Zhu
Despite the development of pre-trained language models (PLMs) significantly raise the performances of various Chinese natural language processing (NLP) tasks, the vocabulary for these Chinese PLMs remain to be the one provided by Google Chinese Bert, which is based on Chinese characters. Second, the masked language model pre-training is based on a single vocabulary, which limits its downstream task performances. In this work, we first propose a novel method, seg_tok, to form the vocabulary of Chinese BERT, with the help of Chinese word segmentation (CWS) and subword tokenization. Then we propose three versions of multi-vocabulary pretraining (MVP) to improve the models expressiveness. Experiments show that: (a) compared with char based vocabulary, seg_tok does not only improves the performances of Chinese PLMs on sentence level tasks, it can also improve efficiency; (b) MVP improves PLMs' downstream performance, especially it can improve seg_tok's performances on sequence labeling tasks.
****

Sparse Quantized Spectral Clustering
Zhenyu Liao,Romain Couillet,Michael W. Mahoney
Given a large data matrix, sparsifying, quantizing, and/or performing other entry-wise nonlinear operations can have numerous benefits, ranging from speeding up iterative algorithms for core numerical linear algebra problems to providing nonlinear filters to design state-of-the-art neural network models. Here, we exploit tools from random matrix theory to make precise statements about how the eigenspectrum of a matrix changes under such nonlinear transformations. In particular, we show that very little change occurs in the informative eigenstructure, even under drastic sparsification/quantization, and consequently that very little downstream performance loss occurs when working with very aggressively sparsified or quantized spectral clustering problems.
We illustrate how these results depend on the nonlinearity, we characterize a phase transition beyond which spectral clustering becomes possible, and we show when such nonlinear transformations can introduce spurious non-informative eigenvectors.

****
Factor Normalization for Deep Neural Network Models
Haobo Qi,Jing Zhou,Hansheng Wang
Deep neural network (DNN) models often involve features of high dimensions. In most cases, the high-dimensional features can be decomposed into two parts. The first part is a low-dimensional factor. The second part is the residual feature, with much-reduced variability and inter-feature correlation. This leads to a number of interesting theoretical findings for deep neural network training. Accordingly, we are inspired to develop a new factor normalization method for better performance. The proposed method leads to a new deep learning model with two important features. First, it allows factor related feature extraction. Second, it allows adaptive learning rates for factors and residuals, respectively. This leads to fast convergence speed on both training and validation datsets. A number of empirical experiments are presented to demonstrate its superior performance. The code is available at https://github.com/HazardNeo4869/FactorNormalization
****
The Unreasonable Effectiveness of Patches in Deep Convolutional Kernels Methods
Louis THIRY,Michael Arbel,Eugene Belilovsky,Edouard Oyallon
A recent line of work showed that various forms of convolutional kernel methods can be competitive with standard supervised deep convolutional networks on datasets like CIFAR-10, obtaining accuracies in the range of 87-90% while being more amenable to theoretical analysis. In this work, we highlight the importance of a data-dependent feature extraction step that is key to the obtain good performance in convolutional kernel methods. This step typically corresponds to a whitened dictionary of patches, and gives rise to a data-driven convolutional kernel methods.We extensively study its effect, demonstrating it is the key ingredient for high performance of these methods. Specifically, we show that one of the simplest instances of such kernel methods, based on a single layer of image patches followed by a linear classifier is already obtaining classification accuracies on CIFAR-10 in the same range as previous more sophisticated convolutional kernel methods. We scale this method to the challenging ImageNet dataset, showing such a simple approach can exceed all existing non-learned representation methods. This is a new baseline for object recognition without representation learning methods, that initiates the investigation of convolutional kernel models on ImageNet. We conduct experiments to analyze the dictionary that we used, our ablations showing they exhibit low-dimensional properties.
****
Graph Coarsening with Neural Networks
Chen Cai,Dingkang Wang,Yusu Wang
As large scale-graphs become increasingly more prevalent, it poses significant computational challenges to process, extract and analyze large graph data. Graph coarsening is one popular technique to reduce the size of a graph while maintaining essential properties. Despite rich graph coarsening literature, there is only limited exploration of data-driven method in the field. In this work, we leverage the recent progress of deep learning on graphs for graph coarsening. We first propose a framework for measuring the quality of coarsening algorithm and show that depending on the goal, we need to carefully choose the Laplace operator on the coarse graph and associated projection/lift operators. Motivated by the observation that the current choice of edge weight for the coarse graph may be sub-optimal, we parametrize the weight assignment map with graph neural networks and train it to improve the coarsening quality in an unsupervised way. Through extensive experiments on both synthetic and real networks, we demonstrate that our method significantly improves common graph coarsening methods under various metrics, reduction ratios, graph sizes, and graph types. It generalizes to graphs of larger size (more than $25\times$ of training graphs), adaptive to different losses (both differentiable and non-differentiable), and scales to much larger graphs than previous work.
****
Cross-State Self-Constraint for Feature Generalization in Deep Reinforcement Learning

Guan Ting Liu,Pu-Jen Cheng,GuanYu Lin
Representation learning on visualized input is an important yet challenging task for deep reinforcement learning (RL). The feature space learned from visualized input not only dominates the agent's generalization ability in new environments but also affect the data efficiency during training. To help the RL agent learn general and discriminative representation among various states, we present cross-state self-constraint(CSSC), a novel constraint that regularizes the representation feature space by comparing similarity of different pairs of representations. Based on the representation-behavior connection derived from the agent's experience, this constraint helps reinforce the general feature recognition during the learning process and thus enhance the generalization to unseen environment. We test our proposed method on the OpenAI ProcGen benchmark and see significant improvement on generalization performance across most of ProcGen games.
****
On Size Generalization in Graph Neural Networks
Gilad Yehudai,Ethan Fetaya,Eli Meirom,Gal Chechik,Haggai Maron
Graph neural networks (GNNs) can process graphs of different sizes but their capacity to generalize across sizes is still not well understood. Size generalization is key to numerous GNN applications, from solving combinatorial optimization problems to learning in molecular biology. In such problems, obtaining labels and training on large graphs can be prohibitively expensive, but training on smaller graphs is possible.

This paper puts forward the size-generalization question and characterizes important aspects of that problem theoretically and empirically.
We prove that even for very simple tasks, such as counting the number of nodes or edges in a graph, GNNs do not naturally generalize to graphs of larger size. Instead, their generalization performance is closely related to the distribution of local patterns of connectivity and features and how that distribution changes from small to large graphs. Specifically, we prove that for many tasks, there are weight assignments for GNNs that can perfectly solve the task on small graphs but fail on large graphs, if there is a discrepancy between their local patterns. We further demonstrate on several tasks, that training GNNs on small graphs results in solutions which do not generalize to larger graphs. We then formalize size generalization as a domain-adaption problem and describe two learning setups where size generalization can be improved. First, as a self-supervised learning problem (SSL) over the target domain of large graphs. Second as a semi-supervised learning problem when few samples are available in the target domain. We demonstrate the efficacy of these solutions on a diverse set of benchmark graph datasets.
****
On the Universality of the Double Descent Peak in Ridgeless Regression
David Holzmüller
We prove a non-asymptotic distribution-independent lower bound for the expected mean squared generalization error caused by label noise in ridgeless linear regression. Our lower bound generalizes a similar known result to the overparameterized (interpolating) regime. In contrast to most previous works, our analysis applies to a broad class of input distributions with almost surely full-rank feature matrices, which allows us to cover various types of deterministic or random feature maps. Our lower bound is asymptotically sharp and implies that in the presence of label noise, ridgeless linear regression does not perform well around the interpolation threshold for any of these feature maps. We analyze the imposed assumptions in detail and provide a theory for analytic (random) feature maps. Using this theory, we can show that our assumptions are satisfied for input distributions with a (Lebesgue) density and feature maps given by random deep neural networks with analytic activation functions like sigmoid, tanh, softplus or GELU. As further examples, we show that feature maps from random Fourier features and polynomial kernels also satisfy our assumptions. We complement our theory with further experimental and analytic results.
****

Fast Binarized Neural Network Training with Partial Pre-training
Alex Renda,Joshua Wolff Fromm
Binarized neural networks, networks with weights and activations constrained to lie in a 2-element set, allow for more time- and resource-efficient inference than standard floating-point networks. However, binarized neural networks typically take more training to plateau in accuracy than their floating-point counterparts, in terms of both iteration count and wall clock time. We demonstrate a technique, partial pre-training, that allows for faster from-scratch training of binarized neural networks by first training the network as a standard floating-point network for a short amount of time, then converting the network to a binarized neural network and continuing to train from there. Without tuning any hyperparameters across four networks on three different datasets, partial pre-training is able to train binarized neural networks between $1.26\times$ and $1.61\times$ faster than when training a binarized network from scratch using standard low-precision training.

****
Differentiable Graph Optimization for Neural Architecture Search
Chengyue Huang,Lingfei Wu,Yadong Ding,Siliang Tang,Fangli Xu,Chang Zong,Chilie Tan,Yueting Zhuang
In this paper, we propose Graph Optimized Neural Architecture Learning (GOAL), a novel gradient-based method for Neural Architecture Search (NAS), to find better architectures with fewer evaluated samples. Popular NAS methods usually employ black-box optimization based approaches like reinforcement learning, evolution algorithm or Bayesian optimization, which may be inefficient when having huge combinatorial NAS search spaces. In contrast, we aim to explicitly model the NAS search space as graphs, and then perform gradient-based optimization to learn graph structure with efficient exploitation. To this end, we learn a differentiable graph neural network as a surrogate model to rank candidate architectures, which enable us to obtain gradient w.r.t the input architectures. To cope with the difficulty in gradient-based optimization on the discrete graph structures, we propose to leverage proximal gradient descent to find potentially better architectures.
Our empirical results show that GOAL outperforms mainstream black-box methods on existing NAS benchmarks in terms of search efficiency.
****
Deep Repulsive Clustering of Ordered Data Based on Order-Identity Decomposition
Seon-Ho Lee,Chang-Su Kim
We propose the deep repulsive clustering (DRC) algorithm of ordered data for effective order learning. First, we develop the order-identity decomposition (ORID) network to divide the information of an object instance into an order-related feature and an identity feature. Then, we group object instances into clusters according to their identity features using a repulsive term. Moreover, we estimate the rank of a test instance, by comparing it with references within the same cluster. Experimental results on facial age estimation, aesthetic score regression, and historical color image classification show that the proposed algorithm can cluster ordered data effectively and also yield excellent rank estimation performance.
****
Rapid Task-Solving in Novel Environments
Samuel Ritter,Ryan Faulkner,Laurent Sartran,Adam Santoro,Matthew Botvinick,David Raposo
We propose the challenge of rapid task-solving in novel environments (RTS), wherein an agent must solve a series of tasks as rapidly as possible in an unfamiliar environment. An effective RTS agent must balance between exploring the unfamiliar environment and solving its current task, all while building a model of the new environment over which it can plan when faced with later tasks. While modern deep RL agents exhibit some of these abilities in isolation, none are suitable for the full RTS challenge. To enable progress toward RTS, we introduce two challenge domains: (1) a minimal RTS challenge called the Memory&Planning Game and (

2) One-Shot StreetLearn Navigation, which introduces scale and complexity from r
eal-world data. We demonstrate that state-of-the-art deep RL agents fail at RTS
in both domains, and that this failure is due to an inability to plan over gathe
red knowledge. We develop Episodic Planning Networks (EPNs) and show that deep-R
L agents with EPNs excel at RTS, outperforming the nearest baseline by factors o
f 2-3 and learning to navigate held-out StreetLearn maps within a single episode
. We show that EPNs learn to execute a value iteration-like planning algorithm a
nd that they generalize to situations beyond their training experience.
****

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 o
f a continuous and high-dimensional environment that is suitable for planning. S
uch 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 simple domain where
 the agent learns a compact, lifted representation that generalises across objec
ts. We then apply it to a series of Minecraft tasks to learn object-centric repr
esentations, including object types—directly from pixel data—that can be leverag
ed to solve new tasks quickly. The resulting learned representations enable the
use of a task-level planner, resulting in an agent capable of forming complex, l
ong-term plans with considerably fewer environment interactions.
****

Spherical Motion Dynamics: Learning Dynamics of Neural Network with Normalizatio
n, Weight Decay, and SGD
Ruosi Wan,Zhanxing Zhu,Xiangyu Zhang,Jian Sun
In this work, we comprehensively reveal the learning dynamics of neural network
with normalization, weight decay (WD), and SGD (with momentum), named as Spheric
al Motion Dynamics (SMD). Most related works study SMD by focusing on "effective
 learning rate" in "equilibrium" condition, where weight norm remains unchanged.
 However, their discussions on why equilibrium condition can be reached in SMD i
s either absent or less convincing. Our work investigates SMD by directly explor
ing the cause of equilibrium condition. Specifically, 1) we introduce the assump
tions that can lead to equilibrium condition in SMD, and prove that weight norm
can converge at linear rate with given assumptions; 2) we propose "angular updat
e" as a substitute for effective learning rate to measure the evolving of neural
 network in SMD, and prove angular update can also converge to its theoretical
value at linear rate; 3) we verify our assumptions and theoretical results on va
rious computer vision tasks including ImageNet and MSCOCO with standard settings
. Experiment results show our theoretical findings agree well with empirical obs
ervations.
****

DINO: A Conditional Energy-Based GAN for Domain Translation
Konstantinos Vougioukas,Stavros Petridis,Maja Pantic
Domain translation is the process of transforming data from one domain to anothe
r while preserving the common semantics. Some of the most popular domain transla
tion systems are based on conditional generative adversarial networks, which use
 source domain data to drive the generator and as an input to the discriminator.
 However, this approach does not enforce the preservation of shared semantics si
nce the conditional input can often be ignored by the discriminator. We propose
an alternative method for conditioning and present a new framework, where two ne
tworks are simultaneously trained, in a supervised manner, to perform domain tra
nslation in opposite directions. Our method is not only better at capturing the
shared information between two domains but is more generic and can be applied to
 a broader range of problems. The proposed framework performs well even in chall
enging cross-modal translations, such as video-driven speech reconstruction, for
 which other systems struggle to maintain correspondence.
****

Self-Supervised Video Representation Learning with Constrained Spatiotemporal Ji
gsaw

Yuqi Huo,Mingyu Ding,Haoyu Lu,Zhiwu Lu,Tao Xiang,Ji-Rong Wen,Ziyuan Huang,Jianwen Jiang,Shiwei Zhang,Mingqian Tang,Songfang Huang,Ping Luo
This paper proposes a novel pretext task for self-supervised video representation learning by exploiting spatiotemporal continuity in videos. It is motivated by the fact that videos are spatiotemporal by nature and a representation learned to detect spatiotemporal continuity/discontinuity is thus beneficial for downstream video content analysis tasks. A natural choice of such a pretext task is to construct spatiotemporal (3D) jigsaw puzzles and learn to solve them. However, this task turns out to be intractable. We thus propose Constrained Spatiotemporal Jigsaw (CSJ) whereby the 3D jigsaws are formed in a constrained manner to ensure that large continuous spatiotemporal cuboids exist in a shuffled clip to provide sufficient cues for the model to reason about the continuity. With the constrained jigsaw puzzles, instead of solving them directly, which could still be extremely hard, we carefully design four surrogate tasks that are more solvable but meanwhile still ensure that the learned representation is sensitive to spatiotemporal continuity at both the local and global levels. Extensive experiments show that our CSJ achieves state-of-the-art on two downstream tasks across various benchmarks.
****
Removing Undesirable Feature Contributions Using Out-of-Distribution Data
Saehyung Lee,Changhwa Park,Hyungyu Lee,Jihun Yi,Jonghyun Lee,Sungroh Yoon
Several data augmentation methods deploy unlabeled-in-distribution (UID) data to bridge the gap between the training and inference of neural networks. However, these methods have clear limitations in terms of availability of UID data and dependence of algorithms on pseudo-labels. Herein, we propose a data augmentation method to improve generalization in both adversarial and standard learning by using out-of-distribution (OOD) data that are devoid of the abovementioned issues. We show how to improve generalization theoretically using OOD data in each learning scenario and complement our theoretical analysis with experiments on CIFAR-10, CIFAR-100, and a subset of ImageNet. The results indicate that undesirable features are shared even among image data that seem to have little correlation from a human point of view. We also present the advantages of the proposed method through comparison with other data augmentation methods, which can be used in the absence of UID data. Furthermore, we demonstrate that the proposed method can further improve the existing state-of-the-art adversarial training.
****
Cross-Modal Retrieval Augmentation for Multi-Modal Classification
Shir Gur,Natalia Neverova,Chris Stauffer,Ser-Nam Lim,Douwe Kiela,Austin Reiter
Recent advances in using retrieval components over external knowledge sources have shown impressive results for a variety of downstream tasks in natural language processing. Here, we explore the use of unstructured external knowledge sources of images and their corresponding captions for improving visual question answering (VQA). First, we train a novel alignment model for embedding images and captions in the same space, which achieves state-of-the-art image-caption retrieval performance w.r.t. similar methods.
Second, we show that retrieval-augmented multi-modal transformers using the trained alignment model
significantly improve results on VQA over strong baselines.
We further conduct extensive experiments to establish the promise of this approach, and examine novel applications for inference time such as hot-swapping indices.
****
Self-Supervised Multi-View Learning via Auto-Encoding 3D Transformations
Xiang Gao,Wei Hu,Guo-Jun Qi
3D object representation learning is a fundamental challenge in computer vision to draw inferences about the 3D world. Recent advances in deep learning have shown their efficiency in 3D object recognition, among which view-based methods have performed best so far. However, feature learning of multiple views in existing methods is mostly trained in a supervised fashion, which often requires a large amount of data labels with high cost. Hence, it is critical to learn multi-view

feature representations in a self-supervised fashion. To this end, we propose a novel self-supervised learning paradigm of Multi-View Transformation Equivariant Representations (MV-TER), exploiting the equivariant transformations of a 3D object and its projected multiple views. Specifically, we perform a 3D transformation on a 3D object, and obtain multiple views before and after transformation via projection. Then, we self-train a representation learning module to capture the intrinsic 3D object representation by decoding 3D transformation parameters from the fused feature representations of multiple views before and after transformation. Experimental results demonstrate that the proposed MV-TER significantly outperforms the state-of-the-art view-based approaches in 3D object classification and retrieval tasks.

****

TopoTER: Unsupervised Learning of Topology Transformation Equivariant Representations

Xiang Gao,Wei Hu,Guo-Jun Qi

We present the Topology Transformation Equivariant Representation (TopoTER) learning, a general paradigm of unsupervised learning of node representations of graph data for the wide applicability to Graph Convolutional Neural Networks (GCNNs). We formalize the TopoTER from an information-theoretic perspective, by maximizing the mutual information between topology transformations and node representations before and after the transformations. We derive that maximizing such mutual information can be relaxed to minimizing the cross entropy between the applied topology transformation and its estimation from node representations. In particular, we seek to sample a subset of node pairs from the original graph and flip the edge connectivity between each pair to transform the graph topology. Then, we self-train a representation encoder to learn node representations by reconstructing the topology transformations from the feature representations of the original and transformed graphs. In experiments, we apply the TopoTER to the downstream node and graph classification tasks, and results show that the TopoTER outperforms the state-of-the-art unsupervised approaches.

****

Generalized Gumbel-Softmax Gradient Estimator for Generic Discrete Random Variables

Weonyoung Joo,Dongjun Kim,Seungjae Shin,Il-chul Moon

Estimating the gradients of stochastic nodes, which enables the gradient descent optimization on neural network parameters, is one of the crucial research questions in the deep generative modeling community. When it comes to discrete distributions, Gumbel-Softmax trick reparameterizes Bernoulli and categorical random variables by continuous relaxation. However, gradient estimators of discrete distributions other than the Bernoulli and the categorical have not been explored, and the the Gumbel-Softmax trick is not directly applicable to other discrete distributions. This paper proposes a general version of the Gumbel-Softmax estimator with a theoretical basis, and the proposed estimator is able to reparameterize generic discrete distributions, broader than the Bernoulli and the categorical. In detail, we utilize the truncation of discrete random variables and the Gumbel-Softmax trick with a linear transformation for the relaxed reparameterization. The proposed approach enables the relaxed discrete random variable to be reparameterized through a large-scale stochastic computational graph. Our experiments consist of (1) synthetic data analyses and applications on VAE, which show the efficacy of our methods; and (2) topic models, which demonstrate the value of the proposed estimation in practice.

****

Post-Training Weighted Quantization of Neural Networks for Language Models

Se Jung Kwon,Dongsoo Lee,Yongkweon Jeon,Byeongwook Kim,Bae Seong Park,Yeonju Ro

As a practical model compression technique, parameter quantization is effective especially for language models associated with a large memory footprint. Neural network quantization is usually performed to reduce quantization loss assuming that quantization error of each parameter equally contributes to the overall training loss. The importance of each parameter, however, may highly differ such that for the same number of quantization bits, certain parameters lead to higher tr

aining loss than the others after quantization. In this paper, we consider a non-uniform quantization scheme, specifically binary-coding-based quantization, for high compression ratio and efficient computations while avoiding large accuracy degradation by uniform quantization (e.g., INT8). Then, we derive quantization optimization methods to take into account the importance of each parameter. We demonstrate that for post-training quantization, weight magnitude can represent importance and improve model accuracy significantly compared to the previous schemes lacking importance considerations. For various language models including BERT, DistilBERT, AWD-LSTM, and Transformer, we achieve 2-4 bits per weight by our proposed post-training quantization with reasonable accuracy degradation.
****

Gradient Flow in Sparse Neural Networks and How Lottery Tickets Win
Utku Evci,Yani Ioannou,Cem Keskin,Yann Dauphin
Sparse Neural Networks (NNs) can match the generalization of dense NNs using a fraction of the compute/storage for inference, and also have the potential to enable efficient training. However, naively training unstructured sparse NNs from random initialization results in significantly worse generalization, with the notable exception of Lottery Tickets (LTs) and Dynamic Sparse Training (DST). In this work, we attempt to answer: (1) why training unstructured sparse networks from random initialization performs poorly and; and (2) what makes LTs and DST the exceptions? We show that sparse NNs have poor gradient flow at initialization and propose a modified initialization for unstructured connectivity. Furthermore, we find that DST methods significantly improve gradient flow during training over traditional sparse training methods. Finally, we show that LTs do not improve gradient flow, rather their success lies in re-learning the pruning solution they are derived from – however, this comes at the cost of learning novel solutions.
****

Learning Disconnected Manifolds: Avoiding The No Gan's Land by Latent Rejection
Thibaut Issenhuth,Ugo Tanielian,David Picard,Jeremie Mary
Standard formulations of GANs, where a continuous function deforms a connected latent space, have been shown to be misspecified when fitting disconnected manifolds. In particular, when covering different classes of images, the generator will necessarily sample some low quality images in between the modes. Rather than modify the learning procedure, a line of works aims at improving the sampling quality from trained generators. Thus, it is now common to introduce a rejection step within the generation procedure.
Building on this, we propose to train an additional network and transform the latent space via an adversarial learning of importance weights. This idea has several advantages: 1) it provides a way to inject disconnectedness on any GAN architecture, 2) the rejection avoids going through both the generator and the discriminator saving computation time, 3) this importance weights formulation provides a principled way to estimate the Wasserstein's distance to the true distribution, enabling its minimization. We demonstrate the effectiveness of our method on different datasets, both synthetic and high dimensional, and stress its superiority on highly disconnected data.
****

Temporal Difference Uncertainties as a Signal for Exploration
Sebastian Flennerhag,Jane X Wang,Pablo Sprechmann,Francesco Visin,Alexandre Galashov,Steven Kapturowski,Diana L Borsa,Nicolas Heess,Andre Barreto,Razvan Pascanu
An effective approach to exploration in reinforcement learning is to rely on an agent's uncertainty over the optimal policy, which can yield near-optimal exploration strategies in tabular settings. However, in non-tabular settings that involve function approximators, obtaining accurate uncertainty estimates is almost as challenging as the exploration problem itself. In this paper, we highlight that value estimates are easily biased and temporally inconsistent. In light of this, we propose a novel method for estimating uncertainty over the value function that relies on inducing a distribution over temporal difference errors. This exploration signal controls for state-action transitions so as to isolate uncertainty in value that is due to uncertainty over the agent's parameters. Because our

measure of uncertainty conditions on state-action transitions, we cannot act on this measure directly. Instead, we incorporate it as an intrinsic reward and treat exploration as a separate learning problem, induced by the agent's temporal difference uncertainties. We introduce a distinct exploration policy that learns to collect data with high estimated uncertainty, which gives rise to a curriculum that smoothly changes throughout learning and vanishes in the limit of perfect value estimates. We evaluate our method on hard exploration tasks, including Deep Sea and Atari 2600 environments and find that our proposed form of exploration facilitates efficient exploration.
****

Federated Continual Learning with Weighted Inter-client Transfer
Jaehong Yoon,Wonyong Jeong,Giwoong Lee,Eunho Yang,Sung Ju Hwang
There has been a surge of interest in continual learning and federated learning, both of which are important in deep neural networks in real-world scenarios. Yet little research has been done regarding the scenario where each client learns on a sequence of tasks from a private local data stream. This problem of federated continual learning poses new challenges to continual learning, such as utilizing knowledge from other clients, while preventing interference from irrelevant knowledge. To resolve these issues, we propose a novel federated continual learning framework, Federated Weighted Inter-client Transfer (FedWeIT), which decomposes the network weights into global federated parameters and sparse task-specific parameters, and each client receives selective knowledge from other clients by taking a weighted combination of their task-specific parameters.FedWeITminimizes interference between incompatible tasks, and also allows positive knowledge transfer across clients during learning. We validate ourFedWeITagainst existing federated learning and continual learning methods under varying degrees of task similarity across clients, and our model significantly outperforms them with a large reduction in the communication cost.
****

Accurate Learning of Graph Representations with Graph Multiset Pooling
Jinheon Baek,Minki Kang,Sung Ju Hwang
Graph neural networks have been widely used on modeling graph data, achieving impressive results on node classification and link prediction tasks. Yet, obtaining an accurate representation for a graph further requires a pooling function that maps a set of node representations into a compact form. A simple sum or average over all node representations considers all node features equally without consideration of their task relevance, and any structural dependencies among them. Recently proposed hierarchical graph pooling methods, on the other hand, may yield the same representation for two different graphs that are distinguished by the Weisfeiler-Lehman test, as they suboptimally preserve information from the node features. To tackle these limitations of existing graph pooling methods, we first formulate the graph pooling problem as a multiset encoding problem with auxiliary information about the graph structure, and propose a Graph Multiset Transformer (GMT) which is a multi-head attention based global pooling layer that captures the interaction between nodes according to their structural dependencies. We show that GMT satisfies both injectiveness and permutation invariance, such that it is at most as powerful as the Weisfeiler-Lehman graph isomorphism test. Moreover, our methods can be easily extended to the previous node clustering approaches for hierarchical graph pooling. Our experimental results show that GMT significantly outperforms state-of-the-art graph pooling methods on graph classification benchmarks with high memory and time efficiency, and obtains even larger performance gain on graph reconstruction and generation tasks.
****

AN ONLINE SEQUENTIAL TEST FOR QUALITATIVE TREATMENT EFFECTS
Chengchun Shi,Shikai Luo,Rui Song,Hongtu Zhu
Tech companies (e.g., Google or Facebook) often use randomized online experiments and/or A/B testing primarily based on the average treatment effects to compare their new product with an old one. However, it is also critically important to detect qualitative treatment effects such that the new one may significantly outperform the existing one only under some specific circumstances. The aim of this

paper is to develop a powerful testing procedure to efficiently detect such qua
litative treatment effects. We propose a scalable online updating algorithm to i
mplement our test procedure. It has three novelties including adaptive randomiza
tion, sequential monitoring, and online updating with guaranteed type-I error co
ntrol. We also thoroughly examine the theoretical proper- ties of our testing pr
ocedure including the limiting distribution of test statistics and the justifica
tion of an efficient bootstrap method. Extensive empirical studies are conducted
to examine the finite sample performance of our test procedure.
****

Dream and Search to Control: Latent Space Planning for Continuous Control
Anurag Koul,Varun Kumar Vijay,Alan Fern,Somdeb Majumdar
Learning and planning with latent space dynamics has been shown to be useful for
sample efficiency in model-based reinforcement learning (MBRL) for discrete and
continuous control tasks. In particular, recent work, for discrete action space
s, demonstrated the effectiveness of latent-space planning via Monte-Carlo Tree
Search (MCTS) for bootstrapping MBRL during learning and at test time. However,
the potential gains from latent-space tree search have not yet been demonstrated
for environments with continuous action spaces. In this work, we propose and e
xplore an MBRL approach for continuous action spaces based on tree-based plannin
g over learned latent dynamics. We show that it is possible to demonstrate the
types of bootstrapping benefits as previously shown for discrete spaces. In part
icular, the approach achieves improved sample efficiency and performance on a ma
jority of challenging continuous-control benchmarks compared to the state-of-the
-art.
****

Federated Semi-Supervised Learning with Inter-Client Consistency & Disjoint Lear
ning
Wonyong Jeong,Jaehong Yoon,Eunho Yang,Sung Ju Hwang
While existing federated learning approaches mostly require that clients have fu
lly-labeled data to train on, in realistic settings, data obtained at the client
-side often comes without any accompanying labels. Such deficiency of labels may
result from either high labeling cost, or difficulty of annotation due to the r
equirement of expert knowledge. Thus the private data at each client may be eith
er partly labeled, or completely unlabeled with labeled data being available onl
y at the server, which leads us to a new practical federated learning problem, n
amely Federated Semi-Supervised Learning (FSSL). In this work, we study two esse
ntial scenarios of FSSL based on the location of the labeled data. The first sce
nario considers a conventional case where clients have both labeled and unlabele
d data (labels-at-client), and the second scenario considers a more challenging
case, where the labeled data is only available at the server (labels-at-server).
We then propose a novel method to tackle the problems, which we refer to as Fed
erated Matching (FedMatch). FedMatch improves upon naive combinations of federat
ed learning and semi-supervised learning approaches with a new inter-client cons
istency loss and decomposition of the parameters for disjoint learning on labele
d and unlabeled data. Through extensive experimental validation of our method in
the two different scenarios, we show that our method outperforms both local sem
i-supervised learning and baselines which naively combine federated learning wit
h semi-supervised learning.
****

Contrastive Learning with Adversarial Perturbations for Conditional Text Gener
ation
Seanie Lee,Dong Bok Lee,Sung Ju Hwang
Recently, sequence-to-sequence (seq2seq) models with the Transformer architectur
e have achieved remarkable performance on various conditional text generation ta
sks, such as machine translation. However, most of them are trained with teacher
forcing with the ground truth label given at each time step, without being expo
sed to incorrectly generated tokens during training, which hurts its generalizat
ion to unseen inputs, that is known as the "exposure bias" problem. In this work
, we propose to solve the conditional text generation problem by contrasting pos
itive pairs with negative pairs, such that the model is exposed to various valid

or incorrect perturbations of the inputs, for improved generalization. However, training the model with naïve contrastive learning framework using random non-target sequences as negative examples is suboptimal, since they are easily distinguishable from the correct output, especially so with models pretrained with large text corpora. Also, generating positive examples requires domain-specific augmentation heuristics which may not generalize over diverse domains. To tackle this problem, we propose a principled method to generate positive and negative samples for contrastive learning of seq2seq models. Specifically, we generate negative examples by adding small perturbations to the input sequence to minimize its conditional likelihood, and positive examples by adding large perturbations while enforcing it to have a high conditional likelihood. Such `"hard'' positive and negative pairs generated using our method guides the model to better distinguish correct outputs from incorrect ones. We empirically show that our proposed method significantly improves the generalization of the seq2seq on three text generation tasks --- machine translation, text summarization, and question generation.

****

A REINFORCEMENT LEARNING FRAMEWORK FOR TIME DEPENDENT CAUSAL EFFECTS EVALUATION IN A/B TESTING

Chengchun Shi,Xiaoyu Wang,Shikai Luo,Rui Song,Hongtu Zhu,Jieping Ye

A/B testing, or online experiment is a standard business strategy to compare a new product with an old one in pharmaceutical, technological, and traditional industries. The aim of this paper is to introduce a reinforcement learn- ing framework for carrying A/B testing in two-sided marketplace platforms, while characterizing the long-term treatment effects. Our proposed testing procedure allows for sequential monitoring and online updating. It is generally applicable to a variety of treatment designs in different industries. In addition, we systematically investigate the theoretical properties (e.g., size and power) of our testing procedure. Finally, we apply our framework to both synthetic data and a real-world data example obtained from a technological company to illustrate its advantage over the current practice.

****

EM-RBR: a reinforced framework for knowledge graph completion from reasoning perspective

Bozhou Chen,Zhaochong An,Houde Quan,Qihui Lin,Hongzhi Wang

Knowledge graph completion aims to predict the new links in given entities among the knowledge graph (KG). Most mainstream embedding methods focus on fact triplets contained in the given KG, however, ignoring the rich background information provided by logic rules driven from knowledge base implicitly. To solve this problem, in this paper, we propose a general framework, named EM-RBR(embedding and rule-based reasoning), capable of combining the advantages of reasoning based on rules and the state-of-the-art models of embedding. EM-RBR aims to utilize relational background knowledge contained in rules to conduct multi-relation reasoning link prediction rather than superficial vector triangle linkage in embedding models. By this way, we can explore relation between two entities in deeper context to achieve higher accuracy. In experiments, we demonstrate that EM-RBR achieves better performance compared with previous models on FB15k, WN18 and our new dataset FB15k-R, especially the new dataset where our model perform futher better than those state-of-the-arts. We make the implementation of EM-RBR available at https://github.com/1173710224/link-prediction-with-rule-based-reasoning.

****

Learning Intrinsic Symbolic Rewards in Reinforcement Learning

Hassam Sheikh,Shauharda Khadka,Santiago Miret,Somdeb Majumdar

Learning effective policies for sparse objectives is a key challenge in Deep Reinforcement Learning (RL). A common approach is to design task-related dense rewards to improve task learnability. While such rewards are easily interpreted, they rely on heuristics and domain expertise. Alternate approaches that train neural networks to discover dense surrogate rewards avoid heuristics, but are high-dimensional, black-box solutions offering little interpretability. In this paper,

we present a method that discovers dense rewards in the form of low-dimensional symbolic trees - thus making them more tractable for analysis. The trees use simple functional operators to map an agent's observations to a scalar reward, which then supervises the policy gradient learning of a neural network policy. We test our method on continuous action spaces in Mujoco and discrete action spaces in Atari and Pygames environments. We show that the discovered dense rewards are an effective signal for an RL policy to solve the benchmark tasks. Notably, we significantly outperform a widely used, contemporary neural-network based reward-discovery algorithm in all environments considered.
****
Safety Aware Reinforcement Learning (SARL)
Santiago Miret,Somdeb Majumdar,Carroll Wainwright
As reinforcement learning agents become increasingly integrated into complex, real-world environments, designing for safety becomes a critical consideration. We specifically focus on researching scenarios where agents can cause undesired side effects while executing a policy on a primary task. Since one can define multiple tasks for a given environment dynamics, there are two important challenges. First, we need to abstract the concept of safety that applies broadly to that environment independent of the specific task being executed. Second, we need a mechanism for the abstracted notion of safety to modulate the actions of agents executing different policies to minimize their side-effects. In this work, we propose Safety Aware Reinforcement Learning (SARL) - a framework where a virtual safe agent modulates the actions of a main reward-based agent to minimize side effects. The safe agent learns a task-independent notion of safety for a given environment. The main agent is then trained with a regularization loss given by the distance between the native action probabilities of the two agents. Since the safe agent effectively abstracts a task-independent notion of safety via its action probabilities, it can be ported to modulate multiple policies solving different tasks within the given environment without further training. We contrast this with solutions that rely on task-specific regularization metrics and test our framework on the SafeLife Suite, based on Conway's Game of Life, comprising a number of complex tasks in dynamic environments. We show that our solution is able to match the performance of solutions that rely on task-specific side-effect penalties on both the primary and safety objectives while additionally providing the benefit of generalizability and portability.
****
Optimal Conversion of Conventional Artificial Neural Networks to Spiking Neural Networks
Shikuang Deng,Shi Gu
Spiking neural networks (SNNs) are biology-inspired artificial neural networks (ANNs) that comprise of spiking neurons to process asynchronous discrete signals. While more efficient in power consumption and inference speed on the neuromorphic hardware, SNNs are usually difficult to train directly from scratch with spikes due to the discreteness. As an alternative, many efforts have been devoted to converting conventional ANNs into SNNs by copying the weights from ANNs and adjusting the spiking threshold potential of neurons in SNNs. Researchers have designed new SNN architectures and conversion algorithms to diminish the conversion error. However, an effective conversion should address the difference between the SNN and ANN architectures with an efficient approximation of the loss function, which is missing in the field. In this work, we analyze the conversion error by recursive reduction to layer-wise summation and propose a novel strategic pipeline that transfers the weights to the target SNN by combining threshold balance and soft-reset mechanisms. This pipeline enables almost no accuracy loss between the converted SNNs and conventional ANNs with only $\sim1/10$ of the typical SNN simulation time. Our method is promising to get implanted onto embedded platforms with better support of SNNs with limited energy and memory. Codes are available at https://github.com/Jackn0/snn_optimal_conversion_pipeline.
****
Efficient Continual Learning with Modular Networks and Task-Driven Priors
Tom Veniat,Ludovic Denoyer,MarcAurelio Ranzato

Existing literature in Continual Learning (CL) has focused on overcoming catastr
ophic forgetting, the inability of the learner to recall how to perform tasks ob
served in the past.
There are however other desirable properties of a CL system, such as the ability
 to transfer knowledge from previous tasks and to scale memory and compute sub-l
inearly with the number of tasks. Since most current benchmarks focus only on fo
rgetting using short streams of tasks, we first propose a new suite of benchmark
s to probe CL algorithms across these new axes.
Finally, we introduce a new modular architecture, whose modules represent atomic
 skills that can be composed to perform a certain task. Learning a task reduces
to figuring out which past modules to re-use, and which new modules to instantia
te to solve the current task. Our learning algorithm leverages a task-driven pri
or over the exponential search space of all possible ways to combine modules, en
abling efficient learning on long streams of tasks.
Our experiments show that this modular architecture and learning algorithm perfo
rm competitively on widely used CL benchmarks while yielding superior performanc
e on the more challenging benchmarks we introduce in this work. The Benchmark is
 publicly available at https://github.com/facebookresearch/CTrLBenchmark.
****

Neighbor Class Consistency on Unsupervised Domain Adaptation
Chang Liu,Kai Li,Yun Fu
Unsupervised domain adaptation (UDA) is to make predictions for unlabeled data i
n a target domain with labeled data from source domain available. Recent advance
s exploit entropy minimization and self-training to align the feature of two dom
ains. However, as decision boundary is largely biased towards source data, class
-wise pseudo labels generated by target predictions are usually very noisy, and
trusting those noisy supervisions might potentially deteriorate the intrinsic ta
rget discriminative feature. Motivated by agglomerative clustering which assumes
 that features in the near neighborhood should be clustered together, we observe
 that target features from source pre-trained model are highly intrinsic discrim
inative and have a high probability of sharing the same label with their neighbo
rs. Based on those observations, we propose a simple but effective method to imp
ose Neighbor Class Consistency on target features to preserve and further streng
then the intrinsic discriminative nature of target data while regularizing the u
nified classifier less biased towards source data. We also introduce an entropy-
based weighting scheme to help our framework more robust to the potential noisy
neighbor supervision. We conduct ablation studies and extensive experiments on t
hree UDA image classification benchmarks. Our method outperforms all existing UD
A state-of-the-art.
****

On the Universality of Rotation Equivariant Point Cloud Networks
Nadav Dym,Haggai Maron
Learning functions on point clouds has applications in many fields, including co
mputer vision, computer graphics, physics, and chemistry. Recently, there has be
en a growing interest in neural architectures that are invariant or equivariant
to all three shape-preserving transformations of point clouds: translation, rota
tion, and permutation. In this paper, we present a first study of the approximat
ion power of these architectures. We first derive two sufficient conditions for
an equivariant architecture to have the universal approximation property, based
on a novel characterization of the space of equivariant polynomials. We then use
 these conditions to show that two recently suggested models, Tensor field Netwo
rks and SE3-Transformers, are universal, and for devising two other novel univer
sal architectures.
****

Neural Learning of One-of-Many Solutions for Combinatorial Problems in Structure
d Output Spaces
Yatin Nandwani,Deepanshu Jindal,Mausam .,Parag Singla
Recent research has proposed neural architectures for solving combinatorial prob
lems in structured output spaces. In many such problems, there may exist multipl
e solutions for a given input, e.g. a partially filled Sudoku puzzle may have ma

ny completions satisfying all constraints. Further, we are often interested in f inding any "one" of the possible solutions, without any preference between them. Existing approaches completely ignore this solution multiplicity. In this paper , we argue that being oblivious to the presence of multiple solutions can severe ly hamper their training ability. Our contribution is two-fold. First, we formal ly define the task of learning one-of-many solutions for combinatorial problems in structured output spaces, which is applicable for solving several problems of interest such as N-Queens, and Sudoku. Second, we present a generic learning fr amework that adapts an existing prediction network for a combinatorial problem t o handle solution multiplicity. Our framework uses a selection module, whose goa l is to dynamically determine, for every input, the solution that is most effect ive for training the network parameters in any given learning iteration. We prop ose an RL based approach to jointly train the selection module with the predicti on network. Experiments on three different domains, and using two different pred iction networks, demonstrate that our framework significantly improves the accu racy in our setting, obtaining up to 21 pt gain over the baselines.

****

Understanding Classifiers with Generative Models
Laëtitia Shao,Yang Song,Stefano Ermon
Although deep neural networks are effective on supervised learning tasks, they h ave been shown to be brittle. They are prone to overfitting on their training di stribution and are easily fooled by small adversarial perturbations. In this pap er, we leverage generative models to identify and characterize instances where c lassifiers fail to generalize.
We propose a generative model of the features extracted by a classifier, and sho w using rigorous hypothesis testing that errors tend to occur when features are assigned low-probability by our model. From this observation, we develop a detec tion criteria that we test against different sources of classification mistakes: mistakes made on the test set due to poor model generalization, adversarial sam ples and out-of-distribution samples. Our approach is agnostic to class labels f rom the training set which makes it applicable to models trained in a semi-super vised way.
****

LambdaNetworks: Modeling long-range Interactions without Attention
Irwan Bello
We present lambda layers -- an alternative framework to self-attention -- for ca pturing long-range interactions between an input and structured contextual infor mation (e.g. a pixel surrounded by other pixels). Lambda layers capture such int eractions by transforming available contexts into linear functions, termed lambd as, and applying these linear functions to each input separately. Similar to lin ear attention, lambda layers bypass expensive attention maps, but in contrast, t hey model both content and position-based interactions which enables their appli cation to large structured inputs such as images. The resulting neural network a rchitectures, LambdaNetworks, significantly outperform their convolutional and a ttentional counterparts on ImageNet classification, COCO object detection and in stance segmentation, while being more computationally efficient. Additionally, w e design LambdaResNets, a family of hybrid architectures across different scales , that considerably improves the speed-accuracy tradeoff of image classification models. LambdaResNets reach excellent accuracies on ImageNet while being 3.2 - 4.4x faster than the popular EfficientNets on modern machine learning accelerato rs. In large-scale semi-supervised training with an additional 130M pseudo-label ed images, LambdaResNets achieve up to 86.7% ImageNet accuracy while being 9.5x faster than EfficientNet NoisyStudent and 9x faster than a Vision Transformer wi th comparable accuracies.
****

Understanding and Mitigating Accuracy Disparity in Regression
Jianfeng Chi,Han Zhao,Geoff Gordon,Yuan Tian
With the widespread deployment of large-scale prediction systems in high-stakes domains, e.g., face recognition, criminal justice, etc., disparity on prediction

accuracy between different demographic subgroups has called for fundamental understanding on the source of such disparity and algorithmic intervention to mitigate it. In this paper, we study the accuracy disparity problem in regression. To begin with, we first propose an error decomposition theorem, which decomposes the accuracy disparity into the distance between label populations and the distance between conditional representations, to help explain why such accuracy disparity appears in practice. Motivated by this error decomposition and the general idea of distribution alignment with statistical distances, we then propose an algorithm to reduce this disparity, and analyze its game-theoretic optima of the proposed objective function. We conduct experiments on four real-world datasets. The experimental results suggest that our proposed algorithms can effectively mitigate accuracy disparity while maintaining the predictive power of the regression models.
****

Removing Dimensional Restrictions on Complex/Hyper-complex Convolutions
Chase John Gaudet,Anthony S. Maida
It has been shown that the core reasons that complex and hypercomplex valued neural networks offer improvements over their real-valued counterparts is the fact that aspects of their algebra forces treating multi-dimensional data as a single entity (forced local relationship encoding) with an added benefit of reducing parameter count via weight sharing. However, both are constrained to a set number of dimensions, two for complex and four for quaternions. These observations motivate us to introduce novel vector map convolutions which capture both of these properties provided by complex/hypercomplex convolutions, while dropping the unnatural dimensionality constraints their algebra imposes. This is achieved by introducing a system that mimics the unique linear combination of input dimensions via the Hamilton product using a permutation function, as well as batch normalization and weight initialization for the system. We perform three experiments using three different network architectures to show that these novel vector map convolutions seem to capture all the benefits of complex and hyper-complex networks, such as their ability to capture internal latent relations, while avoiding the dimensionality restriction.
****

SHADOWCAST: Controllable Graph Generation with Explainability
Wesley Joon-Wie Tann,Ee-Chien Chang,Bryan Hooi
We introduce the problem of explaining graph generation, formulated as controlling the generative process to produce desired graphs with explainable structures. By directing this generative process, we can explain the observed outcomes. We propose SHADOWCAST, a controllable generative model capable of mimicking networks and directing the generation, as an approach to this novel problem. The proposed model is based on a conditional generative adversarial network for graph data. We design it with the capability to control the conditions using a simple and transparent Markov model. Comprehensive experiments on three real-world network datasets demonstrate our model's competitive performance in the graph generation task. Furthermore, we control SHADOWCAST to generate graphs of different structures to show its effective controllability and explainability. As the first work to pose the problem of explaining generated graphs by controlling the generation, SHADOWCAST paves the way for future research in this exciting area.
****

Continual learning with neural activation importance
Sohee Kim,Seungkyu Lee
Continual learning is a concept of online learning along with multiple sequential tasks. One of the critical barriers of continual learning is that a network should learn a new task keeping the knowledge of old tasks without access to any data of the old tasks. In this paper, we propose a neuron importance based regularization method for stable continual learning.
We propose a comprehensive experimental evaluation framework on existing benchmark data sets to evaluate not just the accuracy of a certain order of continual learning performance also the robustness of the accuracy along with the changes in the order of tasks.

****
It Is Likely That Your Loss Should be a Likelihood
Mark Hamilton,Evan Shelhamer,William T. Freeman
Many common loss functions such as mean-squared-error, cross-entropy, and reconstruction loss are unnecessarily rigid. Under a probabilistic interpretation, these common losses correspond to distributions with fixed shapes and scales. We instead argue for optimizing full likelihoods that include parameters like the normal variance and softmax temperature. Joint optimization of these ``likelihood parameters'' with model parameters can adaptively tune the scales and shapes of losses in addition to the strength of regularization. We explore and systematically evaluate how to parameterize and apply likelihood parameters for robust modeling, outlier-detection, and re-calibration. Additionally, we propose adaptively tuning L_2 and L_1 weights by fitting the scale parameters of normal and Laplace priors and introduce more flexible element-wise regularizers.
****
Reusing Preprocessing Data as Auxiliary Supervision in Conversational Analysis
Joshua Yee Kim,Kalina Yacef
Conversational analysis systems are trained using noisy human labels and often require heavy preprocessing during multi-modal feature extraction. Using noisy labels in single-task learning increases the risk of over-fitting. However, auxiliary tasks could improve the performance of the primary task learning. This approach is known as Primary Multi-Task Learning (MTL). A challenge of MTL is the selection of beneficial auxiliary tasks that avoid negative transfer. In this paper, we explore how the preprocessed data used for feature engineering can be re-used as auxiliary tasks in Primary MTL, thereby promoting the productive use of data in the form of auxiliary supervision learning. Our main contributions are: (1) the identification of sixteen beneficially auxiliary tasks, (2) the method of distributing learning capacity between the primary and auxiliary tasks, and (3) the relative supervision hierarchy between the primary and auxiliary tasks. Extensive experiments on IEMOCAP and SEMAINE data validate the improvements over single-task approaches, and suggest that it may generalize across multiple primary tasks.
****
GAN2GAN: Generative Noise Learning for Blind Denoising with Single Noisy Images
Sungmin Cha,Taeeon Park,Byeongjoon Kim,Jongduk Baek,Taesup Moon
We tackle a challenging blind image denoising problem, in which only single distinct noisy images are available for training a denoiser, and no information about noise is known, except for it being zero-mean, additive, and independent of the clean image. In such a setting, which often occurs in practice, it is not possible to train a denoiser with the standard discriminative training or with the recently developed Noise2Noise (N2N) training; the former requires the underlying clean image for the given noisy image, and the latter requires two independently realized noisy image pair for a clean image. To that end, we propose GAN2GAN (Generated-Artificial-Noise to Generated-Artificial-Noise) method that first learns a generative model that can 1) simulate the noise in the given noisy images and 2) generate a rough, noisy estimates of the clean images, then 3) iteratively trains a denoiser with subsequently synthesized noisy image pairs (as in N2N), obtained from the generative model. In results, we show the denoiser trained with our GAN2GAN achieves an impressive denoising performance on both synthetic and real-world datasets for the blind denoising setting; it almost approaches the performance of the standard discriminatively-trained or N2N-trained models that have more information than ours, and it significantly outperforms the recent baseline for the same setting, \textit{e.g.}, Noise2Void, and a more conventional yet strong one, BM3D. The official code of our method is available at https://github.com/csm9493/GAN2GAN.
****
CPR: Classifier-Projection Regularization for Continual Learning
Sungmin Cha,Hsiang Hsu,Taebaek Hwang,Flavio Calmon,Taesup Moon
We propose a general, yet simple patch that can be applied to existing regularization-based continual learning methods called classifier-projection regularizati

on (CPR). Inspired by both recent results on neural networks with wide local min
ima and information theory, CPR adds an additional regularization term that maxi
mizes the entropy of a classifier's output probability. We demonstrate that this
 additional term can be interpreted as a projection of the conditional probabili
ty given by a classifier's output to the uniform distribution. By applying the P
ythagorean theorem for KL divergence, we then prove that this projection may (in
 theory) improve the performance of continual learning methods. In our extensive
 experimental results, we apply CPR to several state-of-the-art regularization-b
ased continual learning methods and benchmark performance on popular image recog
nition datasets. Our results demonstrate that CPR indeed promotes a wide local m
inima and significantly improves both accuracy and plasticity while simultaneous
ly mitigating the catastrophic forgetting of baseline continual learning methods
. The codes and scripts for this work are available at https://github.com/csm949
3/CPR_CL.
****

Learning Irreducible Representations of Noncommutative Lie Groups
Noah Shutty,Casimir Wierzynski
Recent work has constructed neural networks that are equivariant to continuous s
ymmetry groups such as 2D and 3D rotations. This is accomplished using explicit
group representations to derive the equivariant kernels and nonlinearities. We p
resent two contributions motivated by frontier applications of equivariance beyo
nd rotations and translations. First, we relax the requirement for explicit Lie
group representations, presenting a novel algorithm that finds irreducible repre
sentations of noncommutative Lie groups given only the structure constants of th
e associated Lie algebra. Second, we demonstrate that Lorentz-equivariance is a
useful prior for object-tracking tasks and construct the first object-tracking m
odel equivariant to the Poincaré group.
****

Adaptive Hierarchical Hyper-gradient Descent
RENLONG JIE,Junbin Gao,Andrey Vasnev,Minh-Ngoc Tran
Adaptive learning rates can lead to faster convergence and better final performa
nce
for deep learning models. There are several widely known human-designed adap-
tive optimizers such as Adam and RMSProp, gradient based adaptive methods
such as hyper-descent and L4, and meta learning approaches including learning
to learn. However, the issue of balancing adaptiveness and over-parameterization
is still a topic to be addressed. In this study, we investigate different lev
els of
learning rate adaptation based on the framework of hyper-gradient descent, and
further propose a method that adaptively learns the model parameters for combin-
ing different levels of adaptations. Meanwhile, we show the relationship between
adding regularization on over-parameterized learning rates and building combi-
nations of different levels of adaptive learning rates. The experiments on seve
ral
network architectures including feed-forward networks, LeNet-5 and ResNet-18/34
show that the proposed multi-level adaptive approach can outperform bas
eline
adaptive methods in a variety circumstances with statistical significance.
****

Frequency Decomposition in Neural Processes
Jens Petersen,Paul F Jaeger,Gregor Koehler,David Zimmerer,Fabian Isensee,Klaus M
aier-Hein
Neural Processes are a powerful tool for learning representations of function sp
aces purely from examples, in a way that allows them to perform predictions at t
est time conditioned on so-called context observations. The learned representati
ons are finite-dimensional, while function spaces are infinite-dimensional, and
so far it has been unclear how these representations are learned and what kinds
of functions can be represented. We show that deterministic Neural Processes imp
licitly perform a decomposition of the training signals into different frequency
 components, similar to a Fourier transform. In this context, we derive a theore

tical upper bound on the maximum frequency Neural Processes can reproduce, depen
ding on their representation size. This bound is confirmed empirically. Finally,
 we show that Neural Processes can be trained to only represent a subset of poss
ible frequencies and suppress others, which makes them programmable band-pass or
 band-stop filters.
****
Contrastive Divergence Learning is a Time Reversal Adversarial Game
Omer Yair,Tomer Michaeli
Contrastive divergence (CD) learning is a classical method for fitting unnormali
zed statistical models to data samples. Despite its wide-spread use, the converg
ence properties of this algorithm are still not well understood. The main source
 of difficulty is an unjustified approximation which has been used to derive the
 gradient of the loss. In this paper, we present an alternative derivation of CD
 that does not require any approximation and sheds new light on the objective th
at is actually being optimized by the algorithm. Specifically, we show that CD i
s an adversarial learning procedure, where a discriminator attempts to classify
whether a Markov chain generated from the model has been time-reversed. Thus, al
though predating generative adversarial networks (GANs) by more than a decade, C
D is, in fact, closely related to these techniques. Our derivation settles well
with previous observations, which have concluded that CD's update steps cannot b
e expressed as the gradients of any fixed objective function. In addition, as a
byproduct, our derivation reveals a simple correction that can be used as an alt
ernative to Metropolis-Hastings rejection, which is required when the underlying
 Markov chain is inexact (e.g., when using Langevin dynamics with a large step).
****
Learning Curves for Analysis of Deep Networks
Derek Hoiem,Tanmay Gupta,Zhizhong Li,Michal M Shlapentokh-Rothman
A learning curve models a classifier’s test error as a function of the number of
 training samples. Prior works show that learning curves can be used to select m
odel parameters and extrapolate performance. We investigate how to use learning
curves to analyze the impact of design choices, such as pre-training, architectu
re, and data augmentation. We propose a method to robustly estimate learning cur
ves, abstract their parameters into error and data-reliance, and evaluate the ef
fectiveness of different parameterizations. We also provide several interesting
observations based on learning curves for a variety of image classification mode
ls.
****
Directional graph networks
Dominique Beaini,Saro Passaro,Vincent Letourneau,William L. Hamilton,Gabriele Co
rso,Pietro Liò
In order to overcome the expressive limitations of graph neural networks (GNNs),
 we propose the first method that exploits vector flows over graphs to develop g
lobally consistent directional and asymmetric aggregation functions.
We show that our directional graph networks (DGNs) generalize convolutional neur
al networks (CNNs) when applied on a grid. Whereas recent theoretical works focu
s on understanding local neighbourhoods, local structures and local isomorphism
with no global information flow, our novel theoretical framework allows directio
nal convolutional kernels in any graph.
First, by defining a vector field in the graph, we develop a method of applying
directional derivatives and smoothing by projecting node-specific messages into
the field.
Then we propose the use of the Laplacian eigenvectors as such vector field, and
we show that the method generalizes CNNs on an n-dimensional grid, and is prov
ably more discriminative than standard GNNs regarding the Weisfeiler-Lehman 1-WL
 test.
Finally, we bring the power of CNN data augmentation to graphs by providing a me
ans of doing reflection, rotation and distortion on the underlying directional f
ield. We evaluate our method on different standard benchmarks and see a relative
 error reduction of 8% on the CIFAR10 graph dataset and 11% to 32% on the molecu
lar ZINC dataset. An important outcome of this work is that it enables to transl

ate any physical or biological problems with intrinsic directional axes into a g raph network formalism with an embedded directional field.
****
Contrastive estimation reveals topic posterior information to linear models
Christopher Tosh,Akshay Krishnamurthy,Daniel Hsu
Contrastive learning is an approach to representation learning that utilizes nat urally occurring similar and dissimilar pairs of data points to find useful embe ddings of data. In the context of document classification under topic modeling a ssumptions, we prove that contrastive learning is capable of recovering a repres entation of documents that reveals their underlying topic posterior information to linear models. We apply this procedure in a semi-supervised setup and demonst rate empirically that linear classifiers with these representations perform well in document classification tasks with very few training examples.
****
Contrasting distinct structured views to learn sentence embeddings
Antoine Simoulin,Benoit Crabbé
We propose a self-supervised method that builds sentence embeddings from the co mbination of diverse explicit syntactic structures of a sentence. We assume stru cture is crucial to build consistent representations as we expect sentence meani ng to be a function from both syntax and semantic aspects. In this perspective, we hypothesize that some linguistic representations might be better adapted give n the considered task or sentence. We, therefore, propose to jointly learn indiv idual representation functions for different syntactic frameworks. Again, by hyp othesis, all such functions should encode similar semantic information different ly and consequently, be complementary for building better sentential semantic em beddings. To assess such hypothesis, we propose an original contrastive multi-vi ew framework that induces an explicit interaction between models during the trai ning phase. We make experiments combining various structures such as dependency, constituency, or sequential schemes. We evaluate our method on standard sentenc e embedding benchmarks. Our results outperform comparable methods on several tas ks.

****
Byzantine-Robust Learning on Heterogeneous Datasets via Resampling
Lie He,Sai Praneeth Karimireddy,Martin Jaggi
In Byzantine-robust distributed optimization, 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 ar bitrary messages to the server. While this problem has received significant atte ntion recently, most current defenses assume that the workers have identical dat a distribution. For realistic cases when the data across workers are heterogeneo us (non-iid), we design new attacks that circumvent these defenses leading to si gnificant loss of performance. We then propose a universal resampling scheme tha t addresses data heterogeneity at a negligible computational cost. We theoretica lly and experimentally validate our approach, showing that combining resampling with existing robust algorithms is effective against challenging attacks.

****
Federated Generalized Bayesian Learning via Distributed Stein Variational Gradi ent Descent
Rahif Kassab,Osvaldo Simeone
This paper introduces Distributed Stein Variational Gradient Descent (DSVGD), a non-parametric generalized Bayesian inference framework for federated learning. DSVGD maintains a number of non-random and interacting particles at a central se rver to represent the current iterate of the model global posterior. The particl es are iteratively downloaded and updated by one of the agents with the end goal of minimizing the global free energy. By varying the number of particles, DSVGD enables a flexible trade-off between per-iteration communication load and numbe

r of communication rounds. DSVGD is shown to compare favorably to benchmark frequentist and Bayesian federated learning strategies, also scheduling a single device per iteration, in terms of accuracy and scalability with respect to the number of agents, while also providing well-calibrated, and hence trustworthy, predictions.
****

On the Dynamics of Training Attention Models
Haoye Lu,Yongyi Mao,Amiya Nayak
The attention mechanism has been widely used in deep neural networks as a model component. By now, it has become a critical building block in many state-of-the-art natural language models. Despite its great success established empirically, the working mechanism of attention has not been investigated at a sufficient theoretical depth to date. In this paper, we set up a simple text classification task and study the dynamics of training a simple attention-based classification model using gradient descent. In this setting, we show that, for the discriminative words that the model should attend to, a persisting identity exists relating its embedding and the inner product of its key and the query. This allows us to prove that training must converge to attending to the discriminative words when the attention output is classified by a linear classifier. Experiments are performed, which validate our theoretical analysis and provide further insights.
****

Which Mutual-Information Representation Learning Objectives are Sufficient for Control?
Kate Rakelly,Abhishek Gupta,Carlos Florensa,Sergey Levine
Mutual information maximization provides an appealing formalism for learning representations of data. In the context of reinforcement learning, such representations can accelerate learning by discarding irrelevant and redundant information, while retaining the information necessary for control. Much of the prior work on these methods has addressed the practical difficulties of estimating mutual information from samples of high-dimensional observations, while comparatively less is understood about \emph{which} mutual information objectives are sufficient for RL from a theoretical perspective. In this paper we identify conditions under which representations that maximize specific mutual-information objectives are theoretically sufficient for learning and representing the optimal policy. Somewhat surprisingly, we find that several popular objectives can yield insufficient representations given mild and common assumptions on the structure of the MDP. We corroborate our theoretical results with deep RL experiments on a simulated game environment with visual observations.
****

Model-Based Offline Planning
Arthur Argenson,Gabriel Dulac-Arnold
Offline learning is a key part of making reinforcement learning (RL) useable in real systems. Offline RL looks at scenarios where there is data from a system's operation, but no direct access to the system when learning a policy. Recent work on training RL policies from offline data has shown results both with model-free policies learned directly from the data, or with planning on top of learnt models of the data. Model-free policies tend to be more performant, but are more opaque, harder to command externally, and less easy to integrate into larger systems. We propose an offline learner that generates a model that can be used to control the system directly through planning. This allows us to have easily controllable policies directly from data, without ever interacting with the system. We show the performance of our algorithm, Model-Based Offline Planning (MBOP) on a series of robotics-inspired tasks, and demonstrate its ability leverage planning to respect environmental constraints. We are able to find near-optimal polices for certain simulated systems from as little as 50 seconds of real-time system interaction, and create zero-shot goal-conditioned policies on a series of environments.
****

Neurosymbolic Deep Generative Models for Sequence Data with Relational Constraints

Halley Young,Maxwell Du,Osbert Bastani
Recently, there has been significant progress designing deep generative models t hat generate realistic sequence data such as text or music. Nevertheless, it rem ains difficult to incorporate high-level structure to guide the generative proce ss. We propose a novel approach for incorporating structure in the form of relat ional 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. To train model (i), we propose a novel program synthesis algorithm that infers the relational constrain ts present in the training data, and then train the models based on the resultin g relational constraints. In our experiments, we show that our approach signifi cantly improves over state-of-the-art approaches in terms of capturing high-leve l structure in the data, while performing comparably or better in terms of low-l evel structure.
****
Modelling Hierarchical Structure between Dialogue Policy and Natural Language Ge nerator with Option Framework for Task-oriented Dialogue System
Jianhong Wang,Yuan Zhang,Tae-Kyun Kim,Yunjie Gu
Designing task-oriented dialogue systems is a challenging research topic, since it needs not only to generate utterances fulfilling user requests but also to gu arantee the comprehensibility. Many previous works trained end-to-end (E2E) mode ls with supervised learning (SL), however, the bias in annotated system utteranc es remains as a bottleneck. Reinforcement learning (RL) deals with the problem t hrough using non-differentiable evaluation metrics (e.g., the success rate) as r ewards. Nonetheless, existing works with RL showed that the comprehensibility of generated system utterances could be corrupted when improving the performance o n fulfilling user requests. In our work, we (1) propose modelling the hierarchic al structure between dialogue policy and natural language generator (NLG) with t he option framework, called HDNO, where the latent dialogue act is applied to av oid designing specific dialogue act representations; (2) train HDNO via hierarch ical reinforcement learning (HRL), as well as suggest the asynchronous updates b etween dialogue policy and NLG during training to theoretically guarantee their convergence to a local maximizer; and (3) propose using a discriminator modelled with language models as an additional reward to further improve the comprehensi bility. We test HDNO on MultiWoz 2.0 and MultiWoz 2.1, the datasets on multi-dom ain dialogues, in comparison with word-level E2E model trained with RL, LaRL and HDSA, showing improvements on the performance evaluated by automatic evaluation metrics and human evaluation. Finally, we demonstrate the semantic meanings of latent dialogue acts to show the explanability for HDNO.
****
Towards Understanding the Cause of Error in Few-Shot Learning
Liang Song,Jinlu Liu,Yongqiang Qin
Few-Shot Learning (FSL) is a challenging task of recognizing novel classes from scarce labeled samples. Many existing researches focus on learning good represen tations that generalize well to new categories. However, given low-data regime, the restricting factors of performance on novel classes has not been well studie d. In this paper, our objective is to understand the cause of error in few-shot classification, as well as exploring the upper limit of error rate. We first int roduce and derive a theoretical upper bound of error rate which is constrained t o 1) linear separability in the learned embedding space and 2) discrepancy of ta sk-specific and task-independent classifier. Quantitative experiment is conducte d and results show that the error in FSL is dominantly caused by classifier disc repancy. We further propose a simple method to confirm our theoretical analysis and observation. The method adds a constraint to reduce classifier discrepancy s o as to lower the upper bound of error rate. Experiments on three benchmarks wit h different base learners verify the effectiveness of our method. It shows that decreasing classifier discrepancy can consistently achieve improvements in most cases.
****

The Heavy-Tail Phenomenon in SGD
Mert Gurbuzbalaban,Umut Simsekli,Lingjiong Zhu
In recent years, various notions of capacity and complexity have been proposed f
or characterizing the generalization properties of stochastic gradient descent (
SGD) in deep learning. Some of the popular notions that correlate well with the
performance on unseen data are (i) the 'flatness' of the local minimum found by
SGD, which is related to the eigenvalues of the Hessian, (ii) the ratio of the s
tepsize η to the batch size b, which essentially controls the magnitude o
f the stochastic gradient noise, and (iii) the 'tail-index', which measures the
heaviness of the tails of the network weights at convergence. In this paper, we
argue that these three seemingly unrelated perspectives for generalization are d
eeply linked to each other. We claim that depending on the structure of the Hess
ian of the loss at the minimum, and the choices of the algorithm parameters $\et
a$ and b, the SGD iterates will converge to a \emph{heavy-tailed} stationary d
istribution. We rigorously prove this claim in the setting of quadratic optimiza
tion: we show that even in a simple linear regression problem with independent a
nd identically distributed Gaussian data, the iterates can be heavy-tailed with
infinite variance. We further characterize the behavior of the tails with respec
t to algorithm parameters, the dimension, and the curvature. We then translate o
ur results into insights about the behavior of SGD in deep learning. We finally
support our theory with experiments conducted on both synthetic data and fully c
onnected neural networks.
****
Neural Synthesis of Binaural Speech From Mono Audio
Alexander Richard,Dejan Markovic,Israel D. Gebru,Steven Krenn,Gladstone Alexande
r Butler,Fernando Torre,Yaser Sheikh
We present a neural rendering approach for binaural sound synthesis that can pro
duce realistic and spatially accurate binaural sound in realtime. The network ta
kes, as input, a single-channel audio source and synthesizes, as output, two-cha
nnel binaural sound, conditioned on the relative position and orientation of the
 listener with respect to the source. We investigate deficiencies of the l2-loss
 on raw waveforms in a theoretical analysis and introduce an improved loss that
overcomes these limitations. In an empirical evaluation, we establish that our a
pproach is the first to generate spatially accurate waveform outputs (as measure
d by real recordings) and outperforms existing approaches by a considerable marg
in, both quantitatively and in a perceptual study. Dataset and code are availabl
e online.
****
Distilling Knowledge from Reader to Retriever for Question Answering
Gautier Izacard,Edouard Grave
The task of information retrieval is an important component of many natural lang
uage processing systems, such as open domain question answering. While tradition
al methods were based on hand-crafted features, continuous representations based
 on neural networks recently obtained competitive results. A challenge of using
such methods is to obtain supervised data to train the retriever model, correspo
nding to pairs of query and support documents. In this paper, we propose a techn
ique to learn retriever models for downstream tasks, inspired by knowledge disti
llation, and which does not require annotated pairs of query and documents. Our
approach leverages attention scores of a reader model, used to solve the task ba
sed on retrieved documents, to obtain synthetic labels for the retriever. We eva
luate our method on question answering, obtaining state-of-the-art results.
****
How to Train Your Super-Net: An Analysis of Training Heuristics in Weight-Sharin
g NAS
Kaicheng Yu,Rene Ranftl,Mathieu Salzmann
Weight sharing promises to make neural architecture search (NAS) tractable even
on commodity hardware. Existing methods in this space rely on a diverse set of h
euristics to design and train the shared-weight backbone network, a.k.a. the sup
er-net. Since heuristics substantially vary across different methods and have no
t been carefully studied, it is unclear to which extent they impact super-net tr

aining and hence the weight-sharing NAS algorithms. In this paper, we disentangle super-net training from the search algorithm, isolate 14 frequently-used training heuristics, and evaluate them over three benchmark search spaces. Our analysis uncovers that several commonly-used heuristics negatively impact the correlation between super-net and stand-alone performance, whereas simple, but often overlooked factors, such as proper hyper-parameter settings, are key to achieve strong performance. Equipped with this knowledge, we show that simple random search achieves competitive performance to complex state-of-the-art NAS algorithms when the super-net is properly trained.

****

Revisiting Loss Modelling for Unstructured Pruning
César Laurent,Camille Ballas,Thomas George,Pascal Vincent,Nicolas Ballas
By removing parameters from deep neural networks, unstructured pruning methods aim at cutting down memory footprint and computational cost, while maintaining prediction accuracy. In order to tackle this otherwise intractable problem, many of these methods model the loss landscape using first or second order Taylor expansions to identify which parameters can be discarded. We revisit loss modelling for unstructured pruning: we show the importance of ensuring locality of the pruning steps, and systematically compare first and second order Taylor expansions. Finally, we show that better preserving the original network function does not necessarily transfer to better performing networks after fine-tuning, suggesting that only considering the impact of pruning on the loss might not be a sufficient objective to design good pruning criteria.
****

Continual learning using hash-routed convolutional neural networks
Ahmad Berjaoui
Continual learning could shift the machine learning paradigm from data centric to model centric. A continual learning model needs to scale efficiently to handle semantically different datasets, while avoiding unnecessary growth. We introduce hash-routed convolutional neural networks: a group of convolutional units where data flows dynamically. Feature maps are compared using feature hashing and similar data is routed to the same units. A hash-routed network provides excellent plasticity thanks to its routed nature, while generating stable features through the use of orthogonal feature hashing. Each unit evolves separately and new units can be added (to be used only when necessary). Hash-routed networks achieve excellent performance across a variety of typical continual learning benchmarks without storing raw data and train using only gradient descent. Besides providing a continual learning framework for supervised tasks with encouraging results, our model can be used for unsupervised or reinforcement learning.
****

UneVEn: Universal Value Exploration for Multi-Agent Reinforcement Learning
Tarun Gupta,Anuj Mahajan,Bei Peng,Wendelin Boehmer,Shimon Whiteson
This paper focuses on cooperative value-based multi-agent reinforcement learning (MARL) in the paradigm of centralized training with decentralized execution (CTDE). Current state-of-the-art value-based MARL methods leverage CTDE to learn a centralized joint-action value function as a monotonic mixing of each agent's utility function, which enables easy decentralization. However, this monotonic restriction leads to inefficient exploration in tasks with nonmonotonic returns due to suboptimal approximations of the values of joint actions. To address this, we present a novel MARL approach called Universal Value Exploration (UneVEn), which uses universal successor features (USFs) to learn policies of tasks related to the target task, but with simpler reward functions in a sample efficient manner. UneVEn uses novel action-selection schemes between randomly sampled related tasks during exploration, which enables the monotonic joint-action value function of the target task to place more importance on useful joint actions. Empirical results on a challenging cooperative predator-prey task requiring significant coordination amongst agents show that UneVEn significantly outperforms state-of-the-art baselines.
****

Variance Based Sample Weighting for Supervised Deep Learning
Paul Novello,Gaël Poëtte,David Lugato,Pietro Congedo
In the context of supervised learning of a function by a Neural Network (NN), we claim and empirically justify that a NN yields better results when the distribution of the data set focuses on regions where the function to learn is steeper. We first traduce this assumption in a mathematically workable way using Taylor expansion. Then, theoretical derivations allow to construct a methodology that we call Variance Based Samples Weighting (VBSW). VBSW uses local variance of the labels to weight the training points. This methodology is general, scalable, cost effective, and significantly increases the performances of a large class of models for various classification and regression tasks on image, text and multivariate data. We highlight its benefits with experiments involving NNs from shallow linear NN to ResNet or Bert.
****
Learning from Demonstrations with Energy based Generative Adversarial Imitation Learning
Kaifeng Zhang
Traditional reinforcement learning methods usually deal with the tasks with explicit reward signals. However, for vast majority of cases, the environment wouldn't feedback a reward signal immediately. It turns out to be a bottleneck for modern reinforcement learning approaches to be applied into more realistic scenarios. Recently, inverse reinforcement learning has made great progress in making full use of the expert demonstrations to recover the reward signal for reinforcement learning. And generative adversarial imitation learning is one promising approach. In this paper, we propose a new architecture for training generative adversarial imitation learning which is so called energy based generative adversarial imitation learning (EB-GAIL). It views the discriminator as an energy function that attributes low energies to the regions near the expert demonstrations and high energies to other regions. Therefore, a generator can be seen as a reinforcement learning procedure to sample trajectories with minimal energies (cost), while the discriminator is trained to assign high energies to these generated trajectories. In detail, EB-GAIL uses an auto-encoder architecture in place of the discriminator, with the energy being the reconstruction error. Theoretical analysis shows our EB-GAIL could match the occupancy measure with expert policy during the training process. Meanwhile, the experiments depict that EB-GAIL outperforms other SoTA methods while the training process for EB-GAIL can be more stable.
****
Efficient Certified Defenses Against Patch Attacks on Image Classifiers
Jan Hendrik Metzen,Maksym Yatsura
Adversarial patches pose a realistic threat model for physical world attacks on autonomous systems via their perception component. Autonomous systems in safety-critical domains such as automated driving should thus contain a fail-safe fallback component that combines certifiable robustness against patches with efficient inference while maintaining high performance on clean inputs. We propose BagCert, a novel combination of model architecture and certification procedure that allows efficient certification. We derive a loss that enables end-to-end optimization of certified robustness against patches of different sizes and locations. On CIFAR10, BagCert certifies 10.000 examples in 43 seconds on a single GPU and obtains 86% clean and 60% certified accuracy against 5x5 patches.
****
Differentiable Weighted Finite-State Transducers
Awni Hannun,Vineel Pratap,Jacob Kahn,Wei-Ning Hsu
We introduce a framework for automatic differentiation with weighted finite-state transducers (WFSTs) allowing them to be used dynamically at training time. Through the separation of graphs from operations on graphs, this framework enables the exploration of new structured loss functions which in turn eases the encoding of prior knowledge into learning algorithms. We show how the framework can combine pruning and back-off in transition models with various sequence-level loss functions. We also show how to learn over the latent decomposition of phrases into word pieces. Finally, to demonstrate that WFSTs can be used in the interior o

f a deep neural network, we propose a convolutional WFST layer which maps lower-level representations to higher-level representations and can be used as a drop-in replacement for a traditional convolution. We validate these algorithms with experiments in handwriting recognition and speech recognition.
****
Meta Adversarial Training
Jan Hendrik Metzen,Nicole Finnie,Robin Hutmacher
Recently demonstrated physical-world adversarial attacks have exposed vulnerabilities in perception systems that pose severe risks for safety-critical applications such as autonomous driving. These attacks place adversarial artifacts in the physical world that indirectly cause the addition of universal perturbations to inputs of a model that can fool it in a variety of contexts. Adversarial training is the most effective defense against image-dependent adversarial attacks. However, tailoring adversarial training to universal perturbations is computationally expensive since the optimal universal perturbations depend on the model weights which change during training. We propose meta adversarial training (MAT), a novel combination of adversarial training with meta-learning, which overcomes this challenge by meta-learning universal perturbations along with model training. MAT requires little extra computation while continuously adapting a large set of perturbations to the current model. We present results for universal patch and universal perturbation attacks on image classification and traffic-light detection. MAT considerably increases robustness against universal patch attacks compared to prior work.
****
Quantifying Differences in Reward Functions
Adam Gleave,Michael D Dennis,Shane Legg,Stuart Russell,Jan Leike
For many tasks, the reward function is inaccessible to introspection or too complex to be specified procedurally, and must instead be learned from user data. Prior work has evaluated learned reward functions by evaluating policies optimized for the learned reward. However, this method cannot distinguish between the learned reward function failing to reflect user preferences and the policy optimization process failing to optimize the learned reward. Moreover, this method can only tell us about behavior in the evaluation environment, but the reward may incentivize very different behavior in even a slightly different deployment environment. To address these problems, we introduce the Equivalent-Policy Invariant Comparison (EPIC) distance to quantify the difference between two reward functions directly, without a policy optimization step. We prove EPIC is invariant on an equivalence class of reward functions that always induce the same optimal policy. Furthermore, we find EPIC can be efficiently approximated and is more robust than baselines to the choice of coverage distribution. Finally, we show that EPIC distance bounds the regret of optimal policies even under different transition dynamics, and we confirm empirically that it predicts policy training success. Our source code is available at https://github.com/HumanCompatibleAI/evaluating-rewards.
****
Single-Node Attack for Fooling Graph Neural Networks
Ben Finkelshtein,Chaim Baskin,Evgenii Zheltonozhskii,Uri Alon
Graph neural networks (GNNs) have shown broad applicability in a variety of domains.
Some of these domains, such as social networks and product recommendations, are fertile ground for malicious users and behavior.
In this paper, we show that GNNs are vulnerable to the extremely limited scenario of a single-node adversarial example, where the node cannot be picked by the attacker.
That is, an attacker can force the GNN to classify any target node to a chosen label by only slightly perturbing another single arbitrary node in the graph, even when not being able to pick that specific attacker node. When the adversary is allowed to pick a specific attacker node, the attack is even more effective.
We show that this attack is effective across various GNN types (e.g., GraphSAGE, GCN, GAT, and GIN), across a variety of real-world datasets, and as a targeted

and non-targeted attack.
Our code is available anonymously at https://github.com/gnnattack/SINGLE .
****

Dataset Condensation with Gradient Matching
Bo Zhao,Konda Reddy Mopuri,Hakan Bilen
As the state-of-the-art machine learning methods in many fields rely on larger datasets, storing datasets and training models on them become significantly more expensive. This paper proposes a training set synthesis technique for data-efficient learning, called Dataset Condensation, that learns to condense large dataset into a small set of informative synthetic samples for training deep neural networks from scratch. We formulate this goal as a gradient matching problem between the gradients of deep neural network weights that are trained on the original and our synthetic data. We rigorously evaluate its performance in several computer vision benchmarks and demonstrate that it significantly outperforms the state-of-the-art methods. Finally we explore the use of our method in continual learning and neural architecture search and report promising gains when limited memory and computations are available.
****

Taking Notes on the Fly Helps Language Pre-Training
Qiyu Wu,Chen Xing,Yatao Li,Guolin Ke,Di He,Tie-Yan Liu
How to make unsupervised language pre-training more efficient and less resource-intensive is an important research direction in NLP. In this paper, we focus on improving the efficiency of language pre-training methods through providing better data utilization. It is well-known that in language data corpus, words follow a heavy-tail distribution. A large proportion of words appear only very few times and the embeddings of rare words are usually poorly optimized. We argue that such embeddings carry inadequate semantic signals, which could make the data utilization inefficient and slow down the pre-training of the entire model. To mitigate this problem, we propose Taking Notes on the Fly (TNF), which takes notes for rare words on the fly during pre-training to help the model understand them when they occur next time. Specifically, TNF maintains a note dictionary and saves a rare word's contextual information in it as notes when the rare word occurs in a sentence. When the same rare word occurs again during training, the note information saved beforehand can be employed to enhance the semantics of the current sentence. By doing so, TNF provides a better data utilization since cross-sentence information is employed to cover the inadequate semantics caused by rare words in the sentences. We implement TNF on both BERT and ELECTRA to check its efficiency and effectiveness. Experimental results show that TNF's training time is 60% less than its backbone pre-training models when reaching the same performance. When trained with same number of iterations, TNF outperforms its backbone methods on most of downstream tasks and the average GLUE score. Code is attached in the supplementary material.
****

Neural Architecture Search of SPD Manifold Networks
Rhea Sanjay Sukthanker,Zhiwu Huang,Suryansh Kumar,Erik Goron,Yan Wu,Luc Van Gool
In this paper, we propose a new neural architecture search (NAS) problem of Symmetric Positive Definite (SPD) manifold networks. Unlike the conventional NAS problem, our problem requires to search for a unique computational cell called the SPD cell. This SPD cell serves as a basic building block of SPD neural architectures. An efficient solution to our problem is important to minimize the extraneous manual effort in the SPD neural architecture design. To accomplish this goal, we first introduce a geometrically rich and diverse SPD neural architecture search space for an efficient SPD cell design. Further, we model our new NAS problem using the supernet strategy, which models the architecture search problem as a one-shot training process of a single supernet. Based on the supernet modeling, we exploit a differentiable NAS algorithm on our relaxed continuous search space for SPD neural architecture search. Statistical evaluation of our method on drone, action, and emotion recognition tasks mostly provides better results than the state-of-the-art SPD networks and NAS algorithms. Empirical results show that our algorithm excels in discovering better SPD network design and providing mod

els that are more than 3 times lighter than searched by state-of-the-art NAS alg orithms.
****
Fourier Stochastic Backpropagation
Amine Echraibi,Joachim Flocon Cholet,Stéphane Gosselin,Sandrine Vaton
Backpropagating gradients through random variables is at the heart of numerous m achine learning applications. In this paper, we present a general framework for deriving stochastic backpropagation rules for any distribution, discrete or cont inuous. Our approach exploits the link between the characteristic function and t he Fourier transform, to transport the derivatives from the parameters of the di stribution to the random variable. Our method generalizes previously known estim ators, and results in new estimators for the gamma, beta, Dirichlet and Laplace distributions. Furthermore, we show that the classical deterministic backproapag ation rule and the discrete random variable case, can also be interpreted throug h stochastic backpropagation.
****
Predictive Attention Transformer: Improving Transformer with Attention Map Predi ction
Yujing Wang,Yaming Yang,Jiangang Bai,Mingliang Zhang,Jing Bai,Jing Yu,Ce Zhang,Y unhai Tong
Transformer is a ubiquitous model for natural language processing and has also a ttracted wide attentions in other domains such as computer vision. The self-atte ntion maps, learned independently for each layer, are indispensable for a transf ormer model to encode the dependencies among input tokens, however, learning the m effectively is still a challenging problem. In this paper, we address this pro blem and propose a novel approach to improve self-attention through supplementar y prediction modules. The underlying assumption is that the attention structures in the current layer should not be completely independent from those in the pre vious layer. Instead, we model their dependencies via a chain of prediction mode ls that take previous attention maps as input to predict the attention maps of a new layer through convolutional neural networks. Specifically, we propose Predi ctive Attention Transformer and obtain significant performance gains for various kinds of tasks on top of multiple state-of-the-art models. On GLUE benchmark, t he average performances of BERT-Base and BERT-Large are lifted by 4.1 and 2.5 po ints respectively. For machine translation, it improves the BLUE score of a vani lla Transformer consistently on IWSLT'14 De-En dataset with different model size s. For ImageNet classification, we achieve significant improvement over a strong backbone model with comparable capacity.
****
Accurately Solving Rod Dynamics with Graph Learning
Han Shao,Tassilo Kugelstadt,Torsten Hädrich,Wojciech Pa■ubicki,Jan Bender,Soren Pirk,Dominik Michels
Iterative solvers are widely used to accurately simulate physical systems. These solvers require initial guesses to generate a sequence of improving approximate solutions. In this contribution, we introduce a novel method to accelerate iter ative solvers for rod dynamics with graph networks (GNs) by predicting the initi al guesses to reduce the number of iterations. Unlike existing methods that aim to learn physical systems in an end-to-end manner, our approach guarantees long- term stability and therefore leads to more accurate solutions. Furthermore, our method improves the run time performance of traditional iterative solvers for ro d dynamics. To explore our method we make use of position-based dynamics (PBD) a s a common solver for physical systems and evaluate it by simulating the dynamic s of elastic rods. Our approach is able to generalize across different initial c onditions, discretizations, and realistic material properties. We demonstrate th at it also performs well when taking discontinuous effects into account such as collisions between individual rods. Finally, to illustrate the scalability of ou r approach, we simulate complex 3D tree models composed of over a thousand indiv idual branch segments swaying in wind fields.
****
Graph Edit Networks

Benjamin Paassen,Daniele Grattarola,Daniele Zambon,Cesare Alippi,Barbara Eva Hammer

While graph neural networks have made impressive progress in classification and regression, few approaches to date perform time series prediction on graphs, and those that do are mostly limited to edge changes. We suggest that graph edits are a more natural interface for graph-to-graph learning. In particular, graph edits are general enough to describe any graph-to-graph change, not only edge changes; they are sparse, making them easier to understand for humans and more efficient computationally; and they are local, avoiding the need for pooling layers in graph neural networks. In this paper, we propose a novel output layer - the graph edit network - which takes node embeddings as input and generates a sequence of graph edits that transform the input graph to the output graph. We prove that a mapping between the node sets of two graphs is sufficient to construct training data for a graph edit network and that an optimal mapping yields edit scripts that are almost as short as the graph edit distance between the graphs. We further provide a proof-of-concept empirical evaluation on several graph dynamical systems, which are difficult to learn for baselines from the literature.

****

Learn Robust Features via Orthogonal Multi-Path

Kun Fang,Xiaolin Huang,Yingwen Wu,Tao Li,Jie Yang

■It is now widely known that by adversarial attacks, clean images with invisible perturbations can fool deep neural networks.

■To defend adversarial attacks, we design a block containing multiple paths to learn robust features and the parameters of these paths are required to be orthogonal with each other.

■The so-called Orthogonal Multi-Path (OMP) block could be posed in any layer of a neural network.

■Via forward learning and backward correction, one OMP block makes the neural networks learn features that are appropriate for all the paths and hence are expected to be robust. With careful design and thorough experiments on e.g., the positions of imposing orthogonality constraint, and the trade-off between the variety and accuracy,

■the robustness of the neural networks is significantly improved.

■For example, under white-box PGD attack with l_∞ bound ${8}/{255}$ (this is a fierce attack that can make the accuracy of many vanilla neural networks drop to nearly 10% on CIFAR10), VGG16 with the proposed OMP block could keep over 50% accuracy. For black-box attacks, neural networks equipped with an OMP block have accuracy over 80%. The performance under both white-box and black-box attacks is much better than the existing state-of-the-art adversarial defenders.

****

TAM: Temporal Adaptive Module for Video Recognition

Zhaoyang Liu,Limin Wang,Wayne Wu,Chen Qian,Tong Lu

Temporal modeling is crucial for capturing spatiotemporal structure in videos for action recognition. Video data is with extremely complex dynamics along its temporal dimension due to various factors such as camera motion, speed variation, and different activities. To effectively capture this diverse motion pattern, this paper presents a new temporal adaptive module ({\bf TAM}) to generate video-specific kernels based on its own feature maps. TAM proposes a unique two-level adaptive modeling scheme by decoupling dynamic kernels into a location sensitive importance map and a location invariant aggregation weight. The importance map is learned in a local temporal window to capture short term information, while the aggregation weight is generated from a global view with a focus on long-term structure. TAM is a principled module and could be integrated into 2D CNNs to yield a powerful video architecture (TANet) with a very small extra computational cost. The extensive experiments on Kinetics-400 and Something-Something datasets, demonstrate that the TAM outperforms other temporal modeling methods consistently owing to its temporal adaptive modeling strategy.

****

FOCAL: Efficient Fully-Offline Meta-Reinforcement Learning via Distance Metric L

earning and Behavior Regularization
Lanqing Li,Rui Yang,Dijun Luo
We study the offline meta-reinforcement learning (OMRL) problem, a paradigm whic h enables reinforcement learning (RL) algorithms to quickly adapt to unseen task s without any interactions with the environments, making RL truly practical in m any real-world applications. This problem is still not fully understood, for whi ch two major challenges need to be addressed. First, offline RL usually suffers from bootstrapping errors of out-of-distribution state-actions which leads to di vergence of value functions. Second, meta-RL requires efficient and robust task inference learned jointly with control policy. In this work, we enforce behavior regularization on learned policy as a general approach to offline RL, combined with a deterministic context encoder for efficient task inference. We propose a novel negative-power distance metric on bounded context embedding space, whose g radients propagation is detached from the Bellman backup. We provide analysis an d insight showing that some simple design choices can yield substantial improvem ents over recent approaches involving meta-RL and distance metric learning. To t he best of our knowledge, our method is the first model-free and end-to-end OMRL algorithm, which is computationally efficient and demonstrated to outperform pr ior algorithms on several meta-RL benchmarks.
****
CURI: A Benchmark for Productive Concept Learning Under Uncertainty
Shanmukha Ramakrishna Vedantam,Arthur Szlam,Maximilian Nickel,Ari S. Morcos,Bren den M. Lake
Humans can learn and reason under substantial uncertainty in a space of infinite ly many concepts, including structured relational concepts (“a scene with object s that have the same color”) and ad-hoc categories defined through goals (“objec ts that could fall on one’s head”). In contrast, standard classification benchma rks: 1) consider only a fixed set of category labels, 2) do not evaluate composi tional concept learning and 3) do not explicitly capture a notion of reasoning u nder uncertainty. We introduce a new few-shot, meta-learning benchmark, Composit ional Reasoning Under Uncertainty (CURI) to bridge this gap. CURI evaluates diff erent aspects of productive and systematic generalization, including abstract un derstandings of disentangling, productive generalization, learning boolean opera tions, variable binding, etc. Importantly, it also defines a model-independent “ compositionality gap” to evaluate difficulty of generalizing out-of-distribution along each of these axes. Extensive evaluations across a range of modeling choi ces spanning different modalities (image, schemas, and sounds), splits, privileg ed auxiliary concept information, and choices of negatives reveal substantial sc ope for modeling advances on the proposed task. All code and datasets will be av ailable online.
****
Effective and Efficient Vote Attack on Capsule Networks
Jindong Gu,Baoyuan Wu,Volker Tresp
Standard Convolutional Neural Networks (CNNs) can be easily fooled by images wit h small quasi-imperceptible artificial perturbations. As alternatives to CNNs, t he recently proposed Capsule Networks (CapsNets) are shown to be more robust to white-box attack than CNNs under popular attack protocols. Besides, the class-co nditional reconstruction part of CapsNets is also used to detect adversarial exa mples. In this work, we investigate the adversarial robustness of CapsNets, espe cially how the inner workings of CapsNets change when the output capsules are at tacked. The first observation is that adversarial examples misled CapsNets by ma nipulating the votes from primary capsules. Another observation is the high comp utational cost, when we directly apply multi-step attack methods designed for CN Ns to attack CapsNets, due to the computationally expensive routing mechanism. M otivated by these two observations, we propose a novel vote attack where we atta ck votes of CapsNets directly. Our vote attack is not only effective, but also e fficient by circumventing the routing process. Furthermore, we integrate our vot e attack into the detection-aware attack paradigm, which can successfully bypass the class-conditional reconstruction based detection method. Extensive experime nts demonstrate the superior attack performance of our vote attack on CapsNets.

****
Rapid Neural Architecture Search by Learning to Generate Graphs from Datasets
Hayeon Lee,Eunyoung Hyung,Sung Ju Hwang
Despite the success of recent Neural Architecture Search (NAS) methods on various tasks which have shown to output networks that largely outperform human-designed networks, conventional NAS methods have mostly tackled the optimization of searching for the network architecture for a single task (dataset), which does not generalize well across multiple tasks (datasets). Moreover, since such task-specific methods search for a neural architecture from scratch for every given task, they incur a large computational cost, which is problematic when the time and monetary budget are limited. In this paper, we propose an efficient NAS framework that is trained once on a database consisting of datasets and pretrained networks and can rapidly search for a neural architecture for a novel dataset. The proposed MetaD2A (Meta Dataset-to-Architecture) model can stochastically generate graphs (architectures) from a given set (dataset) via a cross-modal latent space learned with amortized meta-learning. Moreover, we also propose a meta-performance predictor to estimate and select the best architecture without direct training on target datasets. The experimental results demonstrate that our model meta-learned on subsets of ImageNet-1K and architectures from NAS-Bench 201 search space successfully generalizes to multiple unseen datasets including CIFAR-10 and CIFAR-100, with an average search time of 33 GPU seconds. Even under MobileNetV3 search space, MetaD2A is 5.5K times faster than NSGANetV2, a transferable NAS method, with comparable performance. We believe that the MetaD2A proposes a new research direction for rapid NAS as well as ways to utilize the knowledge from rich databases of datasets and architectures accumulated over the past years. Code is available at https://github.com/HayeonLee/MetaD2A.
****
Reservoir Transformers
Sheng Shen,Alexei Baevski,Ari S. Morcos,Kurt Keutzer,Michael Auli,Douwe Kiela
We demonstrate that transformers obtain impressive performance even when some of the layers are randomly initialized and never updated. Inspired by old and well-established ideas in machine learning, we explore a variety of non-linear reservoir layers interspersed with regular transformer layers, and show improvements in wall-clock compute time until convergence, as well as overall performance, on various machine translation and (masked) language modelling tasks.
****
AUTOSAMPLING: SEARCH FOR EFFECTIVE DATA SAMPLING SCHEDULES
Ming Sun,Haoxuan Dou,Baopu Li,Junjie Yan,Wanli Ouyang
Data sampling acts as a pivotal role in training deep learning models. However, an effective sampling schedule is difficult to learn due to its inherent high-dimension as a hyper-parameter. In this paper, we propose the AutoSampling method to automatically learn sampling schedules for model training, which consists of the multi-exploitation step aiming for optimal local sampling schedules and the exploration step for the ideal sampling distribution. More specifically, we achieve sampling schedule search with shortened exploitation cycle to provide enough supervision. In addition, we periodically estimate the sampling distribution from the learned sampling schedules and perturb it to search in the distribution space. The combination of two searches allows us to learn a robust sampling schedule. We apply our AutoSampling method to a variety of image classification tasks illustrating the effectiveness of the proposed method.
****
Deep Reinforcement Learning With Adaptive Combined Critics
Huihui Zhang,Wu Huang
The overestimation problem has long been popular in deep value learning, because function approximation errors may lead to amplified value estimates and suboptimal policies. There have been several methods to deal with the overestimation problem, however, further problems may be induced, for example, the underestimation bias and instability. In this paper, we focus on the overestimation issues on continuous control through deep reinforcement learning, and propose a novel algorithm that can minimize the overestimation, avoid the underestimation bias and r

etain the policy improvement during the whole training process. Specifically, we add a weight factor to adjust the influence of two independent critics, and use the combined value of weighted critics to update the policy. Then the updated policy is involved in the update of the weight factor, in which we propose a novel method to provide theoretical and experimental guarantee for future policy improvement. We evaluate our method on a set of classical control tasks, and the results show that the proposed algorithms are more computationally efficient and stable than several existing algorithms for continuous control.
****
Impact of Representation Learning in Linear Bandits
Jiaqi Yang,Wei Hu,Jason D. Lee,Simon Shaolei Du
We study how representation learning can improve the efficiency of bandit problems. We study the setting where we play T linear bandits with dimension d concurrently, and these T bandit tasks share a common $k (\ll d)$ dimensional linear representation. For the finite-action setting, we present a new algorithm which achieves $\widetilde{O}(T\sqrt{kN} + \sqrt{dkNT})$ regret, where N is the number of rounds we play for each bandit. When T is sufficiently large, our algorithm significantly outperforms the naive algorithm (playing T bandits independently) that achieves $\widetilde{O}(T\sqrt{d N})$ regret. We also provide an $\Omega(T\sqrt{kN} + \sqrt{dkNT})$ regret lower bound, showing that our algorithm is minimax-optimal up to poly-logarithmic factors. Furthermore, we extend our algorithm to the infinite-action setting and obtain a corresponding regret bound which demonstrates the benefit of representation learning in certain regimes. We also present experiments on synthetic and real-world data to illustrate our theoretical findings and demonstrate the effectiveness of our proposed algorithms.
****
BRAC+: Going Deeper with Behavior Regularized Offline Reinforcement Learning
Chi Zhang,Sanmukh Rao Kuppannagari,Viktor Prasanna
Online interactions with the environment to collect data samples for training a Reinforcement Learning agent is not always feasible due to economic and safety concerns. The goal of Offline Reinforcement Learning (RL) is to address this problem by learning effective policies using previously collected datasets. Standard off-policy RL algorithms are prone to overestimations of the values of out-of-distribution (less explored) actions and are hence unsuitable for Offline RL. Behavior regularization, which constraints the learned policy within the support set of the dataset, has been proposed to tackle the limitations of standard off-policy algorithms. In this paper, we improve the behavior regularized offline reinforcement learning and propose \emph{BRAC+}. We use an analytical upper bound on KL divergence as the behavior regularizor to reduce variance associated with sample based estimations. Additionally, we employ state-dependent Lagrange multipliers for the regularization term to avoid distributing KL divergence penalty across all states of the sampled batch. The proposed Lagrange multipliers allow more freedom of deviation to high probability (more explored) states leading to better rewards while simultaneously restricting low probability (less explored) states to prevent out-of-distribution actions. \textcolor{blue}{To prevent catastrophic performance degradation due to rare out-of-distribution actions, we add a gradient penalty term to the policy evaluation objective to penalize the gradient of the Q value w.r.t the out-of-distribution actions. By doing so, the Q values evaluated at the out-of-distribution actions are bounded.} On challenging offline RL benchmarks, BRAC+ outperforms the state-of-the-art model-free and model-based approaches.
****
Creative Sketch Generation
Songwei Ge,Vedanuj Goswami,Larry Zitnick,Devi Parikh
Sketching or doodling is a popular creative activity that people engage in. However, most existing work in automatic sketch understanding or generation has focused on sketches that are quite mundane. In this work, we introduce two datasets of creative sketches -- Creative Birds and Creative Creatures -- containing 10k sketches each along with part annotations. We propose DoodlerGAN -- a part-based

Generative Adversarial Network (GAN) -- to generate unseen compositions of novel part appearances. Quantitative evaluations as well as human studies demonstrate that sketches generated by our approach are more creative and of higher quality than existing approaches. In fact, in Creative Birds, subjects prefer sketches generated by DoodlerGAN over those drawn by humans!
****
Self-supervised Representation Learning with Relative Predictive Coding
Yao-Hung Hubert Tsai,Martin Q. Ma,Muqiao Yang,Han Zhao,Louis-Philippe Morency,Ruslan Salakhutdinov
This paper introduces Relative Predictive Coding (RPC), a new contrastive representation learning objective that maintains a good balance among training stability, minibatch size sensitivity, and downstream task performance. The key to the success of RPC is two-fold. First, RPC introduces the relative parameters to regularize the objective for boundedness and low variance. Second, RPC contains no logarithm and exponential score functions, which are the main cause of training instability in prior contrastive objectives. We empirically verify the effectiveness of RPC on benchmark vision and speech self-supervised learning tasks. Lastly, we relate RPC with mutual information (MI) estimation, showing RPC can be used to estimate MI with low variance.
****
BBRefinement: an universal scheme to improve precision of box object detectors
Petr Hurtik,Marek Vajgl
We present a conceptually simple yet powerful and flexible scheme for refining predictions of bounding boxes. Our approach is trained standalone on GT boxes and can then be combined with an object detector to improve its predictions. The method, called BBRefinement, uses mixture data of image information and the object's class and center. Due to the transformation of the problem into a domain where BBRefinement does not care about multiscale detection, recognition of the object's class, computing confidence, or multiple detections, the training is much more effective. It results in the ability to refine even COCO's ground truth labels into a more precise form. BBRefinement improves the performance of SOTA architectures up to 2mAP points on the COCO dataset in the benchmark. The refinement process is fast; it adds 50-80ms overhead to a standard detector using RTX2080, so it can run in real-time on standard hardware. The code is available at https://gitlab.com/irafm-ai/bb-refinement.
****
Differentiable Spatial Planning using Transformers
Devendra Singh Chaplot,Deepak Pathak,Jitendra Malik
We consider the problem of spatial path planning. In contrast to the classical solutions which optimize a new plan from scratch and assume access to the full map with ground truth obstacle locations, we learn a planner from the data in a differentiable manner that allows us to leverage statistical regularities from past data. We propose Spatial Planning Transformers (SPT), which given an obstacle map learns to generate actions by planning over long-range spatial dependencies, unlike prior data-driven planners that propagate information locally via convolutional structure in an iterative manner. In the setting where the ground truth map is not known to the agent, we leverage pre-trained SPTs to in an end-to-end framework that has the structure of mapper and planner built into it which allows seamless generalization to out-of-distribution maps and goals. SPTs outperform prior state-of-the-art across all the setups for both manipulation and navigation tasks, leading to an absolute improvement of 7-19%.
****
Feature-Robust Optimal Transport for High-Dimensional Data
Mathis Petrovich,Chao Liang,Ryoma Sato,Yanbin Liu,Yao-Hung Hubert Tsai,Linchao Zhu,Yi Yang,Ruslan Salakhutdinov,Makoto Yamada
Optimal transport is a machine learning problem with applications including distribution comparison, feature selection, and generative adversarial networks. In this paper, we propose feature-robust optimal transport (FROT) for high-dimensional data, which solves high-dimensional OT problems using feature selection to avoid the curse of dimensionality. Specifically, we find a transport plan with di

scriminative features. To this end, we formulate the FROT problem as a min--max optimization problem. We then propose a convex formulation of the FROT problem a nd solve it using a Frank--Wolfe-based optimization algorithm, whereby the subpr oblem can be efficiently solved using the Sinkhorn algorithm. Since FROT finds t he transport plan from selected features, it is robust to noise features. To sho w the effectiveness of FROT, we propose using the FROT algorithm for the layer s election problem in deep neural networks for semantic correspondence. By conduct ing synthetic and benchmark experiments, we demonstrate that the proposed method can find a strong correspondence by determining important layers. We show that the FROT algorithm achieves state-of-the-art performance in real-world semantic correspondence datasets.

****

Model-based Navigation in Environments with Novel Layouts Using Abstract 2-D M aps

Linfeng Zhao,Lawson L. S. Wong

Efficiently training agents with planning capabilities has long been one of the major challenges in decision-making. In this work, we focus on zero-shot navigat ion ability on a given abstract 2-D occupancy map, like human navigation by read ing a paper map, by treating it as an image. To learn this ability, we need to e fficiently train an agent on environments with a small proportion of training ma ps and share knowledge effectively across the environments. We hypothesize that model-based navigation can better adapt an agent's behaviors to a task, since it disentangles the variations in map layout and goal location and enables longer- term planning ability on novel locations compared to reactive policies. We propo se to learn a hypermodel that can understand patterns from a limited number of a bstract maps and goal locations, to maximize alignment between the hypermodel pr edictions and real trajectories to extract information from multi-task off-polic y experiences, and to construct denser feedback for planners by n-step goal re labelling. We train our approach on DeepMind Lab environments with layouts from different maps, and demonstrate superior performance on zero-shot transfer to no vel maps and goals.

****

PURE: An Uncertainty-aware Recommendation Framework for Maximizing Expected Post erior Utility of Platform

Haokun Chen,Zhaoyang Liu,Chen Xu,Ziqian Chen,Jinyang Gao,Bolin Ding

Commercial recommendation can be regarded as an interactive process between the recommendation platform and its target users. One crucial problem for the platfo rm is how to make full use of its advantages so as to maximize its utility, i.e. , the commercial benefits from recommendation. In this paper, we propose a novel recommendation framework which effectively utilizes the information of user unc ertainty over different item dimensions and explicitly takes into consideration the impact of display policy on user in order to achieve maximal expected poster ior utility for the platform. We formulate the problem of deriving optimal polic y to achieve maximal expected posterior utility as a constrained non-convex opti mization problem and further propose an ADMM-based solution to derive an approxi mately optimal policy. Extensive experiments are conducted over data collected f rom a real-world recommendation platform and demonstrate the effectiveness of th e proposed framework. Besides, we also adopt the proposed framework to conduct e xperiments with an intent to reveal how the platform achieves its commercial ben efits. The results suggest that the platform should cater to the user's preferen ce for item dimensions that the user prefers, while for item dimensions where th e user is with high uncertainty, the platform can achieve more commercial benefi ts by recommending items with high utilities.

****

Transfer among Agents: An Efficient Multiagent Transfer Learning Framework

Tianpei Yang,Jianye HAO,Weixun Wang,Hongyao Tang,Zhaopeng Meng,Hangyu Mao,Dong L i,Wulong Liu,Yujing Hu,Yingfeng Chen,Changjie Fan

Transfer Learning has shown great potential to enhance the single-agent Reinforc ement Learning (RL) efficiency, by sharing learned policies of previous tasks. S imilarly, in multiagent settings, the learning performance can also be promoted

if agents can share knowledge between each other. However, it remains an open qu
estion of how an agent should learn from other agents' knowledge. In this paper,
 we propose a novel multiagent option-based policy transfer (MAOPT) framework to
 improve multiagent learning efficiency. Our framework learns what advice to giv
e to each agent and when to terminate it by modeling multiagent policy transfer
as the option learning problem. MAOPT provides different kinds of variants which
 can be classified into two types in terms of the experience used during trainin
g. One type is the MAOPT with the Global Option Advisor which has the access to
the global information of the environment. However, in many realistic scenarios,
 we can only obtain each agent's local information due to the partial observatio
n. The other type contains MAOPT with the Local Option Advisor and MAOPT with th
e Successor Representation Option (SRO) which are suitable for this setting and
collect each agent's local experience for the update. In many cases, each agent'
s experience is inconsistent with each other which causes the option-value estim
ation to oscillate and to become inaccurate. SRO is used to handle the experienc
e inconsistency by decoupling the dynamics of the environment from the rewards t
o learn the option-value function under each agent's preference. MAOPT can be ea
sily combined with existing deep RL approaches. Experimental results show it sig
nificantly boosts the performance of existing deep RL methods in both discrete a
nd continuous state spaces.
****

Neighbor2Seq: Deep Learning on Massive Graphs by Transforming Neighbors to Seque
nces

Meng Liu,Shuiwang Ji

Modern Graph Neural Networks (GNNs) follow a recursive neighbor-wise message pas
sing scheme and have achieved great success in many fields. However, this recurs
ive design brings expensive computation and huge memory usage, making it difficu
lt to deploy on large-scale graphs. In this work, we propose Neighbor2Seq, which
 transforms the hierarchical neighborhood of each node into an ordered sequence
and enables the subsequent utilization of general deep learning operations, such
 as convolution and attention. Neighbor2Seq grants our proposed models, i.e., Ne
ighbor2Seq-Conv and Neighbor2Seq-Attn, the ability to learn on arbitrarily large
 graphs as long as the Neighbor2Seq step can be precomputed. Another potential a
dvantage obtained by the way is that Neighbor2Seq can alleviate the over-squashi
ng issue existing in modern GNNs. We conduct thorough experiments on a massive g
raph with more than 111 million nodes and 1.6 billion edges, as well as several
medium-scale graphs, to evaluate our proposed method. Experimental results demon
strate that our proposed method is scalable to the massive graph and achieves su
perior performance across datasets.
****

Additive Poisson Process: Learning Intensity of Higher-Order Interaction in Stoc
hastic 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 stochastic pr
ocesses using lower dimensional projections. Our model combines the techniques i
n information geometry to model higher-order interactions on a statistical manif
old and in generalized additive models to use lower-dimensional projections to o
vercome the effects from the curse of dimensionality. Our approach solves a conv
ex optimization problem by minimizing the KL divergence from a sample distributi
on in lower dimensional projections to the distribution modeled by an intensity
function in the stochastic process. Our empirical results show that our model is
 able to use samples observed in the lower dimensional space to estimate the hig
her-order intensity function with extremely sparse observations.
****

Hierarchical Probabilistic Model for Blind Source Separation via Legendre Transf
ormation

Simon Luo,Lamiae Azizi,Mahito Sugiyama

We present a novel blind source separation (BSS) method, called information geom
etric blind source separation (IGBSS). Our formulation is based on the log-linea

r model equipped with a hierarchically structured sample space, which has theore tical guarantees to uniquely recover a set of source signals by minimizing the K L divergence from a set of mixed signals. Source signals, received signals, and mixing matrices are realized as different layers in our hierarchical sample spac e. Our empirical results have demonstrated on images and time series data that o ur approach is superior to well established techniques and is able to separate s ignals with complex interactions.
****
Cubic Spline Smoothing Compensation for Irregularly Sampled Sequences
Jing Shi,Jing Bi,Yingru Liu,Chenliang Xu
The marriage of recurrent neural networks and neural ordinary differential netwo rks (ODE-RNN) is effective in modeling irregularly sampled sequences.
While ODE produces the smooth hidden states between observation intervals, the R NN will trigger a hidden state jump when a new observation arrives and thus caus e the interpolation discontinuity problem.
To address this issue, we propose the cubic spline smoothing compensation, which is a stand-alone module upon either the output or the hidden state of ODE-RNN a nd can be trained end-to-end.
We derive its analytical solution and provide its theoretical interpolation erro r bound.
Extensive experiments indicate its merits over both ODE-RNN and cubic spline int erpolation.
****
Human Perception-based Evaluation Criterion for Ultra-high Resolution Cell Membr ane Segmentation
Ruohua Shi,Wenyao Wang,Zhixuan Li,Liuyuan He,Kaiwen Sheng,Lei Ma,Kai Du,Tingting Jiang,Tiejun Huang
Computer vision technology is widely used in biological and medical data analysi s and understanding. However, there are still two major bottlenecks in the field of cell membrane segmentation, which seriously hinder further research: lack of sufficient high-quality data and lack of suitable evaluation criteria. In order to solve these two problems, this paper first introduces an Ultra-high Resoluti on Image Segmentation dataset for the Cell membrane, called U-RISC, the largest annotated EM dataset for the Cell membrane with multiple iterative annotations a nd uncompressed high-resolution raw data. During the analysis process of the U-R ISC, we found that the current popular segmentation evaluation criteria are inco nsistent with human perception. This interesting phenomenon is confirmed by a su bjective experiment involving twenty people. Furthermore, to resolve this incons istency, we propose a Perceptual Hausdorff Distance (PHD) evaluation criterion t o measure the quality of cell membrane segmentation results. Detailed performanc e comparison and discussion of classic segmentation methods along with two itera tive manual annotation results under existing criteria and PHD is given.
****
With False Friends Like These, Who Can Have Self-Knowledge?
Lue Tao,Songcan Chen
Adversarial examples arise from excessive sensitivity of a model. Commonly studi ed adversarial examples are malicious inputs, crafted by an adversary from corre ctly classified examples, to induce misclassification. This paper studies an int riguing, yet far overlooked consequence of the excessive sensitivity, that is, a misclassified example can be easily perturbed to help the model to produce corr ect output. Such perturbed examples look harmless, but actually can be malicious ly utilized by a false friend to make the model self-satisfied. Thus we name the m hypocritical examples. With false friends like these, a poorly performed model could behave like a state-of-the-art one. Once a deployer trusts the hypocritic al performance and uses the "well-performed" model in real-world applications, p otential security concerns appear even in benign environments. In this paper, we formalize the hypocritical risk for the first time and propose a defense method specialized for hypocritical examples by minimizing the tradeoff between natura l risk and an upper bound of hypocritical risk. Moreover, our theoretical analys is reveals connections between adversarial risk and hypocritical risk. Extensive

experiments verify the theoretical results and the effectiveness of our proposed methods.
****
A Sharp Analysis of Model-based Reinforcement Learning with Self-Play
Qinghua Liu,Tiancheng Yu,Yu Bai,Chi Jin
Model-based algorithms---algorithms that explore the environment through building and utilizing an estimated model---are widely used in reinforcement learning practice and theoretically shown to achieve optimal sample efficiency for single-agent reinforcement learning in Markov Decision Processes (MDPs). However, for multi-agent reinforcement learning in Markov games, the current best known sample complexity for model-based algorithms is rather suboptimal and compares unfavorably against recent model-free approaches. In this paper, we present a sharp analysis of model-based self-play algorithms for multi-agent Markov games. We design an algorithm \emph{Optimistic Nash Value Iteration} (Nash-VI) for two-player zero-sum Markov games that is able to output an ϵ-approximate Nash policy in $\tilde{\mathcal{O}}(H^3SAB/\epsilon^2)$ episodes of game playing, where S is the number of states, A,B are the number of actions for the two players respectively, and H is the horizon length. This significantly improves over the best known model-based guarantee of $\tilde{\mathcal{O}}(H^4S^2AB/\epsilon^2)$, and is the first that matches the information-theoretic lower bound $\Omega(H^3S(A+B)/\epsilon^2)$ except for a $\min\{A,B\}$ factor. In addition, our guarantee compares favorably against the best known model-free algorithm if $\min\{A,B\}=o(H^3)$, and outputs a single Markov policy while existing sample-efficient model-free algorithms output a nested mixture of Markov policies that is in general non-Markov and rather inconvenient to store and execute. We further adapt our analysis to designing a provably efficient task-agnostic algorithm for zero-sum Markov games, and designing the first line of provably sample-efficient algorithms for multi-player general-sum Markov games.
****
Contrastive Video Textures
Medhini Narasimhan,Shiry Ginosar,Andrew Owens,Alexei A Efros,Trevor Darrell
Existing methods for video generation struggle to generate more than a short sequence of frames. We introduce a non-parametric approach for infinite video generation based on learning to resample frames from an input video. Our work is inspired by Video Textures, a classic method relying on pixel similarity to stitch sequences of frames, which performs well for videos with a high degree of regularity but fails in less constrained settings. Our method learns a distance metric to compare frames in a manner that scales to more challenging dynamics and allows for conditioning on heterogeneous data, such as audio. We learn representations for video frames and probabilities of transitioning by fitting a video-specific bi-gram model trained using contrastive learning. To synthesize the texture, we represent the video as a graph where the nodes are frames and edges are transitions with probabilities predicted by our video-specific model. By randomly traversing edges with high transition probabilities, we generate diverse temporally smooth videos with novel sequences and transitions. The model naturally extends with no additional training to handle the task of Audio Conditioned Video Synthesis, when conditioned on an audio signal. Our model outperforms baselines on human perceptual scores, can handle a diverse range of input videos, and can combine semantic and audio-visual cues in order to synthesize videos that synchronize well with an audio signal.
****
HALMA: Humanlike Abstraction Learning Meets Affordance in Rapid Problem Solving
Sirui Xie,Xiaojian Ma,Peiyu Yu,Yixin Zhu,Ying Nian Wu,Song-Chun Zhu
Humans learn compositional and causal abstraction, \ie, knowledge, in response to the structure of naturalistic tasks. When presented with a problem-solving task involving some objects, toddlers would first interact with these objects to reckon what they are and what can be done with them. Leveraging these concepts, they could understand the internal structure of this task, without seeing all of the problem instances. Remarkably, they further build cognitively executable strategies to \emph{rapidly} solve novel problems. To empower a learning agent with

similar capability, we argue there shall be three levels of generalization in how an agent represents its knowledge: perceptual, conceptual, and algorithmic. In this paper, we devise the very first systematic benchmark that offers joint evaluation covering all three levels. This benchmark is centered around a novel task domain, HALMA, for visual concept development and rapid problem solving. Uniquely, HALMA has a minimum yet complete concept space, upon which we introduce a novel paradigm to rigorously diagnose and dissect learning agents' capability in understanding and generalizing complex and structural concepts. We conduct extensive experiments on reinforcement learning agents with various inductive biases and carefully report their proficiency and weakness.
****

Learning to Search for Fast Maximum Common Subgraph Detection
Yunsheng Bai,Derek Qiang Xu,Yizhou Sun,Wei Wang
Detecting the Maximum Common Subgraph (MCS) between two input graphs is fundamental for applications in biomedical analysis, malware detection, cloud computing, etc. This is especially important in the task of drug design, where the successful extraction of common substructures in compounds can reduce the number of experiments needed to be conducted by humans. However, MCS computation is NP-hard, and state-of-the-art MCS solvers rely on heuristics in search which in practice cannot find good solution for large graph pairs under a limited search budget. Here we propose GLSearch, a Graph Neural Network based model for MCS detection, which learns to search. Our model uses a state-of-the-art branch and bound algorithm as the backbone search algorithm to extract subgraphs by selecting one node pair at a time. In order to make better node selection decision at each step, we replace the node selection heuristics with a novel task-specific Deep Q-Network (DQN), allowing the search process to find larger common subgraphs faster. To enhance the training of DQN, we leverage the search process to provide supervision in a pre-training stage and guide our agent during an imitation learning stage. Therefore, our framework allows search and reinforcement learning to mutually benefit each other. Experiments on synthetic and real-world large graph pairs demonstrate that our model outperforms state-of-the-art MCS solvers and neural graph matching network models.
****

One Network Fits All? Modular versus Monolithic Task Formulations in Neural Networks
Atish Agarwala,Abhimanyu Das,Brendan Juba,Rina Panigrahy,Vatsal Sharan,Xin Wang,Qiuyi Zhang
Can deep learning solve multiple, very different tasks simultaneously? We investigate how the representations of the underlying tasks affect the ability of a single neural network to learn them jointly. We present theoretical and empirical findings that a single neural network is capable of simultaneously learning multiple tasks from a combined data set, for a variety of methods for representing tasks---for example, when the distinct tasks are encoded by well-separated clusters or decision trees over some task-code attributes. Indeed, more strongly, we present a novel analysis that shows that families of simple programming-like constructs for the codes encoding the tasks are learnable by two-layer neural networks with standard training. We study more generally how the complexity of learning such combined tasks grows with the complexity of the task codes; we find that learning many tasks can be provably hard, even though the individual tasks are easy to learn. We provide empirical support for the usefulness of the learning bounds by training networks on clusters, decision trees, and SQL-style aggregation.
****

Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data
Jonathan Pilault,Amine El hattami,Christopher Pal
Multi-Task Learning (MTL) networks have emerged as a promising method for transferring learned knowledge across different tasks. However, MTL must deal with challenges such as: overfitting to low resource tasks, catastrophic forgetting, and negative task transfer, or learning interference. Often, in Natural Language Pr

ocessing (NLP), a separate model per task is needed to obtain the best performan ce. However, many fine-tuning approaches are both parameter inefficient, i.e., p otentially involving one new model per task, and highly susceptible to losing kn owledge acquired during pretraining. We propose a novel Transformer based Hypern etwork Adapter consisting of a new conditional attention mechanism as well as a set of task-conditioned modules that facilitate weight sharing. Through this con struction, we achieve more efficient parameter sharing and mitigate forgetting b y keeping half of the weights of a pretrained model fixed. We also use a new mul ti-task data sampling strategy to mitigate the negative effects of data imbalanc e across tasks. Using this approach, we are able to surpass single task fine-tun ing methods while being parameter and data efficient (using around 66% of the da ta). Compared to other BERT Large methods on GLUE, our 8-task model surpasses ot her Adapter methods by 2.8% and our 24-task model outperforms by 0.7-1.0% models that use MTL and single task fine-tuning. We show that a larger variant of our single multi-task model approach performs competitively across 26 NLP tasks and yields state-of-the-art results on a number of test and development sets.

****

A Deeper Look at Discounting Mismatch in Actor-Critic Algorithms

Shangtong Zhang,Romain Laroche,Harm van Seijen,Shimon Whiteson,Remi Tachet des C ombes

We investigate the discounting mismatch in actor-critic algorithm implementation s from a representation learning perspective. Theoretically, actor-critic algori thms usually have discounting for both actor and critic, i.e., there is a $\gamm a^t$ term in the actor update for the transition observed at time t in a traje ctory and the critic is a discounted value function. Practitioners, however, usu ally ignore the discounting (γ^t) for the actor while using a discounted critic. We investigate this mismatch in two scenarios. In the first scenario, we consider optimizing an undiscounted objective $(\gamma = 1)$ where γ^t d isappears naturally $(1^t = 1)$. We then propose to interpret the discounting in critic in terms of a bias-variance-representation trade-off and provide support ing empirical results. In the second scenario, we consider optimizing a discount ed objective ($\gamma < 1$) and propose to interpret the omission of the discoun ting in the actor update from an auxiliary task perspective and provide supporti ng empirical results.

****

A Universal Representation Transformer Layer for Few-Shot Image Classification

Lu Liu,William L. Hamilton,Guodong Long,Jing Jiang,Hugo Larochelle

Few-shot classification aims to recognize unseen classes when presented with onl y a small number of samples. We consider the problem of multi-domain few-shot im age classification, where unseen classes and examples come from diverse data sou rces. This problem has seen growing interest and has inspired the development of benchmarks such as Meta-Dataset. A key challenge in this multi-domain setting i s to effectively integrate the feature representations from the diverse set of t raining domains. Here, we propose a Universal Representation Transformer (URT) l ayer, that meta-learns to leverage universal features for few-shot classificatio n by dynamically re-weighting and composing the most appropriate domain-specific representations. In experiments, we show that URT sets a new state-of-the-art r esult on Meta-Dataset. Specifically, it achieves top-performance on the highest number of data sources compared to competing methods. We analyze variants of URT and present a visualization of the attention score heatmaps that sheds light on how the model performs cross-domain generalization.

****

Isometric Propagation Network for Generalized Zero-shot Learning

Lu Liu,Tianyi Zhou,Guodong Long,Jing Jiang,Xuanyi Dong,Chengqi Zhang

Zero-shot learning (ZSL) aims to classify images of an unseen class only based o n a few attributes describing that class but no access to any training sample. A popular strategy is to learn a mapping between the semantic space of class attr ibutes and the visual space of images based on the seen classes and their data. Thus, an unseen class image can be ideally mapped to its corresponding class att ributes. The key challenge is how to align the representations in the two spaces

. For most ZSL settings, the attributes for each seen/unseen class are only repr
esented by a vector while the seen-class data provide much more information. Thu
s, the imbalanced supervision from the semantic and the visual space can make th
e learned mapping easily overfitting to the seen classes. To resolve this proble
m, we propose Isometric Propagation Network (IPN), which learns to strengthen th
e relation between classes within each space and align the class dependency in t
he two spaces. Specifically, IPN learns to propagate the class representations o
n an auto-generated graph within each space. In contrast to only aligning the re
sulted static representation, we regularize the two dynamic propagation procedur
es to be isometric in terms of the two graphs' edge weights per step by minimizi
ng a consistency loss between them. IPN achieves state-of-the-art performance on
 three popular ZSL benchmarks. To evaluate the generalization capability of IPN,
 we further build two larger benchmarks with more diverse unseen classes and dem
onstrate the advantages of IPN on them.
****

Scalable Learning and MAP Inference for Nonsymmetric Determinantal Point Process
es

Mike Gartrell,Insu Han,Elvis Dohmatob,Jennifer Gillenwater,Victor-Emmanuel Brune
l

Determinantal point processes (DPPs) have attracted significant attention in mac
hine learning for their ability to model subsets drawn from a large item collect
ion. Recent work shows that nonsymmetric DPP (NDPP) kernels have significant adv
antages over symmetric kernels in terms of modeling power and predictive perform
ance. However, for an item collection of size M, existing NDPP learning and in
ference algorithms require memory quadratic in M and runtime cubic (for learni
ng) or quadratic (for inference) in M, making them impractical for many typica
l subset selection tasks. In this work, we develop a learning algorithm with spa
ce and time requirements linear in M by introducing a new NDPP kernel decompos
ition. We also derive a linear-complexity NDPP maximum a posteriori (MAP) infere
nce algorithm that applies not only to our new kernel but also to that of prior
work. Through evaluation on real-world datasets, we show that our algorithms sca
le significantly better, and can match the predictive performance of prior work.
****

Selective Sensing: A Data-driven Nonuniform Subsampling Approach for Computation
-free On-Sensor Data Dimensionality Reduction

Zhikang Zhang,Kai Xu,Fengbo Ren

Designing an on-sensor data dimensionality reduction scheme for efficient signal
 sensing has always been a challenging task. Compressive sensing is a state-of-t
he-art sensing technique used for on-sensor data dimensionality reduction. Howev
er, the undesired computational complexity involved in the sensing stage of comp
ressive sensing limits its practical application in resource-constrained sensor
devices or high-data-rate sensor devices dealing with high-dimensional signals.
In this paper, we propose a selective sensing framework that adopts the novel co
ncept of data-driven nonuniform subsampling to reduce the dimensionality of acqu
ired signals while retaining the information of interest in a computation-free f
ashion. Selective sensing adopts a co-optimization methodology to co-train a sel
ective sensing operator with a subsequent information decoding neural network. W
e take image as the sensing modality and reconstruction as the information decod
ing task to demonstrate the 1st proof-of-concept of selective sensing. The exper
iment results on CIFAR10, Set5 and Set14 datasets show that selective sensing ca
n achieve an average reconstruction accuracy improvement in terms of PSNR/SSIM b
y 3.73dB/0.07 and 9.43dB/0.16 over compressive sensing and uniform subsampling c
ounterparts across the compression ratios of 4-32x, respectively. Source code is
 available at https://figshare.com/s/519a923fae8f386d7f5b
****

Laplacian Eigenspaces, Horocycles and Neuron Models on Hyperbolic Spaces

Ming-Xi Wang

We use hyperbolic Poisson kernel to construct the horocycle neuron model on hype
rbolic spaces, which is a spectral generalization of the classical neuron model.
 We prove a universal approximation theorem for horocycle neurons. As a corollar

y, this theorem leads to a state-of-the-art result on the expressivity of neuron s of the hyperbolic MLR. Our experiments get state-of-the-art results on the Poi ncare-embedding tree classification task and the two-dimensional visualization o f images.
****

Model-Based Robust Deep Learning: Generalizing to Natural, Out-of-Distribution D ata

Alexander Robey,Hamed Hassani,George J. Pappas

While deep learning (DL) has resulted in major breakthroughs in many application s, the frameworks commonly used in DL remain fragile to seemingly innocuous chan ges in the data. In response, adversarial training has emerged as a principled a pproach for improving the robustness of DL against norm-bounded perturbations. Despite this progress, DL is also known to be fragile to unbounded shifts in the data distribution due to many forms of natural variation, including changes in weather or lighting in images. However, there are remarkably few techniques tha t can address robustness to natural, out-of-distribution shifts in the data dist ribution in a general context. To address this gap, we propose a paradigm shift from perturbation-based adversarial robustness to model-based robust deep learn ing. Critical to our paradigm is to obtain models of natural variation, which v ary data over a range of natural conditions. Then by exploiting these models, w e develop three novel model-based robust training algorithms that improve the ro bustness of DL with respect to natural variation. Our extensive experiments show that across a variety of natural conditions in twelve distinct datasets, classi fiers trained with our algorithms significantly outperform classifiers trained v ia ERM, adversarial training, and domain adaptation techniques. Specifically, w hen training on ImageNet and testing on various subsets of ImageNet-c, our algor ithms improve over baseline methods by up to 30 percentage points in top-1 accur acy. Further, we show that our methods provide robustness (1) against natural, out-of-distribution data, (2) against multiple simultaneous distributional shift s, and (3) to domains entirely unseen during training.
****

Long-tail learning via logit adjustment

Aditya Krishna Menon,Sadeep Jayasumana,Ankit Singh Rawat,Himanshu Jain,Andreas V eit,Sanjiv Kumar

Real-world classification problems typically exhibit an imbalanced or long-taile d label distribution, wherein many labels have only a few associated samples. Th is poses a challenge for generalisation on such labels, and also makes naive le arning biased towards dominant labels. In this paper, we present a statistical framework that unifies and generalises several recent proposals to cope with the se challenges. Our framework revisits the classic idea of logit adjustment based on the label frequencies, which encourages a large relative margin between logi ts of rare positive versus dominant negative labels. This yields two techniques for long-tail learning, where such adjustment is either applied post-hoc to a t rained model, or enforced in the loss during training. These techniques are stat istically grounded, and practically effective on four real-world datasets with l ong-tailed label distributions.
****

Explainable Reinforcement Learning Through Goal-Based Interpretability

Gregory Bonaert,Youri Coppens,Denis Steckelmacher,Ann Nowe

Deep Reinforcement Learning agents achieve state-of-the-art performance in many tasks at the cost of making them black-boxes, hard to interpret and understand, making their use difficult in trusted applications, such as robotics or industri al applications. We introduce goal-based interpretability, where the agent produ ces goals which show the reason for its current actions (reach the current goal) and future goals indicate its desired future behavior without having to run the environment, a useful property in environments with no simulator. Additionally, in many environments, the goals can be visualized to make them easier to unders tand for non-experts. To have a goal-producing agent without requiring domain kn owledge, we use 2-layer hierarchical agents where the top layer produces goals a nd the bottom layer attempts to reach those goals.

Most classical reinforcement learning algorithms cannot be used train goal-producing hierarchical agents. We introduce a new algorithm to train these more interpretable agents, called HAC-General with Teacher, an extension of the Hindsight Actor-Critic (HAC) algorithm that adds 2 key improvements: (1) the goals now consist of a state s to be reached and a reward r to be collected, making it possible for the goal-producing policy to incentivize the goal-reaching policy to go through high-reward paths and (2) an expert teacher is leveraged to improve the training of the hierarchical agent, in a process similar but distinct to imitation learning and distillation. Contrarily to HAC, there is no requirement that environments need to provide the desired end state. Additionally, our experiments show that it has better performance and learns faster than HAC, and can solve environments that HAC fails to solve.
****

Fusion 360 Gallery: A Dataset and Environment for Programmatic CAD Reconstruction

Karl Willis,Yewen Pu,Jieliang Luo,Hang Chu,Tao Du,Joseph Lambourne,Armando Solar-Lezama,Wojciech Matusik

Parametric computer-aided design (CAD) is a standard paradigm used for the design of manufactured objects. CAD designers perform modeling operations, such as sketch and extrude, to form a construction sequence that makes up a final design. Despite the pervasiveness of parametric CAD and growing interest from the research community, a dataset of human designed 3D CAD construction sequences has not been available to-date. In this paper we present the Fusion 360 Gallery reconstruction dataset and environment for learning CAD reconstruction. We provide a dataset of 8,625 designs, comprising sequential sketch and extrude modeling operations, together with a complementary environment called the Fusion 360 Gym, to assist with performing CAD reconstruction. We outline a standard CAD reconstruction task, together with evaluation metrics, and present results from a novel method using neurally guided search to recover a construction sequence from a target geometry.
****

Teleport Graph Convolutional Networks

Hongyang Gao,Shuiwang Ji

We consider the limitations in message-passing graph neural networks. In message-passing operations, each node aggregates information from its neighboring nodes. To enlarge the receptive field, graph neural networks need to stack multiple message-passing graph convolution layers, which leads to the over-fitting issue and over-smoothing issue. To address these limitations, we propose a teleport graph convolution layer (TeleGCL) that uses teleport functions to enable each node to aggregate information from a much larger neighborhood. For each node, teleport functions select relevant nodes beyond the local neighborhood, thereby resulting in a larger receptive field. To apply our structure-aware teleport function, we propose a novel method to construct structural features for nodes in the graph. Based on our TeleGCL, we build a family of teleport graph convolutional networks. The empirical results on graph and node classification tasks demonstrate the effectiveness of our proposed methods.
****

Weighted Line Graph Convolutional Networks

Hongyang Gao,Shuiwang Ji

Line graphs have shown to be effective in improving feature learning in graph neural networks. Line graphs can encode topology information of their original graphs and provide a complementary representational perspective. In this work, we show that the encoded information in line graphs is biased. To overcome this issue, we propose a weighted line graph that corrects biases in line graphs by assigning normalized weights to edges. Based on our weighted line graphs, we develop a weighted line graph convolution layer that takes advantage of line graph structures for better feature learning. In particular, it performs message passing operations on both the original graph and its corresponding weighted line graph. To address efficiency issues in line graph neural networks, we propose to use an

incidence matrix to accurately compute the adjacency matrix of the weighted line graph, leading to dramatic reductions in computational resource usage. Experimental results on both real and simulated datasets demonstrate the effectiveness and efficiency of our proposed methods.
****
The Benefit of Distraction: Denoising Remote Vitals Measurements Using Inverse Attention
Ewa Magdalena Nowara,Daniel McDuff,Ashok Veeraraghavan
Attention is a powerful concept in computer vision. End-to-end networks that learn to focus selectively on regions of an image or video often perform strongly. However, other image regions, while not necessarily containing the signal of interest, may contain useful context. We present an approach that exploits the idea that statistics of noise may be shared between the regions that contain the signal of interest and those that do not. Our technique uses the inverse of an attention mask to generate a noise estimate that is then used to denoise temporal observations. We apply this to the task of camera-based physiological measurement. A convolutional attention network is used to learn which regions of a video contain the physiological signal and generate a preliminary estimate. A noise estimate is obtained by using the pixel intensities in the inverse regions of the learned attention mask, this in turn is used to refine the estimate of the physiological signal. We perform experiments on two large benchmark datasets and show that this approach produces state-of-the-art results, increasing the signal-to-noise ratio by up to 5.8 dB, reducing heart rate and breathing rate estimation error by as much as 30%, recovering subtle pulse waveform dynamics, and generalizing from RGB to NIR videos without retraining.
****
Locally Free Weight Sharing for Network Width Search
Xiu Su,Shan You,Tao Huang,Fei Wang,Chen Qian,Changshui Zhang,Chang Xu
Searching for network width is an effective way to slim deep neural networks with hardware budgets. With this aim, a one-shot supernet is usually leveraged as a performance evaluator to rank the performance \wrt~different width. Nevertheless, current methods mainly follow a manually fixed weight sharing pattern, which is limited to distinguish the performance gap of different width. In this paper, to better evaluate each width, we propose a locally free weight sharing strategy (CafeNet) accordingly. In CafeNet, weights are more freely shared, and each width is jointly indicated by its base channels and free channels, where free channels are supposed to locate freely in a local zone to better represent each width. Besides, we propose to further reduce the search space by leveraging our introduced FLOPs-sensitive bins. As a result, our CafeNet can be trained stochastically and get optimized within a min-min strategy. Extensive experiments on ImageNet, CIFAR-10, CelebA and MS COCO dataset have verified our superiority comparing to other state-of-the-art baselines. For example, our method can further boost the benchmark NAS network EfficientNet-B0 by 0.41\% via searching its width more delicately.
****
No Spurious Local Minima: on the Optimization Landscapes of Wide and Deep Neural Networks
Johannes Lederer
Empirical studies suggest that wide neural networks are comparably easy to optimize, but mathematical support for this observation is scarce. In this paper, we analyze the optimization landscapes of deep learning with wide networks. We prove especially that constraint and unconstraint empirical-risk minimization over such networks has no spurious local minima. Hence, our theories substantiate the common belief that increasing network widths not only improves the expressiveness of deep-learning pipelines but also facilitates their optimizations.
****
Accounting for Unobserved Confounding in Domain Generalization
Alexis Bellot,Mihaela van der Schaar
The ability to extrapolate, or generalize, from observed to new related environments is central to any form of reliable machine learning, yet most methods fail

when moving beyond i.i.d data. In some cases, the reason lies in a misappreciation of the causal structure that governs the data, and in particular as a consequence of the influence of unobserved confounders that drive changes in observed distributions and distort correlations. In this paper, we argue for defining generalization with respect to a broader class of distribution shifts (defined as arising from interventions in the underlying causal model), including changes in observed, unobserved and target variable distributions. We propose a new robust learning principle that may be paired with any gradient-based learning algorithm. This learning principle has explicit generalization guarantees, and relates robustness with certain invariances in the causal model, clarifying why, in some cases, test performance lags training performance. We demonstrate the empirical performance of our approach on healthcare data from different modalities, including image and speech data.
****

Dimension reduction as an optimization problem over a set of generalized functions
Rustem Takhanov
We reformulate unsupervised dimension reduction problem (UDR) in the language of tempered distributions, i.e. as a problem of approximating an empirical probability density function $p_{\rm{emp}}({\mathbf x})$ by another tempered distribution $q({\mathbf x})$ whose support is in a k-dimensional subspace. Thus, our problem is reduced to the minimization of the distance between q and $p_{\rm{emp}}$, $D(q, p_{\rm{emp}})$, over a pertinent set of generalized functions.

This infinite-dimensional formulation allows to establish a connection with another classical problem of data science --- the sufficient dimension reduction problem (SDR). Thus, an algorithm for the first problem induces an algorithm for the second and vice versa. In order to reduce an optimization problem over distributions to an optimization problem over ordinary functions we introduce a nonnegative penalty function $R(f)$ that ``forces'' the support of f to be k-dimensional.
Then we present an algorithm for minimization of $I(f)+\lambda R(f)$, based on the idea of two-step iterative computation, briefly described as a) an adaptation to real data and to fake data sampled around a k-dimensional subspace found at a previous iteration, b) calculation of a new k-dimensional subspace. We demonstrate the method on 4 examples (3 UDR and 1 SDR) using synthetic data and standard datasets.
****

Mutual Information State Intrinsic Control
Rui Zhao,Yang Gao,Pieter Abbeel,Volker Tresp,Wei Xu
Reinforcement learning has been shown to be highly successful at many challenging tasks. However, success heavily relies on well-shaped rewards. Intrinsically motivated RL attempts to remove this constraint by defining an intrinsic reward function. Motivated by the self-consciousness concept in psychology, we make a natural assumption that the agent knows what constitutes itself, and propose a new intrinsic objective that encourages the agent to have maximum control on the environment. We mathematically formalize this reward as the mutual information between the agent state and the surrounding state under the current agent policy. With this new intrinsic motivation, we are able to outperform previous methods, including being able to complete the pick-and-place task for the first time without using any task reward. A video showing experimental results is available at https://youtu.be/AUCwc9RThpk.
****

Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows
Kashif Rasul,Abdul-Saboor Sheikh,Ingmar Schuster,Urs M Bergmann,Roland Vollgraf
Time series forecasting is often fundamental to scientific and engineering problems and enables decision making. With ever increasing data set sizes, a trivial solution to scale up predictions is to assume independence between interacting time series. However, modeling statistical dependencies can improve accuracy and

enable analysis of interaction effects. Deep learning methods are well suited fo
r this problem, but multi-variate models often assume a simple parametric distri
bution and do not scale to high dimensions. In this work we model the multi-vari
ate temporal dynamics of time series via an autoregressive deep learning model,
where the data distribution is represented by a conditioned normalizing flow. T
his combination retains the power of autoregressive models, such as good perform
ance in extrapolation into the future, with the flexibility of flows as a genera
l purpose high-dimensional distribution model, while remaining computationally t
ractable. We show that it improves over the state-of-the-art for standard metric
s on many real-world data sets with several thousand interacting time-series.
****

A Mixture of Variational Autoencoders for Deep Clustering
Avi Caciularu,Jacob Goldberger
In this study, we propose a deep clustering algorithm that utilizes a variationa
l autoencoder (VAE) framework with a multi encoder-decoder neural architecture.
This setup enforces a complementary structure that guides the learned latent rep
resentations towards a more meaningful space arrangement. It differs from previo
us VAE-based clustering algorithms by employing a new generative model that uses
 multiple encoder-decoders.
We show that this modeling results in both better clustering capabilities and im
proved data generation. The proposed method is evaluated on standard datasets an
d is shown to outperform state-of-the-art deep clustering methods significantly.
****

Slice, Dice, and Optimize: Measuring the Dimension of Neural Network Class Manif
olds
Stanislav Fort,Ekin Dogus Cubuk,Surya Ganguli,Samuel Stern Schoenholz
Deep neural network classifiers naturally partition input space into regions bel
onging to different classes. The geometry of these class manifolds (CMs) is wide
ly studied and is intimately related to model performance; for example, the marg
in is defined via boundaries between these CMs. We present a simple technique to
 estimate the effective dimension of CMs as well as boundaries between multiple
CMs, by computing their intersection with random affine subspaces of varying dim
ension. We provide a theory for the technique and verify that our theoretical pr
edictions agree with measurements on real neural networks. Through extensive exp
eriments, we leverage this method to show deep connections between the geometry
of CMs, generalization, and robustness. In particular we investigate how CM dime
nsion depends on 1) the dataset, 2) architecture, 3) random initialization, 4) s
tage of training, 5) class, 6) ensemble size, 7) label randomization, 8) trainin
g set size, and 9) model robustness to data corruption. Together a picture emerg
es that well-performing, robust models have higher dimensional CMs than worse pe
rforming models. Moreover, we offer a unique perspective on ensembling via inter
sections of CMs. Our core code is available on Github (link in the PDF abstract)
.
****

Geometry-aware Instance-reweighted Adversarial Training
Jingfeng Zhang,Jianing Zhu,Gang Niu,Bo Han,Masashi Sugiyama,Mohan Kankanhalli
In adversarial machine learning, there was a common belief that robustness and a
ccuracy hurt each other. The belief was challenged by recent studies where we ca
n maintain the robustness and improve the accuracy. However, the other direction
, whether we can keep the accuracy and improve the robustness, is conceptually a
nd practically more interesting, since robust accuracy should be lower than stan
dard accuracy for any model. In this paper, we show this direction is also promi
sing. Firstly, we find even over-parameterized deep networks may still have insu
fficient model capacity, because adversarial training has an overwhelming smooth
ing effect. Secondly, given limited model capacity, we argue adversarial data sh
ould have unequal importance: geometrically speaking, a natural data point close
r to/farther from the class boundary is less/more robust, and the corresponding
adversarial data point should be assigned with larger/smaller weight. Finally, t
o implement the idea, we propose geometry-aware instance-reweighted adversarial
training, where the weights are based on how difficult it is to attack a natural

data point. Experiments show that our proposal boosts the robustness of standard adversarial training; combining two directions, we improve both robustness and accuracy of standard adversarial training.
****
Towards Impartial Multi-task Learning
Liyang Liu,Yi Li,Zhanghui Kuang,Jing-Hao Xue,Yimin Chen,Wenming Yang,Qingmin Liao,Wayne Zhang
Multi-task learning (MTL) has been widely used in representation learning. However, naively training all tasks simultaneously may lead to the partial training issue, where specific tasks are trained more adequately than others. In this paper, we propose to learn multiple tasks impartially. Specifically, for the task-shared parameters, we optimize the scaling factors via a closed-form solution, such that the aggregated gradient (sum of raw gradients weighted by the scaling factors) has equal projections onto individual tasks. For the task-specific parameters, we dynamically weigh the task losses so that all of them are kept at a comparable scale. Further, we find the above gradient balance and loss balance are complementary and thus propose a hybrid balance method to further improve the performance. Our impartial multi-task learning (IMTL) can be end-to-end trained without any heuristic hyper-parameter tuning, and is general to be applied on all kinds of losses without any distribution assumption. Moreover, our IMTL can converge to similar results even when the task losses are designed to have different scales, and thus it is scale-invariant. We extensively evaluate our IMTL on the standard MTL benchmarks including Cityscapes, NYUv2 and CelebA. It outperforms existing loss weighting methods under the same experimental settings.
****
On Learning Universal Representations Across Languages
Xiangpeng Wei,Rongxiang Weng,Yue Hu,Luxi Xing,Heng Yu,Weihua Luo
Recent studies have demonstrated the overwhelming advantage of cross-lingual pre-trained models (PTMs), such as multilingual BERT and XLM, on cross-lingual NLP tasks. However, existing approaches essentially capture the co-occurrence among tokens through involving the masked language model (MLM) objective with token-level cross entropy. In this work, we extend these approaches to learn sentence-level representations and show the effectiveness on cross-lingual understanding and generation. Specifically, we propose a Hierarchical Contrastive Learning (HiCTL) method to (1) learn universal representations for parallel sentences distributed in one or multiple languages and (2) distinguish the semantically-related words from a shared cross-lingual vocabulary for each sentence. We conduct evaluations on two challenging cross-lingual tasks, XTREME and machine translation. Experimental results show that the HiCTL outperforms the state-of-the-art XLM-R by an absolute gain of 4.2% accuracy on the XTREME benchmark as well as achieves substantial improvements on both of the high resource and low-resource English$\rightarrow$X translation tasks over strong baselines.
****
Isotropy in the Contextual Embedding Space: Clusters and Manifolds
Xingyu Cai,Jiaji Huang,Yuchen Bian,Kenneth Church
The geometric properties of contextual embedding spaces for deep language models such as BERT and ERNIE, have attracted considerable attention in recent years. Investigations on the contextual embeddings demonstrate a strong anisotropic space such that most of the vectors fall within a narrow cone, leading to high cosine similarities. It is surprising that these LMs are as successful as they are, given that most of their embedding vectors are as similar to one another as they are. In this paper, we argue that the isotropy indeed exists in the space, from a different but more constructive perspective. We identify isolated clusters and low dimensional manifolds in the contextual embedding space, and introduce tools to both qualitatively and quantitatively analyze them. We hope the study in this paper could provide insights towards a better understanding of the deep language models.
****
A Benchmark for Voice-Face Cross-Modal Matching and Retrieval
Chuyuan Xiong,Deyuan Zhang,Tao Liu,Xiaoyong Du,Jiankun Tian,Songyan Xue

Cross-modal associations between a person's voice and face can be learned algori thmically, and this is a useful functionality in many audio and visual applicati ons. The problem can be defined as two tasks: voice-face matching and retrieval. Recently, this topic has attracted much research attention, but it is still in its early stages of development, and evaluation protocols and test schemes need to be more standardized. Performance metrics for different subtasks are also sca rce, and a benchmark for this problem needs to be established. In this paper, a baseline evaluation framework is proposed for voice-face matching and retrieval tasks. Test confidence is analyzed, and a confidence interval for estimated accu racy is proposed. Various state-of-the-art performances with high test confidenc e are achieved on a series of subtasks using the baseline method (called TriNet) included in this framework. The source code will be published along with the p aper. The results of this study can provide a basis for future research on voice -face cross-modal learning.
****

Federated Learning with Decoupled Probabilistic-Weighted Gradient Aggregation
Jian-hui Duan,Wenzhong Li,Sanglu Lu
In the federated learning paradigm, multiple mobile clients train local models independently based on datasets generated by edge devices, and the server aggreg ates parameters/gradients from local models to form a global model. However, e xisting model aggregation approaches suffer from high bias on both data distribu tion and parameter distribution for non-IID datasets, which result in sever e accuracy drop for increasing number of heterogeneous clients. In this paper , we proposed a novel decoupled probabilistic-weighted gradient aggregation appr oach called FeDEC for federated learning. The key idea is to optimize gradient parameters and statistical parameters in a decoupled way, and aggregate the para meters from local models with probabilistic weights to deal with the heterogenei ty of clients. Since the overall dataset is unaccessible by the central server, we introduce a variational inference method to derive the optimal probabilist ic weights to minimize statistical bias. We further prove the convergence boun d of the proposed approach. Extensive experiments using mainstream convolution al neural network models based on three federated datasets show that FeDEC sign ificantly outperforms the state-of-the-arts in terms of model accuracy and train ing efficiency.
****

Continual Memory: Can We Reason After Long-Term Memorization?
Zhu Zhang,Chang Zhou,Zhou Zhao,Zhijie Lin,Jingren Zhou,Hongxia Yang
Existing reasoning tasks often follow the setting of ‘’end-to-end reasoning'', w hich has an important assumption that the input contents can be always accessed while reasoning. However, human beings frequently adopt another reasoning settin g in daily life, referred to ‘’reasoning after memorizing''. Concretely, human b eings have the ability to unconsciously memorize their experiences within limite d memory capacity, from which they can recall and respond to subsequent tasks. I n this setting, the input contents are no longer available during reasoning, thu s we need to compress and memorize the input stream in one pass, trying to answe r general queries that are unseen before. Memory augmented neural networks intro duce a write-read memory to perform such human-like memorization and reasoning, but they continually update the memory from current information and inevitably f orget the early contents, failing to answer the queries relevant to early inform ation. In this paper, we propose the Continual Memory (CM) to explore this abil ity of reasoning after long-term memorization. To alleviate the gradual forgett ing of early information, we develop self-supervised memorization training with item-level and sequence-level objectives. We demonstrate several interesting cha racteristics of our continual memory via synthetic data, and evaluate its perfor mance by several downstream tasks, including long-term text QA, long-term video QA and recommendation with long sequences.
****

MoPro: Webly Supervised Learning with Momentum Prototypes
Junnan Li,Caiming Xiong,Steven Hoi
We propose a webly-supervised representation learning method that does not suffe

r from the annotation unscalability of supervised learning, nor the computation unscalability of self-supervised learning. Most existing works on webly-supervised representation learning adopt a vanilla supervised learning method without accounting for the prevalent noise in the training data, whereas most prior methods in learning with label noise are less effective for real-world large-scale noisy data. We propose momentum prototypes (MoPro), a simple contrastive learning method that achieves online label noise correction, out-of-distribution sample removal, and representation learning. MoPro achieves state-of-the-art performance on WebVision, a weakly-labeled noisy dataset. MoPro also shows superior performance when the pretrained model is transferred to down-stream image classification and detection tasks. It outperforms the ImageNet supervised pretrained model by +10.5 on 1-shot classification on VOC, and outperforms the best self-supervised pretrained model by +17.3 when finetuned on 1% of ImageNet labeled samples. Furthermore, MoPro is more robust to distribution shifts. Code and pretrained models are available at https://github.com/salesforce/MoPro.
****
Learning from Noisy Data with Robust Representation Learning
Junnan Li,Caiming Xiong,Steven Hoi
Learning from noisy data has attracted much attention, where most methods focus on label noise. In this work, we propose a new framework which simultaneously addresses three types of noise commonly seen in real-world data: label noise, out-of-distribution input, and input corruption. In contrast to most existing methods, we combat noise by learning robust representation. Specifically, we embed images into a low-dimensional subspace by training an autoencoder on the deep features. We regularize the geometric structure of the subspace with robust contrastive learning, which includes an unsupervised consistency loss and a supervised mixup prototypical loss. Furthermore, we leverage the structure of the learned subspace for noise cleaning, by aggregating information from neighboring samples. Experiments on multiple benchmarks demonstrate state-of-the-art performance of our method and robustness of the learned representation. Our code will be released.
****
Effective Subspace Indexing via Interpolation on Stiefel and Grassmann manifolds
Wenqing Hu,Tiefeng Jiang,Zhu Li
We propose a novel local Subspace Indexing Model with Interpolation (SIM-I) for low-dimensional embedding of image datasets. Our SIM-I is constructed via two steps: in the first step we build a piece-wise linear affinity-aware subspace model under a given partition of the dataset; in the second step we interpolate between several adjacent linear subspace models constructed previously using the `"center of mass" calculation on Stiefel and Grassmann manifolds. The resulting subspace indexing model built by SIM-I is a globally non-linear low-dimensional embedding of the original data set. Furthermore, the interpolation step produces a `"smoothed” version of the piece-wise linear embedding mapping constructed in the first step, and can be viewed as a regularization procedure. We provide experimental results validating the effectiveness of SIM-I, that improves PCA recovery for SIFT dataset and nearest-neighbor classification success rates for MNIST and CIFAR-10 datasets.
****
MeshMVS: Multi-view Stereo Guided Mesh Reconstruction
Rakesh Shrestha,Zhiwen Fan,Siyu Zhu,Zuozhuo Dai,Qingkun Su,Ping Tan
Deep learning based 3D shape generation methods generally utilize latent features extracted from color images to encode the objects' semantics and guide the shape generation process. These color image semantics only implicitly encode 3D information, potentially limiting the accuracy of the generated shapes. In this paper we propose a multi-view mesh generation method which incorporates geometry information in the color images explicitly by using the features from intermediate 2.5D depth representations of the input images and regularizing the 3D shapes against these depth images. Our system first predicts a coarse 3D volume from the color images by probabilistically merging voxel occupancy grids from individual views. Depth images corresponding to the multi-view color images are predicted

which along with the rendered depth images of the coarse shape are used as a con
trastive input whose features guide the refinement of the coarse shape through a
 series of graph convolution networks. Attention-based multi-view feature poolin
g is proposed to fuse the contrastive depth features from different viewpoints w
hich are fed to the graph convolution networks.
We validate the proposed multi-view mesh generation method on ShapeNet, where we
 obtain a significant improvement with 34% decrease in chamfer distance to groun
d truth and 14% increase in the F1-score compared with the state-of-the-art mult
i-view shape generation method.
****

GraphCodeBERT: Pre-training Code Representations with Data Flow
Daya Guo,Shuo Ren,Shuai Lu,Zhangyin Feng,Duyu Tang,Shujie LIU,Long Zhou,Nan Duan
,Alexey Svyatkovskiy,Shengyu Fu,Michele Tufano,Shao Kun Deng,Colin Clement,Dawn
Drain,Neel Sundaresan,Jian Yin,Daxin Jiang,Ming Zhou
Pre-trained models for programming language have achieved dramatic empirical imp
rovements on a variety of code-related tasks such as code search, code completio
n, code summarization, etc. However, existing pre-trained models regard a code s
nippet as a sequence of tokens, while ignoring the inherent structure of code, w
hich provides crucial code semantics and would enhance the code understanding pr
ocess. We present GraphCodeBERT, a pre-trained model for programming language th
at considers the inherent structure of code. Instead of taking syntactic-level s
tructure of code like abstract syntax tree (AST), we use data flow in the pre-tr
aining stage, which is a semantic-level structure of code that encodes the relat
ion of "where-the-value-comes-from" between variables. Such a semantic-level str
ucture is neat and does not bring an unnecessarily deep hierarchy of AST, the pr
operty of which makes the model more efficient. We develop GraphCodeBERT based o
n Transformer. In addition to using the task of masked language modeling, we int
roduce two structure-aware pre-training tasks. One is to predict code structure
 edges, and the other is to align representations between source code and code st
ructure. We implement the model in an efficient way with a graph-guided masked a
ttention function to incorporate the code structure. We evaluate our model on fo
ur tasks, including code search, clone detection, code translation, and code ref
inement. Results show that code structure and newly introduced pre-training task
s can improve GraphCodeBERT and achieves state-of-the-art performance on the fou
r downstream tasks. We further show that the model prefers structure-level atten
tions over token-level attentions in the task of code search.
****

Towards Robust Graph Neural Networks against Label Noise
Jun Xia,Haitao Lin,Yongjie Xu,Lirong Wu,Zhangyang Gao,Siyuan Li,Stan Z. Li
Massive labeled data have been used in training deep neural networks, thus label
 noise has become an important issue therein. Although learning with noisy labe
ls has made great progress on image datasets in recent years, it has not yet bee
n studied in connection with utilizing GNNs to classify graph nodes. In this pap
er, we proposed a method, named LPM, to address the problem using Label Propaga
tion (LP) and Meta learning. Different from previous methods designed for image
 datasets, our method is based on a special attribute (label smoothness) of grap
h-structured data, i.e., neighboring nodes in a graph tend to have the same labe
l. A pseudo label is computed from the neighboring labels for each node in the
training set using LP; meta learning is utilized to learn a proper aggregation o
f the original and pseudo label as the final label. Experimental results demons
trate that LPM outperforms state-of-the-art methods in graph node classification
 task with both synthetic and real-world label noise. Source code to reproduce a
ll results will be released.
****

To Learn Effective Features: Understanding the Task-Specific Adaptation of MAML
Zhijie Lin,Zhou Zhao,Zhu Zhang,Huai Baoxing,Jing Yuan
Meta learning, an effective way for learning unseen tasks with few samples, is a
n important research
area in machine learning.
Model Agnostic Meta-Learning~(MAML)~(\cite{finn2017model}) is one of the most we

ll-known gradient-based meta learning algorithms, that learns
the meta-initialization through the inner and outer optimization loop.
The inner loop is to perform fast adaptation in several gradient update steps wi
th the support datapoints,
while the outer loop to generalize the updated model to the query datapoints.
Recently, it has been argued that instead of rapid learning and adaptation, the
learned meta-initialization through MAML
has already absorbed the high-quality features prior, where the task-specific he
ad at training
facilitates the feature learning.
In this work, we investigate the impact of the task-specific adaptation of MAML
and discuss the general formula for
other gradient-based and metric-based meta-learning approaches.
From our analysis, we further devise the Random Decision Planes~(RDP) algorithm
to find a suitable linear classifier
without any gradient descent step and the Meta Contrastive Learning~(MCL) algori
thm to exploit the inter-samples relationship
instead of the expensive inner-loop adaptation.
We conduct sufficient experiments on various datasets to explore our proposed al
gorithms.
****

Enjoy Your Editing: Controllable GANs for Image Editing via Latent Space Navigat
ion
Peiye Zhuang,Oluwasanmi O Koyejo,Alex Schwing
Controllable semantic image editing enables a user to change entire image attrib
utes with a few clicks, e.g., gradually making a summer scene look like it was t
aken in winter. Classic approaches for this task use a Generative Adversarial Ne
t (GAN) to learn a latent space and suitable latent-space transformations. Howev
er, current approaches often suffer from attribute edits that are entangled, glo
bal image identity changes, and diminished photo-realism. To address these conce
rns, we learn multiple attribute transformations simultaneously, integrate attri
bute regression into the training of transformation functions, and apply a conte
nt loss and an adversarial loss that encourages the maintenance of image identit
y and photo-realism. We propose quantitative evaluation strategies for measuring
 controllable editing performance, unlike prior work, which primarily focuses on
 qualitative evaluation. Our model permits better control for both single- and m
ultiple-attribute editing while preserving image identity and realism during tra
nsformation. We provide empirical results for both natural and synthetic images,
 highlighting that our model achieves state-of-the-art performance for targeted
image manipulation.
****

Deep Jump Q-Evaluation for Offline Policy Evaluation in Continuous Action Space
Hengrui Cai,Chengchun Shi,Rui Song,Wenbin Lu
We consider off-policy evaluation (OPE) in continuous action domains, such as dy
namic pricing and personalized dose finding. In OPE, one aims to learn the value
 under a new policy using historical data generated by a different behavior poli
cy. Most existing works on OPE focus on discrete action domains. To handle conti
nuous action space, we develop a brand-new deep jump Q-evaluation method for OPE
. The key ingredient of our method lies in adaptively discretizing the action sp
ace using deep jump Q-learning. This allows us to apply existing OPE methods in
discrete domains to handle continuous actions. Our method is further justified b
y theoretical results, synthetic and real datasets.

****

Fourier Neural Operator for Parametric Partial Differential Equations
Zongyi Li,Nikola Borislavov Kovachki,Kamyar Azizzadenesheli,Burigede liu,Kaushik
 Bhattacharya,Andrew Stuart,Anima Anandkumar
The classical development of neural networks has primarily focused on learning m
appings between finite-dimensional Euclidean spaces. Recently, this has been ge
neralized to neural operators that learn mappings between function spaces. For p

artial differential equations (PDEs), neural operators directly learn the mappin g from any functional parametric dependence to the solution. Thus, they learn an entire family of PDEs, in contrast to classical methods which solve one instanc e of the equation. In this work, we formulate a new neural operator by parameter izing the integral kernel directly in Fourier space, allowing for an expressive and efficient architecture. We perform experiments on Burgers' equation, Darcy f low, and Navier-Stokes equation. The Fourier neural operator is the first ML-bas ed method to successfully model turbulent flows with zero-shot super-resolution. It is up to three orders of magnitude faster compared to traditional PDE solver s. Additionally, it achieves superior accuracy compared to previous learning-bas ed solvers under fixed resolution.

****

Improved Uncertainty Post-Calibration via Rank Preserving Transforms

Yu Bai,Tengyu Ma,Huan Wang,Caiming Xiong

Modern machine learning models with high accuracy often exhibit poor uncertainty calibration: the output probabilities of the model do not reflect its accuracy, and tend to be over-confident. Existing post-calibration methods such as temper ature scaling recalibrate a trained model using rather simple calibrators with o ne or few parameters, which can have a rather limited capacity. In this paper, w e propose Neural Rank Preserving Transforms (NRPT), a new post-calibration metho d that adjusts the output probabilities of a trained classifier using a calibrat or of higher capacity, while maintaining its prediction accuracy. NRPT learns a calibrator that preserves the rank of the probabilities through general monotoni c transforms, individualizes to the original input, and allows learning with any loss function that encourages calibration. We show experimentally that NRPT imp roves the expected calibration error (ECE) significantly over existing post-cali bration methods such as (local) temperature scaling on large-scale image and tex t classification tasks. The performance of NRPT can further match ensemble metho ds such as deep ensembles, while being much more parameter-efficient. We further demonstrate the improved calibration ability of NRPT beyond the ECE metric, suc h as accuracy among top-confidence predictions, as well as optimizing the tradeo ff between calibration and sharpness.

****

Training BatchNorm and Only BatchNorm: On the Expressive Power of Random Feature s in CNNs

Jonathan Frankle,David J. Schwab,Ari S. Morcos

A wide variety of deep learning techniques from style transfer to multitask lear ning rely on training affine transformations of features. Most prominent among t hese is the popular feature normalization technique BatchNorm, which normalizes activations and then subsequently applies a learned affine transform. In this pa per, we aim to understand the role and expressive power of affine parameters use d to transform features in this way. To isolate the contribution of these parame ters from that of the learned features they transform, we investigate the perfor mance achieved when training only these parameters in BatchNorm and freezing all weights at their random initializations. Doing so leads to surprisingly high pe rformance considering the significant limitations that this style of training im poses. For example, sufficiently deep ResNets reach 82% (CIFAR-10) and 32% (Imag eNet, top-5) accuracy in this configuration, far higher than when training an eq uivalent number of randomly chosen parameters elsewhere in the network. BatchNor m achieves this performance in part by naturally learning to disable around a th ird of the random features. Not only do these results highlight the expressive p ower of affine parameters in deep learning, but - in a broader sense - they char acterize the expressive power of neural networks constructed simply by shifting and rescaling random features.

****

Information Laundering for Model Privacy

Xinran Wang,Yu Xiang,Jun Gao,Jie Ding

In this work, we propose information laundering, a novel framework for enhancing model privacy. Unlike data privacy that concerns the protection of raw data inf ormation, model privacy aims to protect an already-learned model that is to be d

eployed for public use. The private model can be obtained from general learning methods, and its deployment means that it will return a deterministic or random response for a given input query. An information-laundered model consists of probabilistic components that deliberately maneuver the intended input and output for queries of the model, so the model's adversarial acquisition is less likely. Under the proposed framework, we develop an information-theoretic principle to quantify the fundamental tradeoffs between model utility and privacy leakage and derive the optimal design.
****
USING OBJECT-FOCUSED IMAGES AS AN IMAGE AUGMENTATION TECHNIQUE TO IMPROVE THE ACCURACY OF IMAGE-CLASSIFICATION MODELS WHEN VERY LIMITED DATA SETS ARE AVAILABLE
Ahmad Melhem Hammoud,Ahmad Rabih Ghandour
Today, many of the machine learning models are extremely data hungry. On the other hand, the accuracy of the algorithms used is very often affected by the amount of the training data available, which is, unfortunately, rarely abundant. Fortunately, image augmentation is one of the very powerful techniques that can be used by computer-vision engineers to expand their existing image data sets. This paper presents an innovative way for creating a variation of existing images and introduces the idea of using an Object-Focused Image (OFI). This is when an image includes only the labeled object and everything else is made transparent. The objective of OFI method is to expand the existing image data set and hence improve the accuracy of the model used to classify images. This paper also elaborates on the OFI approach and compares the accuracy of five different models with the same network design and settings but with different content of the training data set. The experiments presented in this paper show that using OFIs along with the original images can lead to an increase in the validation accuracy of the used model. In fact, when the OFI technique is used, the number of the images supplied nearly doubles.
****
Robust Learning for Congestion-Aware Routing
Sreenivas Gollapudi,Kostas Kollias,Benjamin Plaut,Ameya Velingker
We consider the problem of routing users through a network with unknown congestion functions over an infinite time horizon. On each time step t, the algorithm receives a routing request and must select a valid path. For each edge e in the selected path, the algorithm incurs a cost $c_e^t = f_e(x_e^t) + \eta_e^t$, where x_e^t is the flow on edge e at time t, f_e is the congestion function, and η_e^t is a noise sample drawn from an unknown distribution. The algorithm observes c_e^t, and can use this observation in future routing decisions. The routing requests are supplied adversarially.

We present an algorithm with cumulative regret $\tilde{O}(|E| t^{2/3})$, where the regret on each time step is defined as the difference between the total cost incurred by our chosen path and the minimum cost among all valid paths. Our algorithm has space complexity $O(|E| t^{1/3})$ and time complexity $O(|E| \log t)$. We also validate our algorithm empirically using graphs from New York City road networks.
****
Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis
Bingchen Liu,Yizhe Zhu,Kunpeng Song,Ahmed Elgammal
Training Generative Adversarial Networks (GAN) on high-fidelity images usually requires large-scale GPU-clusters and a vast number of training images. In this paper, we study the few-shot image synthesis task for GAN with minimum computing cost. We propose a light-weight GAN structure that gains superior quality on 1024^2 resolution. Notably, the model converges from scratch with just a few hours of training on a single RTX-2080 GPU, and has a consistent performance, even with less than 100 training samples. Two technique designs constitute our work, a skip-layer channel-wise excitation module and a self-supervised discriminator trained as a feature-encoder. With thirteen datasets covering a wide variety of image domains (The datasets and code are available at https://github.com/odegeasslb

c/FastGAN-pytorch), we show our model's superior performance compared to the sta
te-of-the-art StyleGAN2, when data and computing budget are limited.
****
Task Calibration for Distributional Uncertainty in Few-Shot Classification
Sungnyun Kim,Se-Young Yun
As numerous meta-learning algorithms improve performance when solving few-shot c
lassification problems for practical applications, accurate prediction of uncert
ainty, though challenging, has been considered essential. In this study, we cont
emplate modeling uncertainty in a few-shot classification framework and propose
a straightforward method that appropriately predicts task uncertainty. We suppos
e that the random sampling of tasks can generate those in which it may be hard f
or the model to infer the queries from the support examples. Specifically, measu
ring the distributional mismatch between support and query sets via class-wise s
imilarities, we propose novel meta-training that lets the model predict with car
eful confidence. Moreover, our method is algorithm-agnostic and readily expanded
 to include a range of meta-learning models. Through extensive experiments inclu
ding dataset shift, we present that our training strategy helps the model avoid
being indiscriminately confident, and thereby, produce calibrated classification
 results without the loss of accuracy.
****
Layer-wise Adversarial Defense: An ODE Perspective
Zonghan Yang,Yang Liu,Chenglong Bao,Zuoqiang Shi
Deep neural networks are observed to be fragile against adversarial attacks, whi
ch have dramatically limited their practical applicability. On improving model r
obustness, the adversarial training techniques have proven effective and gained
increasing attention from research communities. Existing adversarial training ap
proaches mainly focus on perturbations to inputs, while the effect of the pertur
bations in hidden layers remains underexplored. In this work, we propose layer-w
ise adversarial defense which improves adversarial training by a noticeable marg
in. The basic idea of our method is to strengthen all of the hidden layers with
perturbations that are proportional to the back-propagated gradients. In order t
o study the layer-wise neural dynamics, we formulate our approach from the persp
ective of ordinary differential equations (ODEs) and build up its extended relat
ionship with conventional adversarial training methods, which tightens the relat
ionship between neural networks and ODEs. In the implementation, we propose two
different training algorithms by discretizing the ODE model with the Lie-Trotter
 and the Strang-Marchuk splitting schemes from the operator-splitting theory. Ex
periments on CIFAR-10 and CIFAR-100 benchmarks show that our methods consistentl
y improve adversarial model robustness on top of widely-used strong adversarial
training techniques.
****
Direct Evolutionary Optimization of Variational Autoencoders with Binary Latents
Enrico Guiraud,Jakob Drefs,Jorg Lucke
Discrete latent variables are considered important to model the generation proce
ss of real world data, which has motivated research on Variational Autoencoders
(VAEs) with discrete latents. However, standard VAE training is not possible in
this case, which has motivated different strategies to manipulate discrete distr
ibutions in order to train discrete VAEs similarly to conventional ones.
Here we ask if it is also possible to keep the discrete nature of the latents fu
lly intact by applying a direct discrete optimization for the encoding model. Th
e studied approach is consequently strongly diverting from standard VAE training
 by altogether sidestepping absolute standard VAE mechanisms such as sampling ap
proximation, reparameterization trick and amortization.

Discrete optimization is realized in a variational setting using truncated poste
riors in conjunction with evolutionary algorithms (using a recently suggested ap
proach). For VAEs with binary latents, we first show how such a discrete variati
onal method (A)~ties into gradient ascent for network weights and (B)~uses the d
ecoder network to select latent states for training.

More conventional amortized training is, as may be expected, more efficient than direct discrete optimization, and applicable to large neural networks.
However, we here find direct optimization to be efficiently scalable to hundreds of latent variables using smaller networks.
More importantly, we find the effectiveness of direct optimization to be highly competitive in 'zero-shot' learning (where high effectiveness for small networks is required).
In contrast to large supervised neural networks, the here investigated VAEs can denoise a single image without previous training on clean data and/or training on large image datasets.

More generally, the studied approach shows that training of VAEs is indeed possible without sampling-based approximation and reparameterization, which may be interesting for the analysis of VAE training in general. In the regime of few data, direct optimization, furthermore, makes VAEs competitive for denoising where they have previously been outperformed by non-generative approaches.
****
The Compact Support Neural Network
Adrian Barbu,Hongyu Mou
Neural networks are popular and useful in many fields, but they have the problem of giving high confidence responses for examples that are away from the training data. This results in the neural networks being very confident while making gross mistakes, thus limiting their reliability for safety critical applications such as autonomous driving, space exploration, etc.
In this paper, we present a more generic neuron formulation that contains the standard dot-product based neuron and the RBF neuron as two extreme cases of a shape parameter. Using ReLU as the activation function we obtain a novel neuron that compact support, which means its output is zero outside a bounded domain. We also show how to avoid difficulties in training a neural network with such neurons, by starting with a trained standard neural network and gradually increasing the shape parameter to the desired value.
Through experiments on standard benchmark datasets, we show the promise of the proposed approach, in that it can have good prediction on in-distribution samples, and it can consistently detect and have low confidence on out of distribution samples.
****
Rethinking the Pruning Criteria for Convolutional Neural Network
Zhongzhan Huang,Xinjiang Wang,Ping Luo
Channel pruning is a popular technique for compressing convolutional neural networks (CNNs), and various pruning criteria have been proposed to remove the redundant filters of CNNs. From our comprehensive experiments, we find some blind spots on pruning criteria: (1) Similarity: There are some strong similarities among several primary pruning criteria that are widely cited and compared. According to these criteria, the ranks of filters' importance in a convolutional layer are almost the same, resulting in similar pruned structures. (2) Applicability: For a large network (each convolutional layer has a large number of filters), some criteria can not distinguish the network redundancy well from their measured filters' importance. In this paper, we theoretically validate these two findings with our assumption that the well-trained convolutional filters in each layer approximately follow a Gaussian-alike distribution. This assumption is verified through systematic and extensive statistical tests.
****
UPDeT: Universal Multi-agent RL via Policy Decoupling with Transformers
Siyi Hu,Fengda Zhu,Xiaojun Chang,Xiaodan Liang
Recent advances in multi-agent reinforcement learning have been largely limited in training one model from scratch for every new task. The limitation is due to the restricted model architecture related to fixed input and output dimensions. This hinders the experience accumulation and transfer of the learned agent over

tasks with diverse levels of difficulty (e.g. 3 vs 3 or 5 vs 6 multi-agent games). In this paper, we make the first attempt to explore a universal multi-agent reinforcement learning pipeline, designing one single architecture to fit tasks with the requirement of different observation and action configurations. Unlike previous RNN-based models, we utilize a transformer-based model to generate a flexible policy by decoupling the policy distribution from the intertwined input observation with an importance weight measured by the merits of the self-attention mechanism. Compared to a standard transformer block, the proposed model, named as Universal Policy Decoupling Transformer (UPDeT), further relaxes the action restriction and makes the multi-agent task's decision process more explainable. UPDeT is general enough to be plugged into any multi-agent reinforcement learning pipeline and equip them with strong generalization abilities that enables the handling of multiple tasks at a time. Extensive experiments on large-scale SMAC multi-agent competitive games demonstrate that the proposed UPDeT-based multi-agent reinforcement learning achieves significant results relative to state-of-the-art approaches, demonstrating advantageous transfer capability in terms of both performance and training speed (10 times faster).
****

Towards Counteracting Adversarial Perturbations to Resist Adversarial Examples
Haimin ZHANG,Min Xu
Studies show that neural networks are susceptible to adversarial attacks. This exposes a potential threat to neural network-based artificial intelligence systems. We observe that the probability of the correct result outputted by the network increases by applying small perturbations generated for class labels other than the original predicted one to adversarial examples. Based on this observation, we propose a method of counteracting adversarial perturbations to resist adversarial examples. In our method, we randomly select a number of class labels and generate small perturbations for these selected labels. The generated perturbations are added together and then clamped to a specified space. The obtained perturbation is finally added to the adversarial example to counteract the adversarial perturbation contained in the example. The proposed method is applied at inference time and does not require retraining or finetuning the model. We validate the proposed method on CIFAR-10 and CIFAR-100. The experimental results demonstrate that our method effectively improves the defense performance of the baseline methods, especially against strong adversarial examples generated using more iterations.
****

Compositional Video Synthesis with Action Graphs
Amir Bar,Roei Herzig,Xiaolong Wang,Gal Chechik,Trevor Darrell,Amir Globerson
Videos of actions are complex signals, containing rich compositional structure. Current video generation models are limited in their ability to generate such videos. To address this challenge, we introduce a generative model (AG2Vid) that can be conditioned on an Action Graph, a structure that naturally represents the dynamics of actions and interactions between objects. Our AG2Vid model disentangles appearance and position features, allowing for more accurate generation. AG2Vid is evaluated on the CATER and Something-Something datasets and outperforms other baselines. Finally, we show how Action Graphs can be used for generating novel compositions of actions.
****

Free Lunch for Few-shot Learning: Distribution Calibration
Shuo Yang,Lu Liu,Min Xu
Learning from a limited number of samples is challenging since the learned model can easily become overfitted based on the biased distribution formed by only a few training examples. In this paper, we calibrate the distribution of these few-sample classes by transferring statistics from the classes with sufficient examples. Then an adequate number of examples can be sampled from the calibrated distribution to expand the inputs to the classifier. We assume every dimension in the feature representation follows a Gaussian distribution so that the mean and the variance of the distribution can borrow from that of similar classes whose statistics are better estimated with an adequate number of samples. Our method can

be built on top of off-the-shelf pretrained feature extractors and classificati on models without extra parameters. We show that a simple logistic regression cl assifier trained using the features sampled from our calibrated distribution can outperform the state-of-the-art accuracy on three datasets (~5% improvement on miniImageNet compared to the next best). The visualization of these generated fe atures demonstrates that our calibrated distribution is an accurate estimation.
****

Learning Stochastic Behaviour from Aggregate Data
Shaojun Ma,Shu Liu,Hongyuan Zha,Hao-Min Zhou
Learning nonlinear dynamics from aggregate data is a challenging problem since t he full trajectory of each individual is not available, namely, the individual o bserved at one time point may not be observed at next time point, or the identit y of individual is unavailable. This is in sharp contrast to learning dynamics w ith trajectory data, on which the majority of existing methods are based. We pro pose a novel method using the weak form of Fokker Planck Equation (FPE) to descr ibe density evolution of data in a sampling form, which is then combined with Wa sserstein generative adversarial network (WGAN) in training process. In such a s ample-based framework we are able to study nonlinear dynamics from aggregate dat a without solving the partial differential equation (PDE). The model can also ha ndle high dimensional cases with the help of deep neural networks. We demonstrat e our approach in the context of a series of synthetic and real-world data sets.
****

Targeted Attack against Deep Neural Networks via Flipping Limited Weight Bits
Jiawang Bai,Baoyuan Wu,Yong Zhang,Yiming Li,Zhifeng Li,Shu-Tao Xia
To explore the vulnerability of deep neural networks (DNNs), many attack paradig ms have been well studied, such as the poisoning-based backdoor attack in the tr aining stage and the adversarial attack in the inference stage. In this paper, w e study a novel attack paradigm, which modifies model parameters in the deployme nt stage for malicious purposes. Specifically, our goal is to misclassify a spec ific sample into a target class without any sample modification, while not signi ficantly reduce the prediction accuracy of other samples to ensure the stealthin ess. To this end, we formulate this problem as a binary integer programming (BIP), since the parameters are stored as binary bits ($i.e.$, 0 and 1) in the memor y. By utilizing the latest technique in integer programming, we equivalently ref ormulate this BIP problem as a continuous optimization problem, which can be eff ectively and efficiently solved using the alternating direction method of multip liers (ADMM) method. Consequently, the flipped critical bits can be easily deter mined through optimization, rather than using a heuristic strategy. Extensive ex periments demonstrate the superiority of our method in attacking DNNs.
****

FedBN: Federated Learning on Non-IID Features via Local Batch Normalization
Xiaoxiao Li,Meirui JIANG,Xiaofei Zhang,Michael Kamp,Qi Dou
The emerging paradigm of federated learning (FL) strives to enable collaborative training of deep models on the network edge without centrally aggregating raw d ata and hence improving data privacy. In most cases, the assumption of independe nt and identically distributed samples across local clients does not hold for fe derated learning setups. Under this setting, neural network training performance may vary significantly according to the data distribution and even hurt trainin g convergence. Most of the previous work has focused on a difference in the dist ribution of labels or client shifts. Unlike those settings, we address an import ant problem of FL, e.g., different scanners/sensors in medical imaging, differen t scenery distribution in autonomous driving (highway vs. city), where local cli ents store examples with different distributions compared to other clients, whic h we denote as feature shift non-iid. In this work, we propose an effective meth od that uses local batch normalization to alleviate the feature shift before ave raging models. The resulting scheme, called FedBN, outperforms both classical Fe dAvg, as well as the state-of-the-art for non-iid data (FedProx) on our extensiv e experiments. These empirical results are supported by a convergence analysis t hat shows in a simplified setting that FedBN has a faster convergence rate than FedAvg. Code is available at https://github.com/med-air/FedBN.

****
MixCon: Adjusting the Separability of Data Representations for Harder Data Recov
ery
Xiaoxiao Li,YANGSIBO HUANG,Binghui Peng,Zhao Song,Kai Li
To address the issue that deep neural networks (DNNs) are vulnerable to model in
version attacks, we design an objective function to adjust the separability of t
he hidden data representations as a way to control the trade-off between data ut
ility and vulnerability to inversion attacks. Our method is motivated by the the
oretical insights of data separability in neural networking training and results
 on the hardness of model inversion. Empirically, we show that there exist sweet
-spots by adjusting the separability of data representation, such that it is dif
ficult to recover data during inference while maintaining data utility.
****
Gradient Based Memory Editing for Task-Free Continual Learning
Xisen Jin,Junyi Du,Xiang Ren
Prior work on continual learning often operate in a “task-aware” manner, by assu
ming that the task boundaries and identifies of the data examples are known at a
ll times. While in practice, it is rarely the case that such information are exp
osed to the methods (i.e., thus called “task-free”)–a setting that is relatively
 underexplored. Recent attempts on task-free continual learning build on previou
s memory replay methods and focus on developing memory construction and replay s
trategies such that model performance over previously seen examples can be best
retained. In this paper, looking from a complementary angle, we propose a novel
approach to “edit” memory examples so that the edited memory can better retain p
ast performance when they are replayed. We use gradient updates to edit memory e
xamples so that they are more likely to be “forgotten” in the future. Experiment
s on five benchmark datasets show the proposed method can be seamlessly combined
 with baselines to significantly improve the performance.
****
Stego Networks: Information Hiding on Deep Neural Networks
Youngwoo Cho,Beomsoo Kim,Jaegul Choo
The best way of keeping a secret is to pretend there is not one. In this spirit,
 a class of techniques called steganography aims to hide secret messages on vari
ous media leaving as little detectable trace as possible. This paper considers n
eural networks as novel steganographic cover media, which we call stego networks
, that can be used to hide one's secret messages. Although there have been numer
ous attempts to hide information in the output of neural networks, techniques fo
r hiding information in the neural network parameters themselves have not been a
ctively studied in the literature. The widespread use of deep learning models in
 various cloud computing platforms and millions of mobile devices as of today im
plies the importance of safety issues regarding stego networks among deep learni
ng researchers and practitioners. In response, this paper presents the advantage
s of stego networks over other types of stego media in terms of security and cap
acity. We provide observations that the fraction bits of some typical network pa
rameters in a floating-point representation tend to follow uniform distributions
 and explain how it can help a secret sender to encrypt messages that are indist
inguishable from the original content. We demonstrate that network parameters ca
n embed a large amount of secret information. Even the most significant fraction
 bits can be used for hiding secrets without inducing noticeable performance deg
radation while making it significantly hard to remove secrets by perturbing insi
gnificant bits. Finally, we discuss possible use cases of stego networks and met
hods to detect or remove secrets from stego networks.
****
AdamP: Slowing Down the Slowdown for Momentum Optimizers on Scale-invariant Weig
hts
Byeongho Heo,Sanghyuk Chun,Seong Joon Oh,Dongyoon Han,Sangdoo Yun,Gyuwan Kim,You
ngjung Uh,Jung-Woo Ha
Normalization techniques, such as batch normalization (BN), are a boon for moder
n deep learning. They let weights converge more quickly with often better genera
lization performances. It has been argued that the normalization-induced scale i

nvariance among the weights provides an advantageous ground for gradient descent (GD) optimizers: the effective step sizes are automatically reduced over time, stabilizing the overall training procedure. It is often overlooked, however, that the additional introduction of momentum in GD optimizers results in a far more rapid reduction in effective step sizes for scale-invariant weights, a phenomenon that has not yet been studied and may have caused unwanted side effects in the current practice. This is a crucial issue because arguably the vast majority of modern deep neural networks consist of (1) momentum-based GD (e.g. SGD or Adam) and (2) scale-invariant parameters (e.g. more than 90% of the weights in ResNet are scale-invariant due to BN). In this paper, we verify that the widely-adopted combination of the two ingredients lead to the premature decay of effective step sizes and sub-optimal model performances. We propose a simple and effective remedy, SGDP and AdamP: get rid of the radial component, or the norm-increasing direction, at each optimizer step. Because of the scale invariance, this modification only alters the effective step sizes without changing the effective update directions, thus enjoying the original convergence properties of GD optimizers. Given the ubiquity of momentum GD and scale invariance in machine learning, we have evaluated our methods against the baselines on 13 benchmarks. They range from vision tasks like classification (e.g. ImageNet), retrieval (e.g. CUB and SOP), and detection (e.g. COCO) to language modelling (e.g. WikiText) and audio classification (e.g. DCASE) tasks. We verify that our solution brings about uniform gains in performances in those benchmarks. Source code is available at https://github.com/clovaai/adamp

****

One Reflection Suffice

Alexander Mathiasen,Frederik Hvilshøj

Orthogonal weight matrices are used in many areas of deep learning. Much previous work attempt to alleviate the additional computational resources it requires to constrain weight matrices to be orthogonal. One popular approach utilizes *many* Householder reflections. The only practical drawback is that many reflections cause low GPU utilization. We mitigate this final drawback by proving that *one* reflection is sufficient, if the reflection is computed by an auxiliary neural network.

****

Learning Subgoal Representations with Slow Dynamics

Siyuan Li,Lulu Zheng,Jianhao Wang,Chongjie Zhang

In goal-conditioned Hierarchical Reinforcement Learning (HRL), a high-level policy periodically sets subgoals for a low-level policy, and the low-level policy is trained to reach those subgoals. A proper subgoal representation function, which abstracts a state space to a latent subgoal space, is crucial for effective goal-conditioned HRL, since different low-level behaviors are induced by reaching subgoals in the compressed representation space. Observing that the high-level agent operates at an abstract temporal scale, we propose a slowness objective to effectively learn the subgoal representation (i.e., the high-level action space). We provide a theoretical grounding for the slowness objective. That is, selecting slow features as the subgoal space can achieve efficient hierarchical exploration. As a result of better exploration ability, our approach significantly outperforms state-of-the-art HRL and exploration methods on a number of benchmark continuous-control tasks. Thanks to the generality of the proposed subgoal representation learning method, empirical results also demonstrate that the learned representation and corresponding low-level policies can be transferred between distinct tasks.

****

A Unified Approach to Interpreting and Boosting Adversarial Transferability

Xin Wang,Jie Ren,Shuyun Lin,Xiangming Zhu,Yisen Wang,Quanshi Zhang

In this paper, we use the interaction inside adversarial perturbations to explain and boost the adversarial transferability. We discover and prove the negative correlation between the adversarial transferability and the interaction inside adversarial perturbations. The negative correlation is further verified through different DNNs with various inputs. Moreover, this negative correlation can be re

garded as a unified perspective to understand current transferability-boosting m
ethods. To this end, we prove that some classic methods of enhancing the transfe
rability essentially decease interactions inside adversarial perturbations. Base
d on this, we propose to directly penalize interactions during the attacking pro
cess, which significantly improves the adversarial transferability. We will rele
ase the code when the paper is accepted.
****

NeurWIN: Neural Whittle Index Network for Restless Bandits via Deep RL
Khaled Nakhleh,Santosh Ganji,Ping-Chun Hsieh,I-Hong Hou,Srinivas Shakkottai
Whittle index policy is a powerful tool to obtain asymptotically optimal solutio
ns for the notoriously intractable problem of restless bandits. However, finding
 the Whittle indices remains a difficult problem for many practical restless ban
dits with convoluted transition kernels. This paper proposes NeurWIN, a neural W
hittle index network that seeks to learn the Whittle indices for any restless ba
ndits by leveraging mathematical properties of the Whittle indices. We show that
 a neural network that produces the Whittle index is also one that produces the
optimal control for a set of Markov decision problems. This property motivates u
sing deep reinforcement learning for the training of NeurWIN. We demonstrate the
 utility of NeurWIN by evaluating its performance for three recently studied res
tless bandit problems. Our experiment results show that the performance of NeurW
IN is either better than, or as good as, state-of-the-art policies for all three
 problems.
****

Differentiable Optimization of Generalized Nondecomposable Functions using Linea
r Programs
Zihang Meng,Lopamudra Mukherjee,Vikas Singh,Sathya N. Ravi
We propose a framework which makes it feasible to directly train deep neural net
works with respect to popular families of task-specific non-decomposable per- fo
rmance measures such as AUC, multi-class AUC, F -measure and others, as well as
models such as non-negative matrix factorization. A common feature of the optimi
zation model that emerges from these tasks is that it involves solving a Linear
Programs (LP) during training where representations learned by upstream layers i
nfluence the constraints. The constraint matrix is not only large but the constr
aints are also modified at each iteration. We show how adopting a set of influen
tial ideas proposed by Mangasarian for 1-norm SVMs – which advocates for solving
 LPs with a generalized Newton method – provides a simple and effective solution
. In particular, this strategy needs little unrolling, which makes it more effic
ient during backward pass. While a number of specialized algorithms have been pr
oposed for the models that we de- scribe here, our module turns out to be applic
able without any specific adjustments or relaxations. We describe each use case,
 study its properties and demonstrate the efficacy of the approach over alternat
ives which use surrogate lower bounds and often, specialized optimization scheme
s. Frequently, we achieve superior computational behavior and performance improv
ements on common datasets used in the literature.

****

Out-of-Distribution Generalization via Risk Extrapolation (REx)
David Krueger,Ethan Caballero,Joern-Henrik Jacobsen,Amy Zhang,Jonathan Binas,Rém
i LE PRIOL,Dinghuai Zhang,Aaron Courville
Distributional shift is one of the major obstacles when transferring machine lea
rning prediction systems from the lab to the real world. To tackle this problem,
 we assume that variation across training domains is representative of the varia
tion we might encounter at test time, but also that shifts at test time may be m
ore extreme in magnitude. In particular, we show that reducing differences in ri
sk across training domains can reduce a model’s sensitivity to a wide range of e
xtreme distributional shifts, including the challenging setting where the input
contains both causal and anti-causal elements. We motivate this approach, Risk E
xtrapolation (REx), as a form of robust optimization over a perturbation set of
extrapolated domains (MM-REx), and propose a penalty on the variance of training
 risks (V-REx) as a simpler variant. We prove that variants of REx can recover t

he causal mechanisms of the targets, while also providing some robustness to cha
nges in the input distribution (``covariate shift''). By appropriately trading-o
ff robustness to causally induced distributional shifts and covariate shift, REx
 is able to outperform alternative methods such as Invariant Risk Minimization i
n situations where these types of shift co-occur.
****

Hidden Incentives for Auto-Induced Distributional Shift
David Krueger,Tegan Maharaj,Jan Leike
Decisions made by machine learning systems have increasing influence on the worl
d, yet it is common for machine learning algorithms to assume that no such influ
ence exists. An example is the use of the i.i.d. assumption in content recommen
dation. In fact, the (choice of) content displayed can change users' perception
s 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 desc
ribe the phenomenon of an algorithm causing a change in the distribution of its
own inputs. Our goal is to ensure that machine learning systems do not leverage
 ADS to increase performance when doing so could be undesirable. We demonstrate
 that changes to the learning algorithm, such as the introduction of meta-learni
ng, can cause hidden incentives for auto-induced distributional shift (HI-ADS) t
o be revealed. To address this issue, we introduce `unit tests' and a mitigatio
n strategy for HI-ADS, as well as a toy environment for modelling real-world iss
ues with HI-ADS in content recommendation, where we demonstrate that strong meta
-learners achieve gains in performance via ADS. We show meta-learning and Q-lea
rning both sometimes fail unit tests, but pass when using our mitigation strateg
y.
****

Orthogonal Subspace Decomposition: A New Perspective of Learning Discriminative
Features for Face Clustering
Jianfeng Wang,Thomas Lukasiewicz,zhongchao shi
Face clustering is an important task, due to its wide applications in
  practice. Graph-based face clustering methods have recently made a gre
at progress and achieved new state-of-the-art results. Learning discriminative
 node features is the key to further improve the performance of graph-based face
 clustering. To this end, we propose subspace learning as a new way to learn di
scriminative node features, which is implemented by a new orthogonal subspace de
composition (OSD) module. In graph-based face clustering, OSD leads to more disc
riminative node features, which better reflect the relationship between each pa
ir of faces, thereby boosting the accuracy of face clustering. Extensive experi
ments show that OSD outperforms state-of-the-art results with a healthy margin.
****

Efficiently Disentangle Causal Representations
Yuanpeng Li,Joel Hestness,Mohamed Elhoseiny,Liang Zhao,Kenneth Church
In this paper, we propose a novel approach to efficiently learning disentangled
representations with causal mechanisms, based on the difference of conditional p
robabilities in original and new distributions. We approximate the difference wi
th model's generalization abilities so that it fits in standard machine learning
 framework and can be efficiently computed. In contrast to the state-of-the-art
approach, which relies on learner's adaptation speed to new distribution, the pr
oposed approach only requires evaluating the generalization ability of the model
. We provide theoretical explanation for the advantage of the proposed method, a
nd our experiments show that the proposed technique is 1.9-11.0x more sample eff
icient and 9.4-32.4x quicker than the previous method on various tasks.
****

Perceptual Adversarial Robustness: Defense Against Unseen Threat Models
Cassidy Laidlaw,Sahil Singla,Soheil Feizi
A key challenge in adversarial robustness is the lack of a precise mathematical
characterization of human perception, used in the definition of adversarial atta
cks that are imperceptible to human eyes. Most current attacks and defenses try
to get around this issue by considering restrictive adversarial threat models su
ch as those bounded by L_2 or L_∞ distance, spatial perturbations, etc.

However, models that are robust against any of these restrictive threat models are still fragile against other threat models, i.e. they have poor generalization to unforeseen attacks. Moreover, even if a model is robust against the union of several restrictive threat models, it is still susceptible to other imperceptible adversarial examples that are not contained in any of the constituent threat models. To resolve these issues, we propose adversarial training against the set of all imperceptible adversarial examples. Since this set is intractable to compute without a human in the loop, we approximate it using deep neural networks. We call this threat model the neural perceptual threat model (NPTM); it includes adversarial examples with a bounded neural perceptual distance (a neural network-based approximation of the true perceptual distance) to natural images. Through an extensive perceptual study, we show that the neural perceptual distance correlates well with human judgements of perceptibility of adversarial examples, validating our threat model.

Under the NPTM, we develop novel perceptual adversarial attacks and defenses. Because the NPTM is very broad, we find that Perceptual Adversarial Training (PAT) against a perceptual attack gives robustness against many other types of adversarial attacks. We test PAT on CIFAR-10 and ImageNet-100 against five diverse adversarial attacks: L_2, L_∞, spatial, recoloring, and JPEG. We find that PAT achieves state-of-the-art robustness against the union of these five attacks—more than doubling the accuracy over the next best model—without training against any of them. That is, PAT generalizes well to unforeseen perturbation types. This is vital in sensitive applications where a particular threat model cannot be assumed, and to the best of our knowledge, PAT is the first adversarial training defense with this property.

Code and data are available at https://github.com/cassidylaidlaw/perceptual-advex
****
Better Fine-Tuning by Reducing Representational Collapse
Armen Aghajanyan,Akshat Shrivastava,Anchit Gupta,Naman Goyal,Luke Zettlemoyer,Sonal Gupta
Although widely adopted, existing approaches for fine-tuning pre-trained language models have been shown to be unstable across hyper-parameter settings, motivating recent work on trust region methods. In this paper, we present a simplified and efficient method rooted in trust region theory that replaces previously used adversarial objectives with parametric noise (sampling from either a normal or uniform distribution), thereby discouraging representation change during fine-tuning when possible without hurting performance. We also introduce a new analysis to motivate the use of trust region methods more generally, by studying representational collapse; the degradation of generalizable representations from pre-trained models as they are fine-tuned for a specific end task. Extensive experiments show that our fine-tuning method matches or exceeds the performance of previous trust region methods on a range of understanding and generation tasks (including DailyMail/CNN, Gigaword, Reddit TIFU, and the GLUE benchmark), while also being much faster. We also show that it is less prone to representation collapse; the pre-trained models maintain more generalizable representations every time they are fine-tuned.
****
Cross-Modal Domain Adaptation for Reinforcement Learning
Xiong-Hui Chen,Shengyi Jiang,Feng Xu,Yang Yu
Domain adaptation is a promising direction for deploying RL agents in real-world applications, where vision-based robotics tasks constitute an important part. Current methods that train polices on simulated images not only require a delicately crafted simulator, but also add extra burdens to the training process. In this paper, we propose a method that can learn a mapping from high-dimensional images to low-level simulator states, allowing agents trained on the source domain of state input to transfer well to the target domain of image input. By fully leveraging the sequential information in the trajectories and incorporating the p

olicy to guide the training process, our method overcomes the intrinsic ill-pose dness in cross-modal domain adaptation when structural constraints from the same modality are unavailable. Experiments on MuJoCo environments show that the poli cy, once combined with the mapping function, can be deployed directly in the ta rget domain with only a small performance gap, while current methods designed fo r same-modal domain adaptation fail on this problem.
****
DeeperGCN: Training Deeper GCNs with Generalized Aggregation Functions
Guohao Li,Chenxin Xiong,Ali Thabet,Bernard Ghanem
Graph Convolutional Networks (GCNs) have been drawing significant attention with the power of representation learning on graphs. Recent works developed framewor ks to train deep GCNs. Such works show impressive results in tasks like point cl oud classification and segmentation, and protein interaction prediction. In this work, we study the performance of such deep models in large scale graph dataset s from the Open Graph Benchmark (OGB). In particular, we look at the effect of a dequately choosing an aggregation function, and its effect on final performance. Common choices of aggregation are mean, max, and sum. It has shown that GCNs ar e sensitive to such aggregations when applied to different datasets. We further validate this point and propose to alleviate it by introducing a novel Generaliz ed Aggregation Function. Our new aggregation not only covers all commonly used o nes, but also can be tuned to learn customized functions for different tasks. Ou r generalized aggregation is fully differentiable, and thus its parameters can b e learned in an end-to-end fashion. We add our generalized aggregation into a de ep GCN framework and show it achieves state-of-the-art results in six benchmarks from OGB.
****
DiffAutoML: Differentiable Joint Optimization for Efficient End-to-End Automated Machine Learning
Kaichen Zhou,Lanqing HONG,Fengwei Zhou,Binxin Ru,Zhenguo Li,Trigoni Niki,Jiashi Feng
The automated machine learning (AutoML) pipeline comprises several crucial compo nents such as automated data augmentation (DA), neural architecture search (NAS) and hyper-parameter optimization (HPO). Although many strategies have been deve loped for automating each component in separation, joint optimization of these c omponents remains challenging due to the largely increased search dimension and different input types required for each component. While conducting these compon ents in sequence is usually adopted as a workaround, it often requires careful c oordination by human experts and may lead to sub-optimal results. In parallel to this, the common practice of searching for the optimal architecture first and t hen retraining it before deployment in NAS often suffers from architecture perfo rmance difference in the search and retraining stages. An end-to-end solution th at integrates the two stages and returns a ready-to-use model at the end of the search is desirable. In view of these, we propose a differentiable joint optimiz ation solution for efficient end-to-end AutoML (DiffAutoML). Our method performs co-optimization of the neural architectures, training hyper-parameters and data augmentation policies in an end-to-end fashion without the need of model retrai ning. Experiments show that DiffAutoML achieves state-of-the-art results on Ima geNet compared with end-to-end AutoML algorithms, and achieves superior performa nce compared with multi-stage AutoML algorithms with higher computational effici ency.
To the best of our knowledge, we are the first to jointly optimize automated DA, NAS and HPO in an en-to-end manner without retraining.
****
Learning N:M Fine-grained Structured Sparse Neural Networks From Scratch
Aojun Zhou,Yukun Ma,Junnan Zhu,Jianbo Liu,Zhijie Zhang,Kun Yuan,Wenxiu Sun,Hongs heng Li
Sparsity in Deep Neural Networks (DNNs) has been widely studied to compress and accelerate the models on resource-constrained environments. It can be generally categorized into unstructured fine-grained sparsity that zeroes out multiple ind ividual weights distributed across the neural network, and structured coarse-gra

ined sparsity which prunes blocks of sub-networks of a neural network. Fine-grai ned sparsity can achieve a high compression ratio but is not hardware friendly a nd hence receives limited speed gains. On the other hand, coarse-grained sparsit y cannot simultaneously achieve both apparent acceleration on modern GPUs and decent performance. In this paper, we are the first to study training from scrat ch an N:M fine-grained structured sparse network, which can maintain the advanta ges of both unstructured fine-grained sparsity and structured coarse-grained spa rsity simultaneously on specifically designed GPUs. Specifically, a 2 : 4 sparse network could achieve 2× speed-up without performance drop on Nvidia A100 GPUs. Furthermore, we propose a novel and effective ingredient, sparse-refined straig ht-through estimator (SR-STE), to alleviate the negative influence of the approx imated gradients computed by vanilla STE during optimization. We also define a m etric, Sparse Architecture Divergence (SAD), to measure the sparse network’s top ology change during the training process. Finally, We justify SR-STE’s advantage s with SAD and demonstrate the effectiveness of SR-STE by performing comprehensive experiments on various tasks. Anonymous code and model will be at available at https://github.com/anonymous-NM-sparsity/NM-sparsity.
****

Vision at A Glance: Interplay between Fine and Coarse Information Processing Pat hways
Zilong Ji,Xiaolong Zou,Tiejun Huang,Si Wu
Object recognition is often viewed as a feedforward, bottom-up process in machin e learning, but in real neural systems, object recognition is a complicated proc ess which involves the interplay between two signal pathways. One is the parvoce llular pathway (P-pathway), which is slow and extracts fine features of objects; the other is the magnocellular pathway (M-pathway), which is fast and extracts coarse features of objects. It has been suggested that the interplay between the two pathways endows the neural system with the capacity of processing visual in formation rapidly, adaptively, and robustly. However, the underlying computation al mechanism remains largely unknown. In this study, we build a two-pathway mode l to elucidate the computational properties associated with the interactions bet ween two visual pathways. The model consists of two convolution neural networks: one mimics the P-pathway, referred to as FineNet, which is deep, has small-size kernels, and receives detailed visual inputs; the other mimics the M-pathway, r eferred to as CoarseNet, which is shallow, has large-size kernels, and receives blurred visual inputs.
The two pathways interact with each other to facilitate information processing. Specifically, we show that CoarseNet can learn from FineNet through imitation to improve its performance considerably, and that through feedback from CoarseNet, the performnace of FineNet is improved and becomes robust to noises. Using visu al backward masking as an example, we demonstrate that our model can explain vis ual cognitive behaviors that involve the interplay between two pathways. We hope that this study will provide insight into understanding visual information proc essing and inspire the development of new object recognition architectures in ma chine learning.
****

EqCo: Equivalent Rules for Self-supervised Contrastive Learning
Benjin Zhu,Junqiang Huang,Zeming Li,Xiangyu Zhang,Jian Sun
In this paper, we propose a method, named EqCo (Equivalent Rules for Contrastive Learning), to make self-supervised learning irrelevant to the number of negativ e samples in the contrastive learning framework. Inspired by the InfoMax princip le, we point that the margin term in contrastive loss needs to be adaptively sca led according to the number of negative pairs in order to keep steady mutual inf ormation bound and gradient magnitude. EqCo bridges the performance gap among a wide range of negative sample sizes, so that for the first time, we can use only a few negative pairs (e.g. 16 per query) to perform self-supervised contrastive training on large-scale vision datasets like ImageNet, while with almost no acc uracy drop. This is quite a contrast to the widely used large batch training or memory bank mechanism in current practices. Equipped with EqCo, our simplified M oCo (SiMo) achieves comparable accuracy with MoCo v2 on ImageNet (linear evaluat

ion protocol) while only involves 16 negative pairs per query instead of 65536, suggesting that large quantities of negative samples might not be a critical fac tor in contrastive learning frameworks.
****

Synthesising Realistic Calcium Traces of Neuronal Populations Using GAN
Bryan M. Li,Theoklitos Amvrosiadis,Nathalie Rochefort,Arno Onken
Calcium imaging has become a powerful and popular technique to monitor the activ ity of large populations of neurons in vivo. However, for ethical considerations and despite recent technical developments, recordings are still constrained to a limited number of trials and animals. This limits the amount of data available from individual experiments and hinders the development of analysis techniques and models for more realistic sizes of neuronal populations. The ability to arti ficially synthesize realistic neuronal calcium signals could greatly alleviate t his problem by scaling up the number of trials. Here, we propose a Generative Ad versarial Network (GAN) model to generate realistic calcium signals as seen in n euronal somata with calcium imaging. To this end, we propose CalciumGAN, a model based on the WaveGAN architecture and train it on calcium fluorescent signals w ith the Wasserstein distance. We test the model on artificial data with known gr ound-truth and show that the distribution of the generated signals closely resem bles the underlying data distribution. Then, we train the model on real calcium traces recorded from the primary visual cortex of behaving mice and confirm that the deconvolved spike trains match the statistics of the recorded data. Togethe r, these results demonstrate that our model can successfully generate realistic calcium traces, thereby providing the means to augment existing datasets of neur onal activity for enhanced data exploration and modelling.
****

Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Lear ning
Xue Bin Peng,Aviral Kumar,Grace Zhang,Sergey Levine
In this work, we aim to develop a simple and scalable reinforcement learning alg orithm that uses standard supervised learning methods as subroutines, while also being able to leverage off-policy data. Our proposed approach, which we refer t o as advantage-weighted regression (AWR), consists of two standard supervised le arning steps: one to regress onto target values for a value function, and anothe r to regress onto weighted target actions for the policy. The method is simple a nd general, can accommodate continuous and discrete actions, and can be implemen ted in just a few lines of code on top of standard supervised learning methods. We provide a theoretical motivation for AWR and analyze its properties when inco rporating off-policy data from experience replay. We evaluate AWR on a suite of standard OpenAI Gym benchmark tasks, and show that it achieves competitive perfo rmance compared to a number of well-established state-of-the-art RL algorithms. AWR is also able to acquire more effective policies than most off-policy algorit hms when learning from purely static datasets with no additional environmental i nteractions. Furthermore, we demonstrate our algorithm on challenging continuous control tasks with highly complex simulated characters.
****

Active Contrastive Learning of Audio-Visual Video Representations
Shuang Ma,Zhaoyang Zeng,Daniel McDuff,Yale Song
Contrastive learning has been shown to produce generalizable representations of audio and visual data by maximizing the lower bound on the mutual information (M I) between different views of an instance. However, obtaining a tight lower boun d requires a sample size exponential in MI and thus a large set of negative samp les. We can incorporate more samples by building a large queue-based dictionary, but there are theoretical limits to performance improvements even with a large number of negative samples. We hypothesize that random negative sampling leads t o a highly redundant dictionary that results in suboptimal representations for d ownstream tasks. In this paper, we propose an active contrastive learning approa ch that builds an actively sampled dictionary with diverse and informative items , which improves the quality of negative samples and improves performances on ta sks where there is high mutual information in the data, e.g., video classificati

on. Our model achieves state-of-the-art performance on challenging audio and visual downstream benchmarks including UCF101, HMDB51 and ESC50.
****
Relevance Attack on Detectors
Sizhe Chen,Fan He,Xiaolin Huang,Kun Zhang
This paper focuses on high-transferable adversarial attacks on detectors, which are hard to attack in a black-box manner, because of their multiple-output characteristics and the diversity across architectures. To pursue a high attack transferability, one plausible way is to find a common property across detectors, which facilitates the discovery of common weaknesses. We are the first to suggest that the relevance map for detectors is such a property. Based on it, we design a Relevance Attack on Detectors (RAD), which achieves a state-of-the-art transferability, exceeding existing results by above 20%. On MS COCO, the detection mAPs for all 8 black-box architectures are more than halved and the segmentation mAPs are also significantly influenced. Given the great transferability of RAD, we generate the first adversarial dataset for object detection, i.e., Adversarial Objects in COntext (AOCO), which helps to quickly evaluate and improve the robustness of detectors.
****
Deep Evolutionary Learning for Molecular Design
Yifeng Li,Hsu Kiang Ooi,Alain Tchagang
In this paper, we propose a deep evolutionary learning (DEL) process that integrates fragment-based deep generative model and multi-objective evolutionary computation for molecular design. Our approach enables (1) evolutionary operations in the latent space of the generative model, rather than the structural space, to generate novel promising molecular structures for the next evolutionary generation, and (2) generative model fine-tuning using newly generated high-quality samples. Thus, DEL implements a data-model co-evolution concept which improves both sample population and generative model learning. Experiments on two public datasets indicate that sample population obtained by DEL exhibits improved property distributions, and dominates samples generated by multi-objective Bayesian optimization algorithms.
****
A law of robustness for two-layers neural networks
Sebastien Bubeck,Yuanzhi Li,Dheeraj Mysore Nagaraj
We initiate the study of the inherent tradeoffs between the size of a neural network and its robustness, as measured by its Lipschitz constant. We make a precise conjecture that, for any Lipschitz activation function and for most datasets, any two-layers neural network with k neurons that perfectly fit the data must have its Lipschitz constant larger (up to a constant) than $\sqrt{n/k}$ where n is the number of datapoints. In particular, this conjecture implies that overparametrization is necessary for robustness, since it means that one needs roughly one neuron per datapoint to ensure a $O(1)$-Lipschitz network, while mere data fitting of d-dimensional data requires only one neuron per d datapoints. We prove a weaker version of this conjecture when the Lipschitz constant is replaced by an upper bound on it based on the spectral norm of the weight matrix. We also prove the conjecture in the high-dimensional regime $n \approx d$ (which we also refer to as the undercomplete case, since only $k \leq d$ is relevant here). Finally we prove the conjecture for polynomial activation functions of degree p when $n \approx d^p$. We complement these findings with experimental evidence supporting the conjecture.
****
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 unsupe

rvised
learning method for addressing the relational learning problem where we
learn the underlying relationship between a pair of data irrespective of the nat
ure
of those data. The central idea of the proposed method is to encapsulate the rel
ational
learning problem with a probabilistic graphical model in which we perform
inference to learn about data relationships and other relational processing task
s.
****

PseudoSeg: Designing Pseudo Labels for Semantic Segmentation
Yuliang Zou,Zizhao Zhang,Han Zhang,Chun-Liang Li,Xiao Bian,Jia-Bin Huang,Tomas P
fister
Recent advances in semi-supervised learning (SSL) demonstrate that a combination
 of consistency regularization and pseudo-labeling can effectively improve image
 classification accuracy in the low-data regime. Compared to classification, sem
antic segmentation tasks require much more intensive labeling costs. Thus, these
 tasks greatly benefit from data-efficient training methods. However, structured
 outputs in segmentation render particular difficulties (e.g., designing pseudo-
labeling and augmentation) to apply existing SSL strategies. To address this pro
blem, we present a simple and novel re-design of pseudo-labeling to generate wel
l-calibrated structured pseudo labels for training with unlabeled or weakly-labe
led data. Our proposed pseudo-labeling strategy is network structure agnostic to
 apply in a one-stage consistency training framework. We demonstrate the effecti
veness of the proposed pseudo-labeling strategy in both low-data and high-data r
egimes. Extensive experiments have validated that pseudo labels generated from w
isely fusing diverse sources and strong data augmentation are crucial to consist
ency training for segmentation. The source code will be released.
****

Finding Physical Adversarial Examples for Autonomous Driving with Fast and Diffe
rentiable Image Compositing
Jinghan Yang,Adith Boloor,Ayan Chakrabarti,Xuan Zhang,Yevgeniy Vorobeychik
There is considerable evidence that deep neural networks are vulnerable to adver
sarial perturbations applied directly to their digital inputs. However, it remai
ns an open question whether this translates to vulnerabilities in real-world sys
tems. Specifically, in the context of image inputs to autonomous driving systems
, an attack can be achieved only by modifying the physical environment, so as to
 ensure that the resulting stream of video inputs to the car's controller leads
to incorrect driving decisions. Inducing this effect on the video inputs indirec
tly through the environment requires accounting for system dynamics and tracking
 viewpoint changes. We propose a scalable and efficient approach for finding adv
ersarial physical modifications, using a differentiable approximation for the ma
pping from environmental modifications—namely, rectangles drawn on the road—to t
he corresponding video inputs to the controller network. Given the color, locati
on, position, and orientation parameters of the rectangles, our mapping composit
es them onto pre-recorded video streams of the original environment. Our mapping
 accounts for geometric and color variations, is differentiable with respect to
rectangle parameters, and uses multiple original video streams obtained by varyi
ng the driving trajectory. When combined with a neural network-based controller,
 our approach allows the design of adversarial modifications through end-to-end
gradient-based optimization. We evaluate our approach using the Carla autonomous
 driving simulator, and show that it is significantly more scalable and far more
 effective at generating attacks than a prior black-box approach based on Bayesi
an Optimization.
****

CorrAttack: Black-box Adversarial Attack with Structured Search
Zhichao Huang,Yaowei Huang,Tong Zhang
We present a new method for score-based adversarial attack, where the attacker q
ueries the loss-oracle of the target model.
Our method employs a parameterized search space with a structure that captures t

he relationship of the gradient of the loss function. We show that searching ove
r the structured space can be approximated by a time-varying contextual bandits
problem, where the attacker takes feature of the associated arm to make modifica
tions of the input, and receives an immediate reward as the reduction of the los
s function. The time-varying contextual bandits problem can then be solved by a
Bayesian optimization procedure, which can take advantage of the features of the
 structured action space. The experiments on ImageNet and the Google Cloud Visio
n API demonstrate that the proposed method achieves the state of the art success
 rates and query efficiencies for both undefended and defended models.
****

Correcting experience replay for multi-agent communication
Sanjeevan Ahilan,Peter Dayan
We consider the problem of learning to communicate using multi-agent reinforceme
nt learning (MARL). A common approach is to learn off-policy, using data sampled
 from a replay buffer. However, messages received in the past may not accurately
 reflect the current communication policy of each agent, and this complicates le
arning. We therefore introduce a 'communication correction' which accounts for t
he non-stationarity of observed communication induced by multi-agent learning. I
t works by relabelling the received message to make it likely under the communic
ator's current policy, and thus be a better reflection of the receiver's current
 environment. To account for cases in which agents are both senders and receiver
s, we introduce an ordered relabelling scheme. Our correction is computationally
 efficient and can be integrated with a range of off-policy algorithms. We find
in our experiments that it substantially improves the ability of communicating M
ARL systems to learn across a variety of cooperative and competitive tasks.
****

Counterfactual Generative Networks
Axel Sauer,Andreas Geiger
Neural networks are prone to learning shortcuts -- they often model simple corre
lations, ignoring more complex ones that potentially generalize better. Prior wo
rks on image classification show that instead of learning a connection to object
 shape, deep classifiers tend to exploit spurious correlations with low-level te
xture or the background for solving the classification task. In this work, we ta
ke a step towards more robust and interpretable classifiers that explicitly expo
se the task's causal structure. Building on current advances in deep generative
modeling, we propose to decompose the image generation process into independent
causal mechanisms that we train without direct supervision. By exploiting approp
riate inductive biases, these mechanisms disentangle object shape, object textur
e, and background; hence, they allow for generating counterfactual images. We de
monstrate the ability of our model to generate such images on MNIST and ImageNet
. Further, we show that the counterfactual images can improve out-of-distributio
n robustness with a marginal drop in performance on the original classification
task, despite being synthetic. Lastly, our generative model can be trained effic
iently on a single GPU, exploiting common pre-trained models as inductive biases
.
****

Quantile Regularization : Towards Implicit Calibration of Regression Models
Saiteja Utpala,Piyush Rai
Recent works have shown that most deep learning models are often poorly calibra
ted, i.e., they may produce overconfident
predictions that are wrong, implying that their uncertainty estimates are unreli
able. While a number of approaches have been proposed recently to calibrate cla
ssification models, relatively little work exists on calibrating regression mode
ls. Isotonic Regression has recently been advocated for regression calibration.
We provide a detailed formal analysis of the \emph{side-effects} of Isotonic Reg
ression when used for regression calibration. To address this, we recast quant
ile calibration as entropy estimation, and leverage this idea to construct a nov
el quantile regularizer, which can be used in any optimization based probabilisi
tc regression models. Unlike most of the existing approaches for calibrating reg
ression models, which are based on \emph{post-hoc} processing of the model's out

put, and require an additional dataset, our method is trainable in an end-to-end fashion, without requiring an additional dataset. We provide empirical results demonstrating that our approach improves calibration for regression models trained on diverse architectures that provide uncertainty estimates, such as Dropout VI, Deep Ensembles

****

Multi-Agent Trust Region Learning

Ying Wen,Hui Chen,Yaodong Yang,Zheng Tian,Minne Li,Xu Chen,Jun Wang

Trust-region methods are widely used in single-agent reinforcement learning. One advantage is that they guarantee a lower bound of monotonic payoff improvement for policy optimization at each iteration. Nonetheless, when applied in multi-agent settings, such guarantee is lost because an agent's payoff is also determined by other agents' adaptive behaviors. In fact, measuring agents' payoff improvements in multi-agent reinforcement learning (MARL) scenarios is still challenging. Although game-theoretical solution concepts such as Nash equilibrium can be applied, the algorithm (e.g., Nash-Q learning) suffers from poor scalability beyond two-player discrete games. To mitigate the above measurability and tractability issues, in this paper, we propose Multi-Agent Trust Region Learning (MATRL) method. MATRL augments the single-agent trust-region optimization process with the multi-agent solution concept of stable fixed point that is computed at the policy-space meta-game level. When multiple agents learn simultaneously, stable fixed points at the meta-game level can effectively measure agents' payoff improvements, and, importantly, a meta-game representation enjoys better scalability for multi-player games. We derive the lower bound of agents' payoff improvements for MATRL methods, and also prove the convergence of our method on the meta-game fixed points. We evaluate the MATRL method on both discrete and continuous multi-player general-sum games; results suggest that MATRL significantly outperforms strong MARL baselines on grid worlds, multi-agent MuJoCo, and Atari games.

****

Certified Distributional Robustness via Smoothed Classifiers

Jungang Yang,Liyao Xiang,Ruidong Chen,Yukun Wang,Wei Wang,Xinbing Wang

The robustness of deep neural networks against adversarial example attacks has received much attention recently. We focus on certified robustness of smoothed classifiers in this work, and propose to use the worst-case population loss over noisy inputs as a robustness metric. Under this metric, we provide a tractable upper bound serving as a robustness certificate by exploiting the duality. To improve the robustness, we further propose a noisy adversarial learning procedure to minimize the upper bound following the robust optimization framework. The smoothness of the loss function ensures the problem easy to optimize even for non-smooth neural networks. We show how our robustness certificate compares with others and the improvement over previous works. Experiments on a variety of datasets and models verify that in terms of empirical accuracies, our approach exceeds the state-of-the-art certified/heuristic methods in defending adversarial examples.

****

Global Self-Attention Networks for Image Recognition

Zhuoran Shen,Irwan Bello,Raviteja Vemulapalli,Xuhui Jia,Ching-Hui Chen

Recently, a series of works in computer vision have shown promising results on various image and video understanding tasks using self-attention. However, due to the quadratic computational and memory complexities of self-attention, these works either apply attention only to low-resolution feature maps in later stages of a deep network or restrict the receptive field of attention in each layer to a small local region. To overcome these limitations, this work introduces a new global self-attention module, referred to as the GSA module, which is efficient enough to serve as the backbone component of a deep network. This module consists of two parallel layers: a content attention layer that attends to pixels based only on their content and a positional attention layer that attends to pixels based on their spatial locations. The output of this module is the sum of the outputs of the two layers. Based on the proposed GSA module, we introduce new standalone global attention-based deep networks that use GSA modules instead of convolutions to model pixel interactions. Due to the global extent of the proposed GSA

module, a GSA network has the ability to model long-range pixel interactions th
roughout the network. Our experimental results show that GSA networks outperform
the corresponding convolution-based networks significantly on the CIFAR-100 and
ImageNet datasets while using less number of parameters and computations. The p
roposed GSA networks also outperform various existing attention-based networks o
n the ImageNet dataset.
****
Contemplating Real-World Object Classification
ali borji
Deep object recognition models have been very successful over benchmark
datasets such as ImageNet. How accurate and robust are they to distribution
shifts arising from natural and synthetic variations in datasets? Prior research
on
this problem has primarily focused on ImageNet variations (e.g., ImageNetV2,
ImageNet-A). To avoid potential inherited biases in these studies, we take a
different approach. Specifically, we reanalyze the ObjectNet dataset recently
proposed by Barbu et al. containing objects in daily life situations. They showe
d
a dramatic performance drop of the state of the art object recognition models on
this dataset. Due to the importance and implications of their results regarding
the generalization ability of deep models, we take a second look at their analys
is.
We find that applying deep models to the isolated objects, rather than the entir
e
scene as is done in the original paper, results in around 20-30% performance
improvement. Relative to the numbers reported in Barbu et al., around 10-15%
of the performance loss is recovered, without any test time data augmentation.
Despite this gain, however, we conclude that deep models still suffer drasticall
y
on the ObjectNet dataset. We also investigate the robustness of models against
synthetic image perturbations such as geometric transformations (e.g., scale,
rotation, translation), natural image distortions (e.g., impulse noise, blur) as
well
as adversarial attacks (e.g., FGSM and PGD-5). Our results indicate that limitin
g
the object area as much as possible (i.e., from the entire image to the bounding
box to the segmentation mask) leads to consistent improvement in accuracy and
robustness. Finally, through a qualitative analysis of ObjectNet data, we find t
hat
i) a large number of images in this dataset are hard to recognize even for human
s,
and ii) easy (hard) samples for models match with easy (hard) samples for humans
.
Overall, our analysis shows that ObjecNet is still a challenging test platform t
hat
can be used to measure the generalization ability of models. The code and data
are available in [masked due to blind review].
****
Semantically-Adaptive Upsampling for Layout-to-Image Translation
Hao Tang,Nicu Sebe
We propose the Semantically-Adaptive UpSampling (SA-UpSample), a general and hig
hly effective upsampling method for the layout-to-image translation task. SA-UpS
ample has three advantages: 1) Global view. Unlike traditional upsampling method
s (e.g., Nearest-neighbor) that only exploit local neighborhoods, SA-UpSample ca
n aggregate semantic information in a global view.
2) Semantically adaptive. Instead of using a fixed kernel for all locations (e.
g., Deconvolution), SA-UpSample enables semantic class-specific upsampling via g
enerating adaptive kernels for different locations. 3) Efficient. Unlike Spatial
Attention which uses a fully-connected strategy to connect all the pixels, SA-U
pSample only considers the most relevant pixels, introducing little computationa

l overhead. We observe that SA-UpSample achieves consistent and substantial gain s on six popular datasets. The source code will be made publicly available.
****

Transferability of Compositionality
Yuanpeng Li,Liang Zhao,Joel Hestness,Ka Yee Lun,Kenneth Church,Mohamed Elhoseiny
Compositional generalization is the algebraic capacity to understand and produce large amount of novel combinations from known components. It is a key element o f human intelligence for out-of-distribution generalization. To equip neural net works with such ability, many algorithms have been proposed to extract compositi onal representations from the training distribution. However, it has not been di scussed whether the trained model can still extract such representations in the test distribution. In this paper, we argue that the extraction ability does not transfer naturally, because the extraction network suffers from the divergence o f distributions. To address this problem, we propose to use an auxiliary reconst ruction network with regularized hidden representations as input, and optimize t he representations during inference. The proposed approach significantly improve s accuracy, showing more than a 20% absolute increase in various experiments com pared with baselines. To our best knowledge, this is the first work to focus on the transferability of compositionality, and it is orthogonal to existing effort s of learning compositional representations in training distribution. We hope th is work will help to advance compositional generalization and artificial intelli gence research.
****

Bayesian Online Meta-Learning
Pauching Yap,Hippolyt Ritter,David Barber
Neural networks are known to suffer from catastrophic forgetting when trained on sequential datasets. While there have been numerous attempts to solve this prob lem for large-scale supervised classification, little has been done to overcome catastrophic forgetting for few-shot classification problems. Few-shot meta-lear ning algorithms often require all few-shot tasks to be readily available in a ba tch for training. The popular gradient-based model-agnostic meta-learning algori thm (MAML) is a typical algorithm that suffers from these limitations. This work introduces a Bayesian online meta-learning framework to tackle the catastrophic forgetting and the sequential few-shot tasks problems. Our framework incorporat es MAML into a Bayesian online learning algorithm with Laplace approximation or variational inference. This framework enables few-shot classification on a rang e of sequentially arriving datasets with a single meta-learned model and trainin g on sequentially arriving few-shot tasks. The experimental evaluations demonstr ate that our framework can effectively prevent catastrophic forgetting and is ca pable of online meta-learning in various few-shot classification settings.
****

Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding
David A. Klindt,Lukas Schott,Yash Sharma,Ivan Ustyuzhaninov,Wieland Brendel,Matt hias Bethge,Dylan Paiton
Disentangling the underlying generative factors from complex data has so far bee n limited to carefully constructed scenarios. We propose a path towards natural data by first showing that the statistics of natural data provide enough structu re to enable disentanglement, both theoretically and empirically. Specifically, we provide evidence that objects in natural movies undergo transitions that are typically small in magnitude with occasional large jumps, which is characteristi c of a temporally sparse distribution. To address this finding we provide a nove l proof that relies on a sparse prior on temporally adjacent observations to rec over the true latent variables up to permutations and sign flips, directly provi ding a stronger result than previous work. We show that equipping practical esti mation methods with our prior often surpasses the current state-of-the-art on se veral established benchmark datasets without any impractical assumptions, such a s knowledge of the number of changing generative factors. Furthermore, we contri bute two new benchmarks, Natural Sprites and KITTI Masks, which integrate the me asured natural dynamics to enable disentanglement evaluation with more realistic datasets. We leverage these benchmarks to test our theory, demonstrating improv

ed performance. We also identify non-obvious challenges for current methods in s
caling to more natural domains. Taken together our work addresses key issues in
disentanglement research for moving towards more natural settings.
****
Explainable Deep One-Class Classification
Philipp Liznerski,Lukas Ruff,Robert A. Vandermeulen,Billy Joe Franks,Marius Klof
t,Klaus Robert Muller
Deep one-class classification variants for anomaly detection learn a mapping tha
t concentrates nominal samples in feature space causing anomalies to be mapped a
way. Because this transformation is highly non-linear, finding interpretations p
oses a significant challenge. In this paper we present an explainable deep one-c
lass classification method, Fully Convolutional Data Description (FCDD), where t
he mapped samples are themselves also an explanation heatmap. FCDD yields compet
itive detection performance and provides reasonable explanations on common anoma
ly detection benchmarks with CIFAR-10 and ImageNet. On MVTec-AD, a recent manufa
cturing dataset offering ground-truth anomaly maps, FCDD sets a new state of the
 art in the unsupervised setting. Our method can incorporate ground-truth anomal
y maps during training and using even a few of these (~5) improves performance s
ignificantly. Finally, using FCDD's explanations we demonstrate the vulnerabilit
y of deep one-class classification models to spurious image features such as ima
ge watermarks.
****
Batch Reinforcement Learning Through Continuation Method
Yijie Guo,Shengyu Feng,Nicolas Le Roux,Ed Chi,Honglak Lee,Minmin Chen
Many real-world applications of reinforcement learning (RL) require the agent to
 learn from a fixed set of trajectories, without collecting new interactions. P
olicy optimization under this setting is extremely challenging as: 1) the geomet
ry of the objective function is hard to optimize efficiently; 2) the shift of da
ta distributions causes high noise in the value estimation. In this work, we pro
pose a simple yet effective policy iteration approach to batch RL using global o
ptimization techniques known as continuation. By constraining the difference be
tween the learned policy and the behavior policy that generates the fixed trajec
tories, and continuously relaxing the constraint, our method 1) helps the agent
escape local optima; 2) reduces the error in policy evaluation in the optimizati
on procedure. We present results on a variety of control tasks, game environme
nts, and a recommendation task to empirically demonstrate the efficacy of our pr
oposed method.
****
Protecting DNNs from Theft using an Ensemble of Diverse Models
Sanjay Kariyappa,Atul Prakash,Moinuddin K Qureshi
Several recent works have demonstrated highly effective model stealing (MS) atta
cks on Deep Neural Networks (DNNs) in black-box settings, even when the training
 data is unavailable. These attacks typically use some form of Out of Distributi
on (OOD) data to query the target model and use the predictions obtained to trai
n a clone model. Such a clone model learns to approximate the decision boundary
of the target model, achieving high accuracy on in-distribution examples. We pro
pose Ensemble of Diverse Models (EDM) to defend against such MS attacks. EDM is
made up of models that are trained to produce dissimilar predictions for OOD inp
uts. By using a different member of the ensemble to service different queries, o
ur defense produces predictions that are highly discontinuous in the input space
 for the adversary's OOD queries. Such discontinuities cause the clone model tra
ined on these predictions to have poor generalization on in-distribution example
s. Our evaluations on several image classification tasks demonstrate that EDM de
fense can severely degrade the accuracy of clone models (up to 39.7%). Our de
fense has minimal impact on the target accuracy, negligible computational costs
during inference, and is compatible with existing defenses for MS attacks.
****
Domain Generalization with MixStyle
Kaiyang Zhou,Yongxin Yang,Yu Qiao,Tao Xiang
Though convolutional neural networks (CNNs) have demonstrated remarkable ability

in learning discriminative features, they often generalize poorly to unseen dom
ains. Domain generalization aims to address this problem by learning from a set
of source domains a model that is generalizable to any unseen domain. In this pa
per, a novel approach is proposed based on probabilistically mixing instance-lev
el feature statistics of training samples across source domains. Our method, ter
med MixStyle, is motivated by the observation that visual domain is closely rela
ted to image style (e.g., photo vs.~sketch images). Such style information is ca
ptured by the bottom layers of a CNN where our proposed style-mixing takes place
. Mixing styles of training instances results in novel domains being synthesized
implicitly, which increase the domain diversity of the source domains, and henc
e the generalizability of the trained model. MixStyle fits into mini-batch train
ing perfectly and is extremely easy to implement. The effectiveness of MixStyle
is demonstrated on a wide range of tasks including category classification, inst
ance retrieval and reinforcement learning.
****

Dynamic Backdoor Attacks Against Deep Neural Networks
Ahmed Salem,Rui Wen,Michael Backes,Shiqing Ma,Yang Zhang
Current Deep Neural Network (DNN) backdooring attacks rely on adding static trig
gers (with fixed patterns and locations) on model inputs that are prone to detec
tion. In this paper, we propose the first class of dynamic backdooring techniqu
es: Random Backdoor, Backdoor Generating Network (BaN), and conditional Backdoor
Generating Network (c-BaN). Triggers generated by our techniques have random pa
tterns and locations. In particular, BaN and c-BaN based on a novel generative n
etwork are the first two schemes that algorithmically generate triggers. Moreov
er, c-BaN is the first conditional backdooring technique that given a target lab
el, it can generate a target-specific trigger. Both BaN and c-BaN are essential
ly a general framework which renders the adversary the flexibility for further c
ustomizing backdoor attacks. We extensively evaluate our techniques on three ben
chmark datasets and show that our techniques achieve almost perfect attack perfo
rmance on backdoored data with a negligible utility loss. More importantly, our
techniques can bypass state-of-the-art defense mechanisms.
****

Efficient Inference of Flexible Interaction in Spiking-neuron Networks
Feng Zhou,Yixuan Zhang,Jun Zhu
Hawkes process provides an effective statistical framework for analyzing the tim
e-dependent interaction of neuronal spiking activities. Although utilized in man
y real applications, the classic Hawkes process is incapable of modelling inhibi
tory interactions among neurons. Instead, the nonlinear Hawkes process allows fo
r a more flexible influence pattern with excitatory or inhibitory interactions.
In this paper, three sets of auxiliary latent variables (Polya-Gamma variables,
latent marked Poisson processes and sparsity variables) are augmented to make fu
nctional connection weights in a Gaussian form, which allows for a simple iterat
ive algorithm with analytical updates. As a result, an efficient expectation-max
imization (EM) algorithm is derived to obtain the maximum a posteriori (MAP) est
imate. We demonstrate the accuracy and efficiency performance of our algorithm o
n synthetic and real data. For real neural recordings, we show our algorithm can
estimate the temporal dynamics of interaction and reveal the interpretable func
tional connectivity underlying neural spike trains.
****

Deep Clustering and Representation Learning that Preserves Geometric Structures
Lirong Wu,Zicheng Liu,Zelin Zang,Jun Xia,Siyuan Li,Stan Z. Li
In this paper, we propose a novel framework for Deep Clustering and multimanifol
d Representation Learning (DCRL) that preserves the geometric structure of data.
In the proposed DCRL framework, manifold clustering is done in the latent space
guided by a clustering loss. To overcome the problem that clusteringoriented lo
sses may deteriorate the geometric structure of embeddings in the latent space,
an isometric loss is proposed for preserving intra-manifold structure locally an
d a ranking loss for inter-manifold structure globally. Experimental results on
various datasets show that the DCRL framework leads to performances comparable t
o current state-of-the-art deep clustering algorithms, yet exhibits superior per

formance for manifold representation. Our results also demonstrate the importanc
e and effectiveness of the proposed losses in preserving geometric structure in
terms of visualization and performance metrics. The code is provided in the Supp
lementary Material.
****
RMIX: Risk-Sensitive Multi-Agent Reinforcement Learning
Wei Qiu,Xinrun Wang,Runsheng Yu,Xu He,Rundong Wang,Bo An,Svetlana Obraztsova,Zin
ovi Rabinovich
Centralized training with decentralized execution (CTDE) has become an important
 paradigm in multi-agent reinforcement learning (MARL). Current CTDE-based metho
ds rely on restrictive decompositions of the centralized value function across a
gents, which decomposes the global Q-value into individual Q values to guide ind
ividuals' behaviours. However, such expected, i.e., risk-neutral, Q value decomp
osition is not sufficient even with CTDE due to the randomness of rewards and th
e uncertainty in environments, which causes the failure of these methods to trai
n coordinating agents in complex environments. To address these issues, we propo
se RMIX, a novel cooperative MARL method with the Conditional Value at Risk (CVa
R) measure over the learned distributions of individuals' Q values. Our main con
tributions are in three folds: (i) We first learn the return distributions of in
dividuals to analytically calculate CVaR for decentralized execution; (ii) We th
en propose a dynamic risk level predictor for CVaR calculation to handle the tem
poral nature of the stochastic outcomes during executions; (iii) We finally prop
ose risk-sensitive Bellman equation along with Individual-Global-MAX (IGM) for M
ARL training. Empirically, we show that our method significantly outperforms sta
te-of-the-art methods on many challenging StarCraft II tasks, demonstrating sign
ificantly enhanced coordination and high sample efficiency.
****
Robust Meta-learning with Noise via Eigen-Reptile
Dong Chen,Lingfei Wu,Siliang Tang,Fangli Xu,Juncheng Li,Chang Zong,Chilie Tan,Yu
eting Zhuang
Recent years have seen a surge of interest in meta-learning techniques for tackl
ing the few-shot learning (FSL) problem. However, the meta-learner's initial mod
el is prone to meta-overfit, as there are only a few available samples with samp
ling noise. Besides, when handling the data sampled with label noise for FSL, me
ta-learner could be extremely sensitive to label noise. To address these two cha
llenges that FSL with sampling and label noise. In particular, we first cast the
 meta-overfitting problem (overfitting on sampling and label noise) as a gradien
t noise problem since few available samples cause meta-learner to overfit on exi
sting examples (clean or corrupted) of an individual task at every gradient step
. We present Eigen-Reptile (ER) that updates the meta-parameters with the main d
irection of historical task-specific parameters to alleviate gradient noise. Spe
cifically, the main direction is computed by a special mechanism for the paramet
er's large size. Furthermore, to obtain a more accurate main direction for Eigen
-Reptile in the presence of label noise, we propose Introspective Self-paced Lea
rning (ISPL) that constructs a plurality of prior models to determine which samp
le should be abandoned. We have proved the effectiveness of Eigen-Reptile and IS
PL, respectively, theoretically and experimentally. Moreover, our experiments on
 different tasks demonstrate that the proposed methods outperform or achieve hig
hly competitive performance compared with the state-of-the-art methods with or w
ithout noisy labels.
****
DICE: Diversity in Deep Ensembles via Conditional Redundancy Adversarial Estimat
ion
Alexandre Rame,Matthieu Cord
Deep ensembles perform better than a single network thanks to the diversity amon
g their members. Recent approaches regularize predictions to increase diversity;
 however, they also drastically decrease individual members’ performances. In th
is paper, we argue that learning strategies for deep ensembles need to tackle th
e trade-off between ensemble diversity and individual accuracies. Motivated by a
rguments from information theory and leveraging recent advances in neural estima

tion of conditional mutual information, we introduce a novel training criterion called DICE: it increases diversity by reducing spurious correlations among features. The main idea is that features extracted from pairs of members should only share information useful for target class prediction without being conditionally redundant. Therefore, besides the classification loss with information bottleneck, we adversarially prevent features from being conditionally predictable from each other. We manage to reduce simultaneous errors while protecting class information. We obtain state-of-the-art accuracy results on CIFAR-10/100: for example, an ensemble of 5 networks trained with DICE matches an ensemble of 7 networks trained independently. We further analyze the consequences on calibration, uncertainty estimation, out-of-distribution detection and online co-distillation.
****
Universal Weakly Supervised Segmentation by Pixel-to-Segment Contrastive Learning
Tsung-Wei Ke,Jyh-Jing Hwang,Stella Yu
Weakly supervised segmentation requires assigning a label to every pixel based on training instances with partial annotations such as image-level tags, object bounding boxes, labeled points and scribbles. This task is challenging, as coarse annotations (tags, boxes) lack precise pixel localization whereas sparse annotations (points, scribbles) lack broad region coverage. Existing methods tackle these two types of weak supervision differently: Class activation maps are used to localize coarse labels and iteratively refine the segmentation model, whereas conditional random fields are used to propagate sparse labels to the entire image.

We formulate weakly supervised segmentation as a semi-supervised metric learning problem, where pixels of the same (different) semantics need to be mapped to the same (distinctive) features. We propose 4 types of contrastive relationships between pixels and segments in the feature space, capturing low-level image similarity, semantic annotation, co-occurrence, and feature affinity They act as priors; the pixel-wise feature can be learned from training images with any partial annotations in a data-driven fashion. In particular, unlabeled pixels in training images participate not only in data-driven grouping within each image, but also in discriminative feature learning within and across images. We deliver a universal weakly supervised segmenter with significant gains on Pascal VOC and DensePose. Our code is publicly available at https://github.com/twke18/SPML.
****
Learning the Connections in Direct Feedback Alignment
Matthew Bailey Webster,Jonghyun Choi,changwook Ahn
Feedback alignment was proposed to address the biological implausibility of the backpropagation algorithm which requires the transportation of the weight transpose during the backwards pass. The idea was later built upon with the proposal of direct feedback alignment (DFA), which propagates the error directly from the output layer to each hidden layer in the backward path using a fixed random weight matrix. This contribution was significant because it allowed for the parallelization of the backwards pass by the use of these feedback connections. However, just as feedback alignment, DFA does not perform well in deep convolutional networks. We propose to learn the backward weight matrices in DFA, adopting the methodology of Kolen-Pollack learning, to improve training and inference accuracy in deep convolutional neural networks by updating the direct feedback connections such that they come to estimate the forward path. The proposed method improves the accuracy of learning by direct feedback connections and reduces the gap between parallel training to serial training by means of backpropagation.
****
Discovering Parametric Activation Functions
Garrett Bingham,Risto Miikkulainen
Recent studies have shown that the choice of activation function can significantly affect the performance of deep learning networks. However, the benefits of novel activation functions have been inconsistent and task-dependent, and therefore the rectified linear unit (ReLU) is still the most commonly used. This paper p

roposes a technique for customizing activation functions automatically, resultin g in reliable improvements in performance. Evolutionary search is used to discov er the general form of the function, and gradient descent to optimize its parame ters for different parts of the network and over the learning process. Experimen ts with three different neural network architectures on the CIFAR-100 image clas sification dataset show that this approach is effective. It discovers different activation functions for different architectures, and consistently improves accu racy over ReLU and other recently proposed activation functions by significant m argins. The approach can therefore be used as an automated optimization step in applying deep learning to new tasks.
****

SVMax: A Feature Embedding Regularizer
Ahmed Taha,Alex Hanson,Abhinav Shrivastava,Larry S. Davis
A neural network regularizer (eg, weight decay) boosts performance by explicitly penalizing the complexity of a network. In this paper, we penalize inferior net work activations -- feature embeddings -- which in turn regularize the network's weights implicitly. We propose singular value maximization (SVMax) to learn a u niform feature embedding. The SVMax regularizer integrates seamlessly with both supervised and unsupervised learning. During training, our formulation mitigates model collapse and enables larger learning rates. Thus, our formulation converg es in fewer epochs, which reduces the training computational cost. We evaluate t he SVMax regularizer using both retrieval and generative adversarial networks. W e leverage a synthetic mixture of Gaussians dataset to evaluate SVMax in an unsu pervised setting. For retrieval networks, SVMax achieves significant improvement margins across various ranking losses.
****

Gradient Descent Resists Compositionality
Yuanpeng Li,Liang Zhao,Joel Hestness,Kenneth Church,Mohamed Elhoseiny
In this paper, we argue that gradient descent is one of the reasons that make co mpositionality learning hard during neural network optimization. We find that th e optimization process imposes a bias toward non-compositional solutions. This i s caused by gradient descent, trying to use all available and redundant informat ion from input, violating the conditional independence property of compositional ity. Based on this finding, we suggest that compositionality learning approaches considering only model architecture design are unlikely to achieve complete com positionality. This is the first work to investigate the relation between compos itional learning and gradient descent. We hope this study provides novel insight s into compositional generalization, and forms a basis for new research directio ns to equip machine learning models with such skills for human-level intelligenc e.
****

Shape-Texture Debiased Neural Network Training
Yingwei Li,Qihang Yu,Mingxing Tan,Jieru Mei,Peng Tang,Wei Shen,Alan Yuille,cihan g xie
Shape and texture are two prominent and complementary cues for recognizing objec ts. Nonetheless, Convolutional Neural Networks are often biased towards either t exture or shape, depending on the training dataset. Our ablation shows that such bias degenerates model performance. Motivated by this observation, we develop a simple algorithm for shape-texture debiased learning. To prevent models from ex clusively attending on a single cue in representation learning, we augment train ing data with images with conflicting shape and texture information (eg, an imag e of chimpanzee shape but with lemon texture) and, most importantly, provide the corresponding supervisions from shape and texture simultaneously.

Experiments show that our method successfully improves model performance on seve ral image recognition benchmarks and adversarial robustness. For example, by tra ining on ImageNet, it helps ResNet-152 achieve substantial improvements on Image Net (+1.2%), ImageNet-A (+5.2%), ImageNet-C (+8.3%) and Stylized-ImageNet (+11. 1%), and on defending against FGSM adversarial attacker on ImageNet (+14.4%). Ou r method also claims to be compatible with other advanced data augmentation stra

tegies, eg, Mixup, and CutMix. The code is available here: https://github.com/LiYingwei/ShapeTextureDebiasedTraining.
****
Interactive Weak Supervision: Learning Useful Heuristics for Data Labeling
Benedikt Boecking,Willie Neiswanger,Eric Xing,Artur Dubrawski
Obtaining large annotated datasets is critical for training successful machine learning models and it is often a bottleneck in practice. Weak supervision offers a promising alternative for producing labeled datasets without ground truth annotations by generating probabilistic labels using multiple noisy heuristics. This process can scale to large datasets and has demonstrated state of the art performance in diverse domains such as healthcare and e-commerce. One practical issue with learning from user-generated heuristics is that their creation requires creativity, foresight, and domain expertise from those who hand-craft them, a process which can be tedious and subjective. We develop the first framework for interactive weak supervision in which a method proposes heuristics and learns from user feedback given on each proposed heuristic. Our experiments demonstrate that only a small number of feedback iterations are needed to train models that achieve highly competitive test set performance without access to ground truth training labels. We conduct user studies, which show that users are able to effectively provide feedback on heuristics and that test set results track the performance of simulated oracles.
****
Learning a unified label space
Xingyi Zhou,Vladlen Koltun,Philipp Kraehenbuehl
How do we build a general and broad object detection system? We use all labels of all concepts ever annotated. These labels span many diverse datasets with potentially inconsistent semantic labels. In this paper, we show how to integrate these datasets and their semantic taxonomies in a completely automated fashion. Once integrated, we train an off-the-shelf object detector on the union of the datasets. This unified recognition system performs as well as dataset-specific models on each training domain, but generalizes much better to new unseen domains. Entries based on the presented methodology ranked first in the object detection and instance segmentation tracks of the ECCV 2020 Robust Vision Challenge.
****
MoVie: Revisiting Modulated Convolutions for Visual Counting and Beyond
Duy Kien Nguyen,Vedanuj Goswami,Xinlei Chen
This paper focuses on visual counting, which aims to predict the number of occurrences given a natural image and a query (e.g. a question or a category). Unlike most prior works that use explicit, symbolic models which can be computationally expensive and limited in generalization, we propose a simple and effective alternative by revisiting modulated convolutions that fuse the query and the image locally. Following the design of residual bottleneck, we call our method MoVie, short for Modulated conVolutional bottlenecks. Notably, MoVie reasons implicitly and holistically and only needs a single forward-pass during inference. Nevertheless, MoVie showcases strong performance for counting: 1) advancing the state-of-the-art on counting-specific VQA tasks while being more efficient; 2) outperforming prior-art on difficult benchmarks like COCO for common object counting; 3) helped us secure the first place of 2020 VQA challenge when integrated as a module for ‘number’ related questions in generic VQA models. Finally, we show evidence that modulated convolutions such as MoVie can serve as a general mechanism for reasoning tasks beyond counting.
****
MLR-SNet: Transferable LR Schedules for Heterogeneous Tasks
Jun Shu,Yanwen Zhu,Qian Zhao,Deyu Meng,Zongben Xu
The learning rate (LR) is one of the most important hyper-parameters in stochastic gradient descent (SGD) for deep neural networks (DNN) training and generalization. However, current hand-designed LR schedules need to manually pre-specify a fixed form, which limits their ability to adapt to non-convex optimization problems due to the significant variation of training dynamics. Meanwhile, it always needs to search a proper LR schedule from scratch for new tasks. To address the

se issues, we propose to parameterize LR schedules with an explicit mapping formulation, called MLR-SNet. The learnable structure brings more flexibility for MLR-SNet to learn a proper LR schedule to comply with the training dynamics of DNN. Image and text classification benchmark experiments substantiate the capability of our method for achieving proper LR schedules. Moreover, the meta-learned MLR-SNet is tuning-free plug-and-play to generalize to new heterogeneous tasks. We transfer our meta-trained MLR-SNet to tasks like different training epochs, network architectures, datasets, especially large scale ImageNet dataset, and achieve comparable performance with hand-designed LR schedules. Finally, MLR-Net can achieve better robustness when training data is biased with corrupted noise.
****

IALE: Imitating Active Learner Ensembles
Christoffer Löffler,Christopher Mutschler
Active learning (AL) prioritizes the labeling of the most informative data samples. However, the performance of AL heuristics depends on the structure of the underlying classifier model and the data. We propose an imitation learning scheme that imitates the selection of the best expert heuristic at each stage of the AL cycle in a batch-mode pool-based setting. We use DAGGER to train the policy on a dataset and later apply it to datasets from similar domains. With multiple AL heuristics as experts, the policy is able to reflect the choices of the best AL heuristics given the current state of the AL process. Our experiment on well-known datasets show that we both outperform state of the art imitation learners and heuristics.
****

Sample weighting as an explanation for mode collapse in generative adversarial networks
Aksel Wilhelm Wold Eide,Eilif Solberg,Ingebjørg Kåsen
Generative adversarial networks were introduced with a logistic MiniMax cost formulation, which normally fails to train due to saturation, and a Non-Saturating reformulation. While addressing the saturation problem, NS-GAN also inverts the generator's sample weighting, implicitly shifting emphasis from higher-scoring to lower-scoring samples when updating parameters. We present both theory and empirical results suggesting that this makes NS-GAN prone to mode dropping. We design MM-nsat, which preserves MM-GAN sample weighting while avoiding saturation by rescaling the MM-GAN minibatch gradient such that its magnitude approximates NS-GAN's gradient magnitude. MM-nsat has qualitatively different training dynamics, and on MNIST and CIFAR-10 it is stronger in terms of mode coverage, stability and FID. While the empirical results for MM-nsat are promising and favorable also in comparison with the LS-GAN and Hinge-GAN formulations, our main contribution is to show how and why NS-GAN's sample weighting causes mode dropping and training collapse.
****

Connection- and Node-Sparse Deep Learning: Statistical Guarantees
Johannes Lederer
Neural networks are becoming increasingly popular in applications, but a comprehensive mathematical understanding of their potentials and limitations is still missing. In this paper, we study the prediction accuracies of neural networks from a statistical point of view. In particular, we establish statistical prediction guarantees for deep learning with different types of sparsity-inducing regularization. Our bounds feature a mild dependence on network widths and depths, and, therefore, support the current trend toward wide and deep networks. The tools that we use in our derivations are uncommon in deep learning and, hence, might be of additional interest.
****

Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors
Linfeng Zhang,Kaisheng Ma
Knowledge distillation, in which a student model is trained to mimic a teacher model, has been proved as an effective technique for model compression and model accuracy boosting. However, most knowledge distillation methods, designed for im

age classification, have failed on more challenging tasks, such as object detect ion. In this paper, we suggest that the failure of knowledge distillation on obj ect detection is mainly caused by two reasons: (1) the imbalance between pixels of foreground and background and (2) lack of distillation on the relation betwee n different pixels. Observing the above reasons, we propose attention-guided dis tillation and non-local distillation to address the two problems, respectively. Attention-guided distillation is proposed to find the crucial pixels of foregro und objects with attention mechanism and then make the students take more effort to learn their features. Non-local distillation is proposed to enable students to learn not only the feature of an individual pixel but also the relation betwe en different pixels captured by non-local modules. Experiments show that our met hods achieve excellent AP improvements on both one-stage and two-stage, both anc hor-based and anchor-free detectors. For example, Faster RCNN (ResNet101 backbon e) with our distillation achieves 43.9 AP on COCO2017, which is 4.1 higher than the baseline. Codes have been released on Github.

****

Revisiting Dynamic Convolution via Matrix Decomposition

Yunsheng Li,Yinpeng Chen,Xiyang Dai,mengchen liu,Dongdong Chen,Ye Yu,Lu Yuan,Zic heng Liu,Mei Chen,Nuno Vasconcelos

Recent research in dynamic convolution shows substantial performance boost for e fficient CNNs, due to the adaptive aggregation of K static convolution kernels. It has two limitations: (a) it increases the number of convolutional weights by K-times, and (b) the joint optimization of dynamic attention and static convolut ion kernels is challenging. In this paper, we revisit it from a new perspective of matrix decomposition and reveal the key issue is that dynamic convolution app lies dynamic attention over channel groups after projecting into a higher dimens ional latent space. To address this issue, we propose dynamic channel fusion to replace dynamic attention over channel groups. Dynamic channel fusion not only e nables significant dimension reduction of the latent space, but also mitigates t he joint optimization difficulty. As a result, our method is easier to train and requires significantly fewer parameters without sacrificing accuracy. Source co de is at https://github.com/liyunsheng13/dcd.

****

GraphSAD: Learning Graph Representations with Structure-Attribute Disentanglemen t

Minghao Xu,Hang Wang,Bingbing Ni,Wenjun Zhang,Jian Tang

Graph Neural Networks (GNNs) learn effective node/graph representations by aggre gating the attributes of neighboring nodes, which commonly derives a single repr esentation mixing the information of graph structure and node attributes. Howeve r, these two kinds of information might be semantically inconsistent and could b e useful for different tasks. In this paper, we aim at learning node/graph repre sentations with Structure-Attribute Disentanglement (GraphSAD). We propose to di sentangle graph structure and node attributes into two distinct sets of represen tations, and such disentanglement can be done in either the input or the embeddi ng space. We further design a metric to quantify the extent of such a disentangl ement. Extensive experiments on multiple datasets show that our approach can ind eed disentangle the semantics of graph structure and node attributes, and it ach ieves superior performance on both node and graph classification tasks.

****

Self-supervised Graph-level Representation Learning with Local and Global Struct ure

Minghao Xu,Hang Wang,Bingbing Ni,Hongyu Guo,Jian Tang

This paper focuses on unsupervised/self-supervised whole-graph representation le arning, which is critical in many tasks including drug and material discovery. C urrent methods can effectively model the local structure between different graph instances, but they fail to discover the global semantic structure of the entir e dataset. In this work, we propose a unified framework called Local-instance an d Global-semantic Learning (GraphLoG) for self-supervised whole-graph representa tion learning. Specifically, besides preserving the local instance-level structu re, GraphLoG leverages a nonparametric strategy to learn hierarchical prototypes

of the data. These prototypes capture the semantic clusters in the latent space , and the number of prototypes can automatically adapt to different feature dist ributions. We evaluate GraphLoG by pre-training it on massive unlabeled graphs f ollowed by fine-tuning on downstream tasks. Extensive experiments on both chemic al and biological benchmark datasets demonstrate the effectiveness of our approa ch.
****

Towards Multi-Sense Cross-Lingual Alignment of Contextual Embeddings
Linlin Liu,Thien Hai Nguyen,Shafiq Joty,Lidong Bing,Luo Si
Cross-lingual word embeddings (CLWE) have been proven useful in many cross-lingu al tasks. However, most existing approaches to learn CLWE including the ones wit h contextual embeddings are sense agnostic. In this work, we propose a novel fra mework to align contextual embeddings at the sense level by leveraging cross-lin gual signal from bilingual dictionaries only. We operationalize our framework by first proposing a novel sense-aware cross entropy loss to model word senses exp licitly. The monolingual ELMo and BERT models pretrained with our sense-aware cr oss entropy loss demonstrate significant performance improvement for word sense disambiguation tasks. We then propose a sense alignment objective on top of the sense-aware cross entropy loss for cross-lingual model pretraining, and pretrain cross-lingual models for several language pairs (English to German/Spanish/Japa nese/Chinese). Compared with the best baseline results, our cross-lingual models achieve 0.52%, 2.09% and 1.29% average performance improvements on zero-shot cr oss-lingual NER, sentiment classification and XNLI tasks, respectively. We will release our code.
****

Improving Adversarial Robustness via Channel-wise Activation Suppressing
Yang Bai,Yuyuan Zeng,Yong Jiang,Shu-Tao Xia,Xingjun Ma,Yisen Wang
The study of adversarial examples and their activations have attracted significa nt attention for secure and robust learning with deep neural networks (DNNs). D ifferent from existing works, in this paper, we highlight two new characteristic s of adversarial examples from the channel-wise activation perspective: 1) the activation magnitudes of adversarial examples are higher than that of natural ex amples; and 2) the channels are activated more uniformly by adversarial examples than natural examples. We find that, while the state-of-the-art defense adversa rial training has addressed the first issue of high activation magnitude via tra ining on adversarial examples, the second issue of uniform activation remains. This motivates us to suppress redundant activations from being activated by adve rsarial perturbations during the adversarial training process, via a Channel-wis e Activation Suppressing (CAS) training strategy. We show that CAS can train a model that inherently suppresses adversarial activations, and can be easily appl ied to existing defense methods to further improve their robustness. Our work pr ovides a simplebut generic training strategy for robustifying the intermediate l ayer activations of DNNs.
****

DynaTune: Dynamic Tensor Program Optimization in Deep Neural Network Compilation
Minjia Zhang,Menghao Li,Chi Wang,Mingqin Li
Recently, the DL compiler, together with Learning to Compile has proven to be a powerful technique for optimizing deep learning models. However, existing method s focus on accelerating the convergence speed of the individual tensor operator rather than the convergence speed of the entire model, which results in long opt imization time to obtain a desired latency.

In this paper, we present a new method called DynaTune, which provides significa ntly faster convergence speed to optimize a DNN model. In particular, we conside r a Multi-Armed Bandit (MAB) model for the tensor program optimization problem. We use UCB to handle the decision-making of time-slot-based optimization, and we devise a Bayesian belief model that allows predicting the potential performance gain of each operator with uncertainty quantification, which guides the optimiz ation process. We evaluate and compare DynaTune with the state-of-the-art DL com piler. The experiment results show that DynaTune is 1.2--2.4 times faster to ach

ieve the same optimization quality for a range of models across different hardwa
re architectures.
****
Learning Latent Topology for Graph Matching
Tianshu Yu,Runzhong Wang,Junchi Yan,Baoxin Li
Graph matching (GM) has been traditionally modeled as a deterministic optimizati
on problem characterized by an affinity matrix under pre-defined graph topology.
 Though there have been several attempts on learning more effective node-level a
ffinity/representation for matching, they still heavily rely on the initial grap
h structure/topology which is typically obtained through heuristic ways (e.g. De
launey or k-nearest) and will not be adjusted during the learning process to a
dapt to problem-specific patterns. We argue that a standalone graph representati
on learning is insufficient for GM task, whereby a GM solver may favor some late
nt topology other than pre-defined one. Motivated by this hypothesis, we propose
 to learn latent graph topology in replacement of the fixed topology as input. T
o this end, we devise two types of latent graph generation procedures in determi
nistic and generative fashion, respectively. Particularly, the generative proced
ure emphasizes the across-graph consistency and thus can be viewed as a \textbf{
co-generative} model. Our methods show superior performance over previous state-
of-the-arts on several benchmarks, thus strongly supporting our hypothesis.
****
Decoupling Representation Learning from Reinforcement Learning
Adam Stooke,Kimin Lee,Pieter Abbeel,Michael Laskin
In an effort to overcome limitations of reward-driven feature learning in deep r
einforcement learning (RL) from images, we propose decoupling representation lea
rning from policy learning. To this end, we introduce a new unsupervised learni
ng (UL) task, called Augmented Temporal Contrast (ATC), which trains a convoluti
onal encoder to associate pairs of observations separated by a short time differ
ence, under image augmentations and using a contrastive loss. In online RL expe
riments, we show that training the encoder exclusively using ATC matches or outp
erforms end-to-end RL in most environments. Additionally, we benchmark several
leading UL algorithms by pre-training encoders on expert demonstrations and usin
g them, with weights frozen, in RL agents; we find that agents using ATC-trained
 encoders outperform all others. We also train multi-task encoders on data from
 multiple environments and show generalization to different downstream RL tasks.
  Finally, we ablate components of ATC, and introduce a new data augmentation to
 enable replay of (compressed) latent images from pre-trained encoders when RL r
equires augmentation. Our experiments span visually diverse RL benchmarks in De
epMind Control, DeepMind Lab, and Atari, and our complete code is available at \
url{hidden url}.
****
MALI: A memory efficient and reverse accurate integrator for Neural ODEs
Juntang Zhuang,Nicha C Dvornek,sekhar tatikonda,James s Duncan
Neural ordinary differential equations (Neural ODEs) are a new family of deep-le
arning models with continuous depth. However, the numerical estimation of the gr
adient in the continuous case is not well solved: existing implementations of th
e adjoint method suffer from inaccuracy in reverse-time trajectory, while the na
ive method and the adaptive checkpoint adjoint method (ACA) have a memory cost t
hat grows with integration time. In this project, based on the asynchronous leap
frog (ALF) solver, we propose the Memory-efficient ALF Integrator (MALI), which
has a constant memory cost $w.r.t$ integration time similar to the adjoint metho
d, and guarantees accuracy in reverse-time trajectory (hence accuracy in gradien
t estimation). We validate MALI in various tasks: on image recognition tasks, to
 our knowledge, MALI is the first to enable feasible training of a Neural ODE on
 ImageNet and outperform a well-tuned ResNet, while existing methods fail due to
 either heavy memory burden or inaccuracy; for time series modeling, MALI signif
icantly outperforms the adjoint method; and for continuous generative models, MA
LI achieves new state-of-the-art performance. We provide a pypi package: https:/
/jzkay12.github.io/TorchDiffEqPack
****

CROSS-SUPERVISED OBJECT DETECTION
Zitian Chen,Zhiqiang Shen,Jiahui Yu,Erik Learned-Miller
After learning a new object category from image-level annotations (with no objec t bounding boxes), humans are remarkably good at precisely localizing those obje cts. However, building good object localizers (i.e., detectors) currently requir es expensive instance-level annotations. While some work has been done on learni ng detectors from weakly labeled samples (with only class labels), these detecto rs do poorly at localization. In this work, we show how to build better object d etectors from weakly labeled images of new categories by leveraging knowledge le arned from fully labeled base categories. We call this learning paradigm cross-s upervised object detection. While earlier works investigated this paradigm, they did not apply it to realistic complex images (e.g., COCO), and their performanc e was poor. We propose a unified framework that combines a detection head traine d from instance-level annotations and a recognition head learned from image-leve l annotations, together with a spatial correlation module that bridges the gap b etween detection and recognition. These contributions enable us to better detect novel objects with image-level annotations in complex multi-object scenes such as the COCO dataset.
****
Weak NAS Predictor Is All You Need
Junru Wu,Xiyang Dai,Dongdong Chen,Yinpeng Chen,Mengchen Liu,Ye Yu,Zhangyang Wang ,Zicheng Liu,Mei Chen,Lu Yuan
Neural Architecture Search (NAS) finds the best network architecture by explorin g the architecture-to-performance manifold. It often trains and evaluates a larg e amount of architectures, causing tremendous computation cost. Recent predictor -based NAS approaches attempt to solve this problem with two key steps: sampling some architecture-performance pairs and fitting a proxy accuracy predictor. Exi sting predictors attempt to model the performance distribution over the whole ar chitecture space, which could be too challenging given limited samples. Instead, we envision that this ambitious goal may not be necessary if the final aim is t o find the best architecture. We present a novel framework to estimate weak pred ictors progressively. Rather than expecting a single strong predictor to model t he whole space, we seek a progressive line of weak predictors that can connect a path to the best architecture, thus greatly simplifying the learning task of ea ch predictor. It is based on the key property of the predictors that their proba bilities of sampling better architectures will keep increasing. We thus only sam ple a few well-performed architectures guided by the predictive model, to estima te another better weak predictor. By this coarse-to-fine iteration, the ranking of sampling space is refined gradually, which helps find the optimal architectur es eventually. Experiments demonstrate that our method costs fewer samples to fi nd the top-performance architectures on NAS-Bench-101 and NAS-Bench-201, and it achieves the state-of-the-art ImageNet performance on the NASNet search space.
****
VECO: Variable Encoder-decoder Pre-training for Cross-lingual Understanding and Generation
Fuli Luo,Wei Wang,Jiahao Liu,Yijia Liu,Bin Bi,Songfang Huang,Fei Huang,Luo Si
Recent studies about learning multilingual representations have achieved signifi cant performance gains across a wide range of downstream cross-lingual tasks. Th ey train either an encoder-only Transformer mainly for understanding tasks, or a n encoder-decoder Transformer specifically for generation tasks, ignoring the co rrelation between the two tasks and frameworks. In contrast, this paper presents a variable encoder-decoder (VECO) pre-training approach to unify the two mainst reams in both model architectures and pre-training tasks. VECO splits the standa rd Transformer block into several sub-modules trained with both inner-sequence a nd cross-sequence masked language modeling, and correspondingly reorganizes cert ain sub-modules for understanding and generation tasks during inference. Such a workflow not only ensures to train the most streamlined parameters necessary for two kinds of tasks, but also enables them to boost each other via sharing commo n sub-modules. As a result, VECO delivers new state-of-the-art results on variou s cross-lingual understanding tasks of the XTREME benchmark covering text classi

fication, sequence labeling, question answering, and sentence retrieval. For gen eration tasks, VECO also outperforms all existing cross-lingual models and state -of-the-art Transformer variants on WMT14 English-to-German and English-to-Frenc h translation datasets, with gains of up to 1~2 BLEU.
****

SelfNorm and CrossNorm for Out-of-Distribution Robustness
Zhiqiang Tang,Yunhe Gao,Yi Zhu,Zhi Zhang,Mu Li,Dimitris N. Metaxas
Normalization techniques are crucial in stabilizing and accelerating the trainin g of deep neural networks. However, they are mainly designed for the independent and identically distributed (IID) data, not satisfying many real-world out-of-d istribution (OOD) situations. Unlike most previous works, this paper presents tw o normalization methods, SelfNorm and CrossNorm, to promote OOD generalization. SelfNorm uses attention to recalibrate statistics (channel-wise mean and varianc e), while CrossNorm exchanges the statistics between feature maps. SelfNorm and CrossNorm can complement each other in OOD generalization, though exploring diff erent directions in statistics usage. Extensive experiments on different domains (vision and language), tasks (classification and segmentation), and settings (s upervised and semi-supervised) show their effectiveness.
****

Meta-Model-Based Meta-Policy Optimization
Takuya Hiraoka,Takahisa Imagawa,Voot Tangkaratt,Takayuki Osa,Takashi Onishi,Yosh imasa Tsuruoka
Model-based reinforcement learning (MBRL) has been applied to meta-learning sett ings and has demonstrated its high sample efficiency.
However, in previous MBRL for meta-learning settings, policies are optimized via rollouts that fully rely on a predictive model of an environment.
Thus, its performance in a real environment tends to degrade when the predictive model is inaccurate.
In this paper, we prove that performance degradation can be suppressed by using branched meta-rollouts.
On the basis of this theoretical analysis, we propose Meta-Model-based Meta-Poli cy Optimization (M3PO), in which the branched meta-rollouts are used for policy optimization.
We demonstrate that M3PO outperforms existing meta reinforcement learning method s in continuous-control benchmarks.
****

On the Marginal Regret Bound Minimization of Adaptive Methods
Wenjie Li,Guang Cheng
Numerous adaptive algorithms such as AMSGrad and Radam have been proposed and ap plied to deep learning recently. However, these modifications do not improve the convergence rate of adaptive algorithms and whether a better algorithm exists s till remains an open question. In this work, we propose a new motivation for des igning the proximal function of adaptive algorithms, named as marginal regret bo und minimization. Based on such an idea, we propose a new class of adaptive algo rithms that not only achieves marginal optimality but can also potentially conve rge much faster than any existing adaptive algorithms in the long term. We show the superiority of the new class of adaptive algorithms both theoretically and e mpirically using experiments in deep learning.
****

Bag of Tricks for Adversarial Training
Tianyu Pang,Xiao Yang,Yinpeng Dong,Hang Su,Jun Zhu
Adversarial training (AT) is one of the most effective strategies for promoting model robustness. However, recent benchmarks show that most of the proposed impr ovements on AT are less effective than simply early stopping the training proced ure. This counter-intuitive fact motivates us to investigate the implementation details of tens of AT methods. Surprisingly, we find that the basic settings (e. g., weight decay, training schedule, etc.) used in these methods are highly inco nsistent. In this work, we provide comprehensive evaluations on CIFAR-10, focusi ng on the effects of mostly overlooked training tricks and hyperparameters for a dversarially trained models. Our empirical observations suggest that adversarial

robustness is much more sensitive to some basic training settings than we thoug
ht. For example, a slightly different value of weight decay can reduce the model
robust accuracy by more than 7%, which is probable to override the potential
promotion induced by the proposed methods. We conclude a baseline training sett
ing and re-implement previous defenses to achieve new state-of-the-art results.
These facts also appeal to more concerns on the overlooked confounders when benc
hmarking defenses.
****
Rethinking Sampling in 3D Point Cloud Generative Adversarial Networks
He Wang,Zetian Jiang,Li Yi,Kaichun Mo,Hao Su,Leonidas Guibas
In this paper, we examine the long-neglected yet important effects of point sam-
pling patterns in point cloud GANs. Through extensive experiments, we show that
sampling-insensitive discriminators (e.g. PointNet-Max) produce shape point clo
uds with point clustering artifacts while sampling-oversensitive discriminators
(e.g. PointNet++, DGCNN, PointConv, KPConv) fail to guide valid shape generation
. We propose the concept of sampling spectrum to depict the different sampling s
ensitivities of discriminators. We further study how different evaluation metric
s weigh the sampling pattern against the geometry and propose several perceptual
metrics forming a sampling spectrum of metrics. Guided by the proposed sampling
spectrum, we discover a middle-point sampling-aware baseline discriminator, Poi
ntNet-Mix, which improves all existing point cloud generators by a large margin
on sampling-related metrics. We point out that, given that recent research has b
een focused on the generator design, the discriminator design needs more attenti
on. Our work provides both suggestions and tools for building future discriminat
ors. We will release the code to facilitate future research.
****
Robustness against Relational Adversary
Yizhen Wang,Xiaozhu Meng,Mihai Christodorescu,Somesh Jha,Ke Wang
Test-time adversarial attacks have posed serious challenges to the robustness of
machine-learning models, and in many settings the adversarial perturbation need
not be bounded by small lp-norms. Motivated by the semantics-preserving attack
s in vision and security domain, we investigate relational adversaries, a broad
class of attackers who create adversarial examples that are in a re■exive-transi
tive closure of a logical relation. We analyze the conditions for robustness and
propose normalize-and-predict – a learning framework with provable robustness g
uarantee. We compare our approach with adversarial training and derive an uni■ed
framework that provides bene■ts of both approaches. Guided by our theoretical ■
ndings, we apply our framework to malware detection and image classi■cation. Res
ults of both tasks show that attacks using relational adversaries frequently foo
l existing models, but our uni■ed framework can signi■cantly enhance their robus
tness.

****
Self-Pretraining for Small Datasets by Exploiting Patch Information
Zhang Chunyang
Deep learning tasks with small datasets are often tackled by pretraining model
s with large datasets on relevent tasks. Although pretraining methods mitigate t
he problem of overfitting, it can be difficult to find appropriate pretrained mo
dels sometimes. In this paper, we proposed a self-pretraininng method by exploit
ing patch information in the dataset itself without pretraining on other dataset
s. Our experiments show that the self-pretraining method leads to better perform
ance than training from scratch both in the condition of not using other data.
****
Dual Graph Complementary Network
Chenhua Liu,Kun Zhan
As a powerful representation learning method on graph data, graph neural network
s (GNNs) have shown great popularity in tackling graph analytic problems. Althou
gh many attempts have been made in literatures to find strategies about extracti
ng better embedding of the target nodes, few of them consider this issue from a
comprehensive perspective. Most of current GNNs usually employ some single metho

d which can commendably extract a certain kind of feature but some equally impor tant features are often ignored. In this paper, we develop a novel dual graph co mplementary network (DGCN) to learn representation complementarily. We use two d ifferent branches, and inputs of the two branches are the same, which are compos ed of structure and feature information. At the same time, there is also a compl ementary relationship between the two branches. Beyond that, our extensive exper iments show that DGCN outperforms state-of-the-art methods on five public benchm ark datasets.
****

Graph View-Consistent Learning Network
Zhuolin Liao,Kun Zhan
Recent years, methods based on neural networks have made great achievements in s olving large and complex graph problems. However, high efficiency of these metho ds depends on large training and validation sets, while the acquisition of groun d-truth labels is expensive and time-consuming. In this paper, a graph view-cons istent learning network (GVCLN) is specially designed for the semi-supervised le arning when the number of the labeled samples is very small. We fully exploit th e neighborhood aggregation capability of GVCLN and use dual views to obtain diff erent representations. Although the two views have different viewing angles, the ir observation objects are the same, so their observation representations need t o be consistent. For view-consistent representations between two views, two loss functions are designed besides a supervised loss. The supervised loss uses the known labeled set, while a view-consistent loss is applied to the two views to o btain the consistent representation and a pseudo-label loss is designed by using the common high-confidence predictions. GVCLN with these loss functions can obt ain the view-consistent representations of the original feature. We also find th at preprocessing the node features with specific filter before training is good for subsequent classification tasks. Related experiments have been done on the t hree citation network datasets of Cora, Citeseer, and PubMed. On several node cl assification tasks, GVCLN achieves state-of-the-art performance.
****

Robust Loss Functions for Complementary Labels Learning
Defu Liu,Guowu Yang
In ordinary-label learning, the correct label is given to each training sample. Similarly, a complementary label is also provided for each training sample in co mplementary-label learning. A complementary label indicates a class that the exa mple does not belong to. Robust learning of classifiers has been investigated fr om many viewpoints under label noise, but little attention has been paid to comp lementary-label learning. In this paper, we present a new algorithm of complemen tary-label learning with the robustness of loss function. We also provide two su fficient conditions on a loss function so that the minimizer of the risk for com plementary labels is theoretically guaranteed to be consistent with the minimize r of the risk for ordinary labels. Finally, the empirical results validate our m ethod’s superiority to current state-of-the-art techniques. Especially in cifar1 0, our algorithm achieves a much higher test accuracy than the gradient ascent a lgorithm, and the parameters of our model are less than half of the ResNet-34 th ey used.
****

Learning to Generate 3D Shapes with Generative Cellular Automata
Dongsu Zhang,Changwoon Choi,Jeonghwan Kim,Young Min Kim
In this work, we present a probabilistic 3D generative model, named Generative C ellular Automata, which is able to produce diverse and high quality shapes. We f ormulate the shape generation process as sampling from the transition kernel of a Markov chain, where the sampling chain eventually evolves to the full shape of the learned distribution. The transition kernel employs the local update rules of cellular automata, effectively reducing the search space in a high-resolution 3D grid space by exploiting the connectivity and sparsity of 3D shapes. Our pro gressive generation only focuses on the sparse set of occupied voxels and their neighborhood, thus enables the utilization of an expressive sparse convolutional network. We propose an effective training scheme to obtain the local homogeneou

s rule of generative cellular automata with sequences that are slightly differen
t from the sampling chain but converge to the full shapes in the training data.
Extensive experiments on probabilistic shape completion and shape generation dem
onstrate that our method achieves competitive performance against recent methods
.
****
Learning Graph Normalization for Graph Neural Networks
Yihao Chen,Xin Tang,Xianbiao Qi,Chun-Guang Li,Rong Xiao
Graph Neural Networks (GNNs) have emerged as a useful paradigm to process graph-
structured data. Usually, GNNs are stacked to multiple layers and the node repre
sentations in each layer are computed through propagating and aggre- gating the
neighboring node features with respect to the graph. To effectively train a GNN
with multiple layers, some normalization techniques are necessary. Though the ex
isting normalization techniques have been shown to accelerate train- ing GNNs, t
he structure information on the graph is ignored yet. In this paper, we propose
two graph-aware normalization methods to effectively train GNNs. Then, by taking
 into account that normalization methods for GNNs are highly task-relevant and i
t is hard to know in advance which normalization method is the best, we propose
to learn attentive graph normalization by optimizing a weighted combination of m
ultiple graph normalization methods at different scales. By op- timizing the com
bination weights, we can automatically select the best or the best combination o
f multiple normalization methods for a specific task. We con- duct extensive exp
eriments on benchmark datasets for different tasks and confirm that the graph-aw
are normalization methods lead to promising results and that the learned weights
 suggest the more appropriate normalization methods for specific task
****
SEED: Self-supervised Distillation For Visual Representation
Zhiyuan Fang,Jianfeng Wang,Lijuan Wang,Lei Zhang,Yezhou Yang,Zicheng Liu
This paper is concerned with self-supervised learning for small models. The prob
lem is motivated by our empirical studies that while the widely used contrastive
 self-supervised learning method has shown great progress on large model trainin
g, it does not work well for small models. To address this problem, we propose a
 new learning paradigm, named $\textbf{SE}$lf-Sup$\textbf{E}$rvised $\textbf{D}$
istillation (${\large S}$EED), where we leverage a larger network (as Teacher) t
o transfer its representational knowledge into a smaller architecture (as Studen
t) in a self-supervised fashion. Instead of directly learning from unlabeled dat
a, we train a student encoder to mimic the similarity score distribution inferre
d by a teacher over a set of instances. We show that ${\large S}$EED dramaticall
y boosts the performance of small networks on downstream tasks. Compared with se
lf-supervised baselines, ${\large S}$EED improves the top-1 accuracy from 42.2%
to 67.6% on EfficientNet-B0 and from 36.3% to 68.2% on MobileNet-v3-Large on the
 ImageNet-1k dataset.
****
Long-tailed Recognition by Routing Diverse Distribution-Aware Experts
Xudong Wang,Long Lian,Zhongqi Miao,Ziwei Liu,Stella Yu
Natural data are often long-tail distributed over semantic classes. Existing re
cognition methods tackle this imbalanced classification by placing more emphasis
 on the tail data, through class re-balancing/re-weighting or ensembling over di
fferent data groups, resulting in increased tail accuracies but reduced head acc
uracies.
We take a dynamic view of the training data and provide a principled model bias
and variance analysis as the training data fluctuates: Existing long-tail classi
fiers invariably increase the model variance and the head-tail model bias gap re
mains large, due to more and larger confusion with hard negatives for the tail.
We propose a new long-tailed classifier called RoutIng Diverse Experts (RIDE).
It reduces the model variance with multiple experts, reduces the model bias with
 a distribution-aware diversity loss, reduces the computational cost with a dyna
mic expert routing module. RIDE outperforms the state-of-the-art by 5% to 7% on
 CIFAR100-LT, ImageNet-LT and iNaturalist 2018 benchmarks. It is also a univers
al framework that is applicable to various backbone networks, long-tailed algori

thms and training mechanisms for consistent performance gains. Our code is avail
able at: https://github.com/frank-xwang/RIDE-LongTailRecognition.
****

PDE-Driven Spatiotemporal Disentanglement
Jérémie Donà,Jean-Yves Franceschi,sylvain lamprier,patrick gallinari
A recent line of work in the machine learning community addresses the problem of
 predicting high-dimensional spatiotemporal phenomena by leveraging specific too
ls from the differential equations theory. Following this direction, we propose
in this article a novel and general paradigm for this task based on a resolution
 method for partial differential equations: the separation of variables. This in
spiration allows us to introduce a dynamical interpretation of spatiotemporal di
sentanglement. It induces a principled model based on learning disentangled spat
ial and temporal representations of a phenomenon to accurately predict future ob
servations. We experimentally demonstrate the performance and broad applicabilit
y of our method against prior state-of-the-art models on physical and synthetic
video datasets.
****

Generalization in data-driven models of primary visual cortex
Konstantin-Klemens Lurz,Mohammad Bashiri,Konstantin Willeke,Akshay Jagadish,Eric
 Wang,Edgar Y. Walker,Santiago A Cadena,Taliah Muhammad,Erick Cobos,Andreas S. T
olias,Alexander S Ecker,Fabian H. Sinz
Deep neural networks (DNN) have set new standards at predicting responses of neu
ral populations to visual input. Most such DNNs consist of a convolutional netw
ork (core) shared across all neurons which learns a representation of neural com
putation in visual cortex and a neuron-specific readout that linearly combines t
he relevant features in this representation. The goal of this paper is to test w
hether such a representation is indeed generally characteristic for visual corte
x, i.e. generalizes between animals of a species, and what factors contribute to
 obtaining such a generalizing core. To push all non-linear computations into th
e core where the generalizing cortical features should be learned, we devise a n
ovel readout that reduces the number of parameters per neuron in the readout by
up to two orders of magnitude compared to the previous state-of-the-art. It does
 so by taking advantage of retinotopy and learns a Gaussian distribution over th
e neuron’s receptive field position. With this new readout we train our network
 on neural responses from mouse primary visual cortex (V1) and obtain a gain in
performance of 7% compared to the previous state-of-the-art network. We then in
vestigate whether the convolutional core indeed captures general cortical featur
es by using the core in transfer learning to a different animal. When transferr
ing a core trained on thousands of neurons from various animals and scans we exc
eed the performance of training directly on that animal by 12%, and outperform a
 commonly used VGG16 core pre-trained on imagenet by 33%. In addition, transfer
learning with our data-driven core is more data-efficient than direct training,
achieving the same performance with only 40% of the data. Our model with its nov
el readout thus sets a new state-of-the-art for neural response prediction in mo
use visual cortex from natural images, generalizes between animals, and captures
 better characteristic cortical features than current task-driven pre-training a
pproaches such as VGG16.
****

Non-Inherent Feature Compatible Learning
Yantao Shen,Fanzi Wu,Ying Shan
The need of Feature Compatible Learning (FCL) arises from many large scale retri
eval-based applications, where updating the entire library of embedding vectors
is expensive. When an upgraded embedding model shows potential, it is desired to
 transform the benefit of the new model without refreshing the library. While pr
ogresses have been made along this new direction, existing approaches for featur
e compatible learning mostly rely on old training data and classifiers, which ar
e not available in many industry settings. In this work, we introduce an approac
h for feature compatible learning without inheriting old classifier and training
 data, i.e., Non-Inherent Feature Compatible Learning. Our approach requires onl
y features extracted by \emph{old} model's backbone and \emph{new} training data

, and makes no assumption about the overlap between old and new training data. We propose a unified framework for FCL, and extend it to handle the case where the old model is a black-box. Specifically, we learn a simple pseudo classifier in lieu of the old model, and further enhance it with a random walk algorithm. As a result, the embedding features produced by the new model can be matched with those from the old model without sacrificing performance. Experiments on ImageNet ILSVRC 2012 and Places365 data proved the efficacy of the proposed approach.
****
FLAG: Adversarial Data Augmentation for Graph Neural Networks
Kezhi Kong,Guohao Li,Mucong Ding,Zuxuan Wu,Chen Zhu,Bernard Ghanem,Gavin Taylor,Tom Goldstein
Data augmentation helps neural networks generalize better, but it remains an open question how to effectively augment graph data to enhance the performance of GNNs (Graph Neural Networks). While most existing graph regularizers focus on augmenting graph topological structures by adding/removing edges, we offer a novel direction to augment in the input node feature space for better performance. We propose a simple but effective solution, FLAG (Free Large-scale Adversarial Augmentation on Graphs), which iteratively augments node features with gradient-based adversarial perturbations during training, and boosts performance at test time. Empirically, FLAG can be easily implemented with a dozen lines of code and is flexible enough to function with any GNN backbone, on a wide variety of large-scale datasets, and in both transductive and inductive settings. Without modifying a model's architecture or training setup, FLAG yields a consistent and salient performance boost across both node and graph classification tasks. Using FLAG, we reach state-of-the-art performance on the large-scale ogbg-molpcba, ogbg-ppa, and ogbg-code datasets.
****
VEM-GCN: Topology Optimization with Variational EM for Graph Convolutional Networks
Rui Yang,Wenrui Dai,Chenglin Li,Junni Zou,Hongkai Xiong
Over-smoothing has emerged as a severe problem for node classification with graph convolutional networks (GCNs). In the view of message passing, the over-smoothing issue is caused by the observed noisy graph topology that would propagate information along inter-class edges, and consequently, over-mix the features of nodes in different classes. In this paper, we propose a novel architecture, namely VEM-GCN, to address this problem by employing the variational EM algorithm to jointly optimize the graph topology and learn desirable node representations for classification. Specifically, variational EM approaches a latent adjacency matrix parameterized by the assortative-constrained stochastic block model (SBM) to enhance intra-class connection and suppress inter-class interaction of the observed noisy graph. In the variational E-step, graph topology is optimized by approximating the posterior probability distribution of the latent adjacency matrix with a neural network learned from node embeddings. In the M-step, node representations are learned using the graph convolutional network based on the refined graph topology for the downstream task of classification. VEM-GCN is demonstrated to outperform existing strategies for tackling over-smoothing and optimizing graph topology in node classification on seven benchmark datasets.
****
Do 2D GANs Know 3D Shape? Unsupervised 3D Shape Reconstruction from 2D Image GANs
Xingang Pan,Bo Dai,Ziwei Liu,Chen Change Loy,Ping Luo
Natural images are projections of 3D objects on a 2D image plane. While state-of-the-art 2D generative models like GANs show unprecedented quality in modeling the natural image manifold, it is unclear whether they implicitly capture the underlying 3D object structures. And if so, how could we exploit such knowledge to recover the 3D shapes of objects in the images? To answer these questions, in this work, we present the first attempt to directly mine 3D geometric cues from an off-the-shelf 2D GAN that is trained on RGB images only. Through our investigation, we found that such a pre-trained GAN indeed contains rich 3D knowledge and thus can be used to recover 3D shape from a single 2D image in an unsupervised m

anner. The core of our framework is an iterative strategy that explores and expl oits diverse viewpoint and lighting variations in the GAN image manifold. The fr amework does not require 2D keypoint or 3D annotations, or strong assumptions on object shapes (e.g. shapes are symmetric), yet it successfully recovers 3D shap es with high precision for human faces, cats, cars, and buildings. The recovered 3D shapes immediately allow high-quality image editing like relighting and obje ct rotation. We quantitatively demonstrate the effectiveness of our approach com pared to previous methods in both 3D shape reconstruction and face rotation. Our code is available at https://github.com/XingangPan/GAN2Shape.
****

Rethinking Soft Labels for Knowledge Distillation: A Bias–Variance Tradeoff Pers pective
Helong Zhou,Liangchen Song,Jiajie Chen,Ye Zhou,Guoli Wang,Junsong Yuan,Qian Zhan g
Knowledge distillation is an effective approach to leverage a well-trained netwo rk or an ensemble of them, named as the teacher, to guide the training of a stud ent network. The outputs from the teacher network are used as soft labels for s upervising the training of a new network. Recent studies (M ■uller et al., 2019 ; Yuan et al., 2020) revealed an intriguing property of the soft labels that mak ing labels soft serves as a good regularization to the student network. From t he perspective of statistical learning, regularization aims to reduce the varia nce, however how bias and variance change is not clear for training with soft l abels. In this paper, we investigate the bias-variance tradeoff brought by dis tillation with soft labels. Specifically, we observe that during training the bias-variance tradeoff varies sample-wisely. Further, under the same distillati on temperature setting, we observe that the distillation performance is negative ly associated with the number of some specific samples, which are named as regul arization samples since these samples lead to bias increasing and variance decre asing. Nevertheless, we empirically find that completely filtering out regulari zation samples also deteriorates distillation performance. Our discoveries insp ired us to propose the novel weighted soft labels to help the network adaptively handle the sample-wise bias-variance tradeoff. Experiments on standard evaluat ion benchmarks validate the effectiveness of our method. Our code is available i n the supplementary.
****

Detection Booster Training: A detection booster training method for improving th e accuracy of classifiers.
Ali Ghobadzadeh,Deepak Sridhar,Juwei Lu,Wei Li
Deep learning models owe their success at large, to the availability of a large amount of annotated data. They try to extract features from the data that contai n useful information needed to improve their performance on target applications. Most works focus on directly optimizing the target loss functions to improve th e accuracy by allowing the model to implicitly learn representations from the da ta. There has not been much work on using background/noise data to estimate the statistics of in-domain data to improve the feature representation of deep neura l networks. In this paper, we probe this direction by deriving a relationship be tween the estimation of unknown parameters of the probability density function (pdf) of input data and classification accuracy. Using this relationship, we show that having a better estimate of the unknown parameters using background and in -domain data provides better features which leads to better accuracy. Based on t his result, we introduce a simple but effective detection booster training (DBT) method that applies a detection loss function on the early layers of a neural n etwork to discriminate in-domain data points from noise/background data, to impr ove the classifier accuracy. The background/noise data comes from the same famil y of pdfs of input data but with different parameter sets (e.g., mean, variance) . In addition, we also show that our proposed DBT method improves the accuracy e ven with limited labeled in-domain training samples as compared to normal traini ng. We conduct experiments on face recognition, image classification, and speake r classification problems and show that our method achieves superior performance over strong baselines across various datasets and model architectures.

****
Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent
El Mahdi El Mhamdi,Rachid Guerraoui,Sébastien Rouault
Byzantine-resilient Stochastic Gradient Descent (SGD) aims at shielding model training from Byzantine faults, be they ill-labeled training datapoints, exploited software/hardware vulnerabilities, or malicious worker nodes in a distributed setting.
Two recent attacks have been challenging state-of-the-art defenses though, often successfully precluding the model from even fitting the training set.
The main identified weakness in current defenses is their requirement of a sufficiently low variance-norm ratio for the stochastic gradients.
We propose a practical method which, despite increasing the variance, reduces the variance-norm ratio, mitigating the identified weakness.
We assess the effectiveness of our method over 736 different training configurations, comprising the 2 state-of-the-art attacks and 6 defenses.
For confidence and reproducibility purposes, each configuration is run 5 times with specified seeds (1 to 5), totalling 3680 runs.
In our experiments, when the attack is effective enough to decrease the highest observed top-1 cross-accuracy by at least 20% compared to the unattacked run, our technique systematically increases back the highest observed accuracy, and is able to recover at least 20% in more than 60% of the cases.
****
Can Kernel Transfer Operators Help Flow based Generative Models?
Zhichun Huang,Rudrasis Chakraborty,Xingjian Zhen,Vikas Singh
Flow-based generative models refer to deep generative models with
tractable likelihoods, and offer several attractive properties including
efficient density estimation and sampling. Despite many advantages,
current formulations (e.g., normalizing flow) often have an expensive memory/runtime footprint, which hinders their use in a number of applications.
In this paper, we consider the setting where we have access to an autoencoder, which is
suitably effective for the dataset of interest. Under some mild conditions,
we show that we can calculate a mapping to a RKHS which subsequently enables deploying
mature ideas from the kernel methods literature for flow-based generative models. Specifically, we can explicitly map the RKHS distribution (i.e.,
approximate the flow) to match or align with
a template/well-characterized distribution, via kernel transfer operators. This leads to a direct and resource efficient approximation avoiding iterative optimization. We empirically show that this simple idea yields competitive results on popular datasets such as CelebA,
as well as promising results on a public 3D brain imaging dataset where the sample sizes are much smaller.
****
Stochastic Canonical Correlation Analysis: A Riemannian Approach
Zihang Meng,Rudrasis Chakraborty,Vikas Singh
We present an efficient stochastic algorithm (RSG+) for canonical correlation analysis (CCA) derived via a differential geometric perspective of the underlying optimization task. We show that exploiting the Riemannian structure of the problem reveals natural strategies for modified forms of manifold stochastic gradient descent schemes that have been variously used in the literature for numerical optimization on manifolds. Our developments complement existing methods for this problem which either require $O(d^3)$ time complexity per iteration with $O(\frac{1}{\sqrt{t}})$ convergence rate (where d is the dimensionality) or only extract the top 1 component with $O(\frac{1}{t})$ convergence rate. In contrast, our algorithm achieves $O(d^2k)$ runtime complexity per iteration for extracting top k canonical components with $O(\frac{1}{t})$ convergence rate. We present our theoretical analysis as well as experiments describing the empirical behavior of our algorithm, including a potential application of this idea for training fair models where the label of protected attribute is missing or otherwise unav

ailable.
****
Secure Network Release with Link Privacy
Carl Yang,Haonan Wang,Ke ZHANG,Lichao Sun
Many data mining and analytical tasks rely on the abstraction of networks (graph s) to summarize relational structures among individuals (nodes). Since relationa l data are often sensitive, we aim to seek effective approaches to release utili ty-preserved yet privacy-protected structured data. In this paper, we leverage t he differential privacy (DP) framework, to formulate and enforce rigorous privac y constraints on deep graph generation models, with a focus on edge-DP to guaran tee individual link privacy. In particular, we enforce edge-DP by injecting Gaus sian noise to the gradients of a link reconstruction based graph generation mode l, and ensure data utility by improving structure learning with structure-orient ed graph comparison. Extensive experiments on two real-world network datasets sh ow that our proposed DPGGAN model is able to generate networks with effectively preserved global structure and rigorously protected individual link privacy.
****
Scaling the Convex Barrier with Active Sets
Alessandro De Palma,Harkirat Behl,Rudy R Bunel,Philip Torr,M. Pawan Kumar
Tight and efficient neural network bounding is of critical importance for the sc aling of neural network verification systems. A number of efficient specialised dual solvers for neural network bounds have been presented recently, but they ar e often too loose to verify more challenging properties. This lack of tightness is linked to the weakness of the employed relaxation, which is usually a linear program of size linear in the number of neurons. While a tighter linear relaxati on for piecewise linear activations exists, it comes at the cost of exponentiall y many constraints and thus currently lacks an efficient customised solver. We a lleviate this deficiency via a novel dual algorithm that realises the full poten tial of the new relaxation by operating on a small active set of dual variables. Our method recovers the strengths of the new relaxation in the dual space: tigh tness and a linear separation oracle. At the same time, it shares the benefits o f previous dual approaches for weaker relaxations: massive parallelism, GPU impl ementation, low cost per iteration and valid bounds at any time. As a consequenc e, we obtain better bounds than off-the-shelf solvers in only a fraction of thei r running time and recover the speed-accuracy trade-offs of looser dual solvers if the computational budget is small. We demonstrate that this results in signif icant formal verification speed-ups.
****
Model Compression via Hyper-Structure Network
Shangqian Gao,Feihu Huang,Heng Huang
In this paper, we propose a novel channel pruning method to solve the problem of compression and acceleration of Convolutional Neural Networks (CNNs). Previous channel pruning methods usually ignore the relationships between channels and la yers. Many of them parameterize each channel independently by using gates or sim ilar concepts. To fill this gap, a hyper-structure network is proposed to genera te the architecture of the main network. Like the existing hypernet, our hyper-s tructure network can be optimized by regular backpropagation. Moreover, we use a regularization term to specify the computational resource of the compact networ k. Usually, FLOPs is used as the criterion of computational resource. However, i f FLOPs is used in the regularization, it may over penalize early layers. To add ress this issue, we further introduce learnable layer-wise scaling factors to ba lance the gradients from different terms, and they can be optimized by hyper-gra dient descent. Extensive experimental results on CIFAR-10 and ImageNet show that our method is competitive with state-of-the-art methods.
****
REPAINT: Knowledge Transfer in Deep Actor-Critic Reinforcement Learning
Yunzhe Tao,Sahika Genc,TAO SUN,Sunil Mallya
Accelerating the learning processes for complex tasks by leveraging previously l earned tasks has been one of the most challenging problems in reinforcement lear ning, especially when the similarity between source and target tasks is low or u

nknown. In this work, we propose a REPresentation-And-INstance Transfer algorith m (REPAINT) for deep actor-critic reinforcement learning paradigm. In representa tion transfer, we adopt a kickstarted training method using a pre-trained teache r policy by introducing an auxiliary cross-entropy loss. In instance transfer, w e develop a sampling approach, i.e., advantage-based experience replay, on trans itions collected following the teacher policy, where only the samples with high advantage estimates are retained for policy update. We consider both learning an unseen target task by transferring from previously learned teacher tasks and le arning a partially unseen task composed of multiple sub-tasks by transferring fr om a pre-learned teacher sub-task. In several benchmark experiments, REPAINT sig nificantly reduces the total training time and improves the asymptotic performan ce compared to training with no prior knowledge and other baselines.

****

Wasserstein Distributional Normalization : Nonparametric Stochastic Modeling for Handling Noisy Labels

Sung Woo Park,Junseok Kwon

We propose a novel Wasserstein distributional normalization (WDN) algorithm to h andle noisy labels for accurate classification. In this paper, we split our dat a into uncertain and certain samples based on small loss criteria. We investigat e the geometric relationship between these two different types of samples and en hance this relation to exploit useful information, even from uncertain samples. To this end, we impose geometric constraints on the uncertain samples by normal izing them into the Wasserstein ball centered on certain samples. Experimental r esults demonstrate that our WDN outperforms other state-of-the-art methods on th e Clothing1M and CIFAR-10/100 datasets, which have diverse noisy labels. The pro posed WDN is highly compatible with existing classification methods, meaning it can be easily plugged into various methods to improve their accuracy significant ly.

****

Grounded Compositional Generalization with Environment Interactions

Yuanpeng Li

In this paper, we present a compositional generalization approach in grounded ag ent instruction learning. Compositional generalization is an important part of h uman intelligence, but current neural network models do not have such ability. T his is more complicated in multi-modal problems with grounding. Our proposed app roach has two main ideas. First, we use interactions between agent and the envir onment to find components in the output. Second, we apply entropy regularization to learn corresponding input components for each output component. The results show the proposed approach significantly outperforms baselines in most tasks, wi th more than 25% absolute average accuracy increase. We also investigate the imp act of entropy regularization and other changes with ablation study. We hope thi s work is the first step to address grounded compositional generalization, and i t will be helpful in advancing artificial intelligence research.

****

Distantly Supervised Relation Extraction in Federated Settings

Dianbo Sui,Yubo Chen,Kang Liu,Jun Zhao

Distant supervision is widely used in relation extraction in order to create a l arge-scale training dataset by aligning a knowledge base with unstructured text. Most existing studies in this field have assumed there is a great deal of centr alized unstructured text. However, in practice, text may be distributed on diffe rent platforms and cannot be centralized due to privacy restrictions. Therefore, it is worthwhile to investigate distant supervision in the federated learning p aradigm, which decouples the training of the model from the need for direct acce ss to the raw text. However, overcoming label noise of distant supervision becom es more difficult in federated settings, because the sentences containing the sa me entity pair scatter around different platforms. In this paper, we propose a f ederated denoising framework to suppress label noise in federated settings. The core of this framework is a multiple instance learning based denoising method th at is able to select reliable sentences via cross-platform collaboration. Variou

s experimental results on New York Times dataset and miRNA gene regulation relation dataset demonstrate the effectiveness of the proposed method.
****
Patch-level Neighborhood Interpolation: A General and Effective Graph-based Regularization Strategy
Ke Sun,Bing Yu,Zhouchen Lin,Zhanxing Zhu
Regularization plays a crucial role in machine learning models, especially for deep neural networks. The existing regularization techniques mainly rely on the i.i.d. assumption and only consider the knowledge from the current sample, without the leverage of the neighboring relationship between samples. In this work, we propose a general regularizer called Patch-level Neighborhood Interpolation~(\textbf{Pani}) that conducts a non-local representation in the computation of network. Our proposal explicitly constructs patch-level graphs in different network layers and then linearly interpolates neighborhood patch features, serving as a general and effective regularization strategy. Further, we customize our approach into two kinds of popular regularization methods, namely Virtual Adversarial Training (VAT) and MixUp as well as its variants. The first derived \textbf{Pani VAT} presents a novel way to construct non-local adversarial smoothness by employing patch-level interpolated perturbations. In addition, the second derived \textbf{Pani MixUp} method extends the original MixUp regularization and its variant to the Pani version, achieving a significant improvement in the performance. Finally, extensive experiments are conducted to verify the effectiveness of our Patch-level Neighborhood Interpolation approach in both supervised and semi-supervised settings.
****
BSQ: Exploring Bit-Level Sparsity for Mixed-Precision Neural Network Quantization
Huanrui Yang,Lin Duan,Yiran Chen,Hai Li
Mixed-precision quantization can potentially achieve the optimal tradeoff between performance and compression rate of deep neural networks, and thus, have been widely investigated. However, it lacks a systematic method to determine the exact quantization scheme. Previous methods either examine only a small manually-designed search space or utilize a cumbersome neural architecture search to explore the vast search space. These approaches cannot lead to an optimal quantization scheme efficiently. This work proposes bit-level sparsity quantization (BSQ) to tackle the mixed-precision quantization from a new angle of inducing bit-level sparsity. We consider each bit of quantized weights as an independent trainable variable and introduce a differentiable bit-sparsity regularizer. BSQ can induce all-zero bits across a group of weight elements and realize the dynamic precision reduction, leading to a mixed-precision quantization scheme of the original model. Our method enables the exploration of the full mixed-precision space with a single gradient-based optimization process, with only one hyperparameter to tradeoff the performance and compression. BSQ achieves both higher accuracy and higher bit reduction on various model architectures on the CIFAR-10 and ImageNet datasets comparing to previous methods.
****
Meta-Active Learning in Probabilistically-Safe Optimization
Mariah L Schrum,Mark Connolly,Eric Cole,Mihir Ghetiya,Robert Gross,Matthew C. Gombolay
Learning to control a safety-critical system with latent dynamics (e.g. for deep brain stimulation) requires judiciously taking calculated risks to gain information. We present a probabilistically-safe, meta-active learning approach to efficiently learn system dynamics and optimal configurations. The key to our approach is a novel integration of meta-learning and chance-constrained optimization in which we 1) meta-learn an LSTM-based embedding of the active learning sample history, 2) encode a deep learning-based acquisition function with this embedding into a mixed-integer linear program (MILP), and 3) solve the MILP to find the optimal action trajectory, trading off the predicted information gain from the acquisition function and the likelihood of safe control. We set a new state-of-the-art in active learning to control a high-dimensional system with latent dynamic

s, achieving a 46% increase in information gain and a 20% speedup in computation time. We then outperform baseline methods in learning the optimal parameter settings for deep brain stimulation in rats to enhance the rats' performance on a cognitive task while safely avoiding unwanted side effects (i.e., triggering seizures).

****

Group-Connected Multilayer Perceptron Networks

Mohammad Kachuee,Sajad Darabi,Shayan Fazeli,Majid Sarrafzadeh

Despite the success of deep learning in domains such as image, voice, and graphs, there has been little progress in deep representation learning for domains without a known structure between features. For instance, a tabular dataset of different demographic and clinical factors where the feature interactions are not given as a prior. In this paper, we propose Group-Connected Multilayer Perceptron (GMLP) networks to enable deep representation learning in these domains. GMLP is based on the idea of learning expressive feature combinations (groups) and exploiting them to reduce the network complexity by defining local group-wise operations. During the training phase, GMLP learns a sparse feature grouping matrix using temperature annealing softmax with an added entropy loss term to encourage the sparsity. Furthermore, an architecture is suggested which resembles binary trees, where group-wise operations are followed by pooling operations to combine information; reducing the number of groups as the network grows in depth. To evaluate the proposed method, we conducted experiments on different real-world datasets covering various application areas. Additionally, we provide visualizations on MNIST and synthesized data. According to the results, GMLP is able to successfully learn and exploit expressive feature combinations and achieve state-of-the-art classification performance on different datasets.

****

Continuous Transfer Learning

Jun Wu,Jingrui He

Transfer learning has been successfully applied across many high-impact applications. However, most existing work focuses on the static transfer learning setting, and very little is devoted to modeling the time evolving target domain, such as the online reviews for movies. To bridge this gap, in this paper, we focus on the continuous transfer learning setting with a time evolving target domain. One major challenge associated with continuous transfer learning is the time evolving relatedness of the source domain and the current target domain as the target domain evolves over time. To address this challenge, we first derive a generic generalization error bound on the current target domain with flexible domain discrepancy measures. Furthermore, a novel label-informed C-divergence is proposed to measure the shift of joint data distributions (over input features and output labels) across domains. It could be utilized to instantiate a tighter error upper bound in the continuous transfer learning setting, thus motivating us to develop an adversarial Variational Auto-encoder algorithm named CONTE by minimizing the C-divergence based error upper bound. Extensive experiments on various data sets demonstrate the effectiveness of our CONTE algorithm.

****

Adversarially Robust Federated Learning for Neural Networks

Yao Zhou,Jun Wu,Jingrui He

In federated learning, data is distributed among local clients which collaboratively train a prediction model using secure aggregation. To preserve the privacy of the clients, the federated learning paradigm requires each client to maintain a private local training data set, and only uploads its summarized model updates to the server. In this work, we show that this paradigm could lead to a vulnerable model, which collapses in performance when the corrupted data samples (under adversarial manipulations) are used for prediction after model deployment. To improve model robustness, we first decompose the aggregation error of the central server into bias and variance, and then, propose a robust federated learning framework, named Fed_BVA, that performs on-device adversarial training using the bias-variance oriented adversarial examples supplied by the server via asymmetri

cal communications. The experiments are conducted on multiple benchmark data sets using several prevalent neural network models, and the empirical results show that our framework is robust against white-box and black-box adversarial corruptions under both IID and non-IID settings.
****
Rethinking Content and Style: Exploring Bias for Unsupervised Disentanglement
Xuanchi Ren,Tao Yang,Wenjun Zeng,Yuwang Wang
Content and style (C-S) disentanglement intends to decompose the underlying explanatory factors of objects into two independent latent spaces. Aiming for unsupervised disentanglement, we introduce an inductive bias to our formulation by assigning different and independent roles to content and style when approximating the real data distributions. The content embeddings of individual images are forced to share a common distribution. The style embeddings encoding instance-specific features are used to customize the shared distribution. The experiments on several popular datasets demonstrate that our method achieves the state-of-the-art disentanglement compared to other unsupervised approaches and comparable or even better results than supervised methods. Furthermore, as a new application of C-S disentanglement, we propose to generate multi-view images from a single view image for 3D reconstruction.
****
Dual-mode ASR: Unify and Improve Streaming ASR with Full-context Modeling
Jiahui Yu,Wei Han,Anmol Gulati,Chung-Cheng Chiu,Bo Li,Tara N Sainath,Yonghui Wu,Ruoming Pang
Streaming automatic speech recognition (ASR) aims to emit each hypothesized word as quickly and accurately as possible, while full-context ASR waits for the completion of a full speech utterance before emitting completed hypotheses. In this work, we propose a unified framework, Dual-mode ASR, to train a single end-to-end ASR model with shared weights for both streaming and full-context speech recognition. We show that the latency and accuracy of streaming ASR significantly benefit from weight sharing and joint training of full-context ASR, especially with inplace knowledge distillation during the training. The Dual-mode ASR framework can be applied to recent state-of-the-art convolution-based and transformer-based ASR networks. We present extensive experiments with two state-of-the-art ASR networks, ContextNet and Conformer, on two datasets, a widely used public dataset LibriSpeech and a large-scale dataset MultiDomain. Experiments and ablation studies demonstrate that Dual-mode ASR not only simplifies the workflow of training and deploying streaming and full-context ASR models, but also significantly improves both emission latency and recognition accuracy of streaming ASR. With Dual-mode ASR, we achieve new state-of-the-art streaming ASR results on both LibriSpeech and MultiDomain in terms of accuracy and latency.
****
Policy-Driven Attack: Learning to Query for Hard-label Black-box Adversarial Examples
Ziang Yan,Yiwen Guo,Jian Liang,Changshui Zhang
To craft black-box adversarial examples, adversaries need to query the victim model and take proper advantage of its feedback. Existing black-box attacks generally suffer from high query complexity, especially when only the top-1 decision (i.e., the hard-label prediction) of the victim model is available. In this paper, we propose a novel hard-label black-box attack named Policy-Driven Attack, to reduce the query complexity. Our core idea is to learn promising search directions of the adversarial examples using a well-designed policy network in a novel reinforcement learning formulation, in which the queries become more sensible. Experimental results demonstrate that our method can significantly reduce the query complexity in comparison with existing state-of-the-art hard-label black-box attacks on various image classification benchmark datasets. Code and models for reproducing our results are available at https://github.com/ZiangYan/pda.pytorch
****
Hierarchical Meta Reinforcement Learning for Multi-Task Environments
Dongyang Zhao,Yue Huang,Changnan Xiao,Yue Li,Shihong Deng
Deep reinforcement learning algorithms aim to achieve human-level intelligence b

y solving practical decisions-making problems, which are often composed of multi ple sub-tasks. Complex and subtle relationships between sub-tasks make tradition al methods hard to give a promising solution. We implement a first-person shooti ng environment with random spatial structures to illustrate a typical representa tive of this kind. A desirable agent should be capable of balancing between diff erent sub-tasks: navigation to find enemies and shooting to kill them. To addres s the problem brought by the environment, we propose a Meta Soft Hierarchical re inforcement learning framework (MeSH), in which each low-level sub-policy focuse s on a specific sub-task respectively and high-level policy automatically learns to utilize low-level sub-policies through meta-gradients. The proposed framewor k is able to disentangle multiple sub-tasks and discover proper low-level polici es under different situations. The effectiveness and efficiency of the framework are shown by a series of comparison experiments. Both environment and algorithm code will be provided for open source to encourage further research.
****

BASGD: Buffered Asynchronous SGD for Byzantine Learning
Yi-Rui Yang,Wu-Jun Li
Distributed learning has become a hot research topic due to its wide application in cluster-based large-scale learning, federated learning, edge computing and s o on. Most traditional distributed learning methods typically assume no failure or attack on workers. However, many unexpected cases, such as communication fail ure and even malicious attack, may happen in real applications. Hence, Byzantine learning (BL), which refers to distributed learning with failure or attack, has recently attracted much attention. Most existing BL methods are synchronous, wh ich are impractical in some applications due to heterogeneous or offline workers . In these cases, asynchronous BL (ABL) is usually preferred. In this paper, we propose a novel method, called buffered asynchronous stochastic gradient descent (BASGD), for ABL. To the best of our knowledge, BASGD is the first ABL method t hat can resist malicious attack without storing any instances on server. Compare d with those methods which need to store instances on server, BASGD takes less r isk of privacy leakage. BASGD is proved to be convergent, and be able to resist failure or attack. Empirical results show that BASGD significantly outperforms v anilla ASGD and other ABL baselines when there exists failure or attack on worke rs.
****

DEEP ADAPTIVE SEMANTIC LOGIC (DASL): COMPILING DECLARATIVE KNOWLEDGE INTO DEEP N EURAL NETWORKS
Karan Sikka,Andrew Silberfarb,John Byrnes,Indranil Sur,Ed Chow,Ajay Divakaran,Ri chard Rohwer
We introduce Deep Adaptive Semantic Logic (DASL), a novel framework for automati ng the generation of deep neural networks that incorporates user-provided formal knowledge to improve learning from data. We provide formal semantics that demon strate that our knowledge representation captures all of first order logic and t hat finite sampling from infinite domains converges to correct truth values. DAS L’s representation improves on prior neuro-symbolic work by avoiding vanishing g radients, allowing deeper logical structure, and enabling richer interactions be tween the knowledge and learning components. We illustrate DASL through a toy pr oblem in which we add structure to an image classification problem and demonstra te that knowledge of that structure
reduces data requirements by a factor of 1000. We apply DASL on a visual relatio nship detection task and demonstrate that the addition of commonsense knowledge improves performance by 10.7% in a data scarce setting.
****

Sequential Density Ratio Estimation for Simultaneous Optimization of Speed and A ccuracy
Akinori F Ebihara,Taiki Miyagawa,Kazuyuki Sakurai,Hitoshi Imaoka
Classifying sequential data as early and as accurately as possible is a challeng ing yet critical problem, especially when a sampling cost is high. One algorithm that achieves this goal is the sequential probability ratio test (SPRT), which is known as Bayes-optimal: it can keep the expected number of data samples as sm

all as possible, given the desired error upper-bound. However, the original SPRT makes two critical assumptions that limit its application in real-world scenarios: (i) samples are independently and identically distributed, and (ii) the likelihood of the data being derived from each class can be calculated precisely. Here, we propose the SPRT-TANDEM, a deep neural network-based SPRT algorithm that overcomes the above two obstacles. The SPRT-TANDEM sequentially estimates the log-likelihood ratio of two alternative hypotheses by leveraging a novel Loss function for Log-Likelihood Ratio estimation (LLLR) while allowing correlations up to $N (\in \mathbb{N})$ preceding samples. In tests on one original and two public video databases, Nosaic MNIST, UCF101, and SiW, the SPRT-TANDEM achieves statistically significantly better classification accuracy than other baseline classifiers, with a smaller number of data samples. The code and Nosaic MNIST are publicly available at https://github.com/TaikiMiyagawa/SPRT-TANDEM.
****

Uncertainty Sets for Image Classifiers using Conformal Prediction
Anastasios Nikolas Angelopoulos,Stephen Bates,Michael Jordan,Jitendra Malik
Convolutional image classifiers can achieve high predictive accuracy, but quantifying their uncertainty remains an unresolved challenge, hindering their deployment in consequential settings. Existing uncertainty quantification techniques, such as Platt scaling, attempt to calibrate the network's probability estimates, but they do not have formal guarantees. We present an algorithm that modifies any classifier to output a predictive set containing the true label with a user-specified probability, such as 90%. The algorithm is simple and fast like Platt scaling, but provides a formal finite-sample coverage guarantee for every model and dataset. Our method modifies an existing conformal prediction algorithm to give more stable predictive sets by regularizing the small scores of unlikely classes after Platt scaling. In experiments on both Imagenet and Imagenet-V2 with ResNet-152 and other classifiers, our scheme outperforms existing approaches, achieving coverage with sets that are often factors of 5 to 10 smaller than a stand-alone Platt scaling baseline.
****

L2E: Learning to Exploit Your Opponent
Zhe Wu,Kai Li,Hang Xu,Meng Zhang,Haobo Fu,Bo An,Junliang Xing
Opponent modeling is essential to exploit sub-optimal opponents in strategic interactions. One key challenge facing opponent modeling is how to fast adapt to opponents with diverse styles of strategies. Most previous works focus on building explicit models to predict the opponents' styles or strategies directly. However, these methods require a large amount of data to train the model and lack the adaptability to new opponents of unknown styles. In this work, we propose a novel Learning to Exploit (L2E) framework for implicit opponent modeling. L2E acquires the ability to exploit opponents by a few interactions with different opponents during training so that it can adapt to new opponents with unknown styles during testing quickly. We propose a novel Opponent Strategy Generation (OSG) algorithm that produces effective opponents for training automatically. By learning to exploit the challenging opponents generated by OSG through adversarial training, L2E gradually eliminates its own strategy's weaknesses. Moreover, the generalization ability of L2E is significantly improved by training with diverse opponents, which are produced by OSG through diversity-regularized policy optimization. We evaluate the L2E framework on two poker games and one grid soccer game, which are the commonly used benchmark for opponent modeling. Comprehensive experimental results indicate that L2E quickly adapts to diverse styles of unknown opponents.
****

Shape Defense
ali borji
Humans rely heavily on shape information to recognize objects. Conversely, convolutional
neural networks (CNNs) are biased more towards texture. This fact
is perhaps the main reason why CNNs are susceptible to adversarial examples.
Here, we explore how shape bias can be incorporated into CNNs to improve their

robustness. Two algorithms are proposed, based on the observation that edges are invariant to moderate imperceptible perturbations. In the first one, a classifier is
adversarially trained on images with the edge map as an additional channel. At inference time, the edge map is recomputed and concatenated to the image. In the second algorithm, a conditional GAN is trained to translate the edge maps, from clean and/or perturbed images, into clean images. The inference is done over the generated image corresponding to the input’s edge map. A large number of experiments
with more than 10 data sets have proved the effectiveness of the proposed algorithms against FGSM and `1 PGD-40 attacks. against FGSM and `ℓ_∞ PGD-40 attacks.
Further, we show that edge information can a) benefit other adversarial training methods, b) be even more effective
in conjunction with background subtraction, c) be used to defend against poisoning
attacks, and d) make CNNs more robust against natural image corruptions
such as motion blur, impulse noise, and JPEG compression, than CNNs trained solely on RGB images. From a broader perspective, our study suggests that CNNs do not adequately account for image structures and operations that are crucial for
robustness. The code is available at: https://github.com/[masked].
****
Modal Uncertainty Estimation via Discrete Latent Representations
Di Qiu,Zhanghan Ke,Peng Su,Lok Ming Lui
Many important problems in the real world don't have unique solutions. It is thus important for machine learning models to be capable of proposing different plausible solutions with meaningful probability measures.
In this work we propose a novel deep learning based framework, named {\it modal uncertainty estimation} (MUE), to learn the one-to-many mappings between the inputs and outputs, together with faithful uncertainty estimation.
Motivated by the multi-modal posterior collapse problem in current conditional generative models, MUE uses a set of discrete latent variables, each representing a latent mode hypothesis that explains one type of input-output relationship, to generate the one-to-many mappings. Benefit from the discrete nature of the latent representations, MUE can estimate any input the conditional probability distribution of the outputs effectively. Moreover, MUE is efficient during training since the discrete latent space and its uncertainty estimation are jointly learned.
We also develop the theoretical background of MUE and extensively validate it on both synthetic and realistic tasks. MUE demonstrates (1) significantly more accurate uncertainty estimation than the current state-of-the-art, and (2) its informativeness for practical use.

****
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 noncovex stochastic optimization, which dynamically incorporates the curvature of the loss function by approximating the Hessian via a diagonal matrix. Algorithmically, Apollo requires only first-order gradients and updates the approximation of the Hessian diagonally such that it satisfies the weak secant relation. To handle nonconvexity, we replace the Hessian with its absolute value, the computation of which is also efficient under our diagonal approximation, yielding an optimization algorithm with linear complexity for both time and memory. Experimentally, through three tasks on vision and language we 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.

****
Improving Tail Label Prediction for Extreme Multi-label Learning
Tong Wei,Wei-Wei Tu,Yu-Feng Li
Extreme multi-label learning (XML) works to annotate objects with relevant labels from an extremely large label set. Many previous methods treat labels uniformly such that the learned model tends to perform better on head labels, while the performance is severely deteriorated for tail labels. However, it is often desirable to predict more tail labels in many real-world applications. To alleviate this problem, in this work, we show theoretical and experimental evidence for the inferior performance of representative XML methods on tail labels. Our finding is that the norm of label classifier weights typically follows a long-tailed distribution similar to the label frequency, which results in the over-suppression of tail labels. Base on this new finding, we present two new modules: (1)~\algoa~learns to re-rank the predictions by optimizing a population-aware loss, which predicts tail labels with high rank; (2)~\algob~augments tail labels via a decoupled learning scheme, which can yield more balanced classification boundary. We conduct experiments on commonly used XML benchmarks with hundreds of thousands of labels, showing that the proposed methods improve the performance of many state-of-the-art XML models by a considerable margin (6\% performance gain with respect to PSP@1 on average).
****
GN-Transformer: Fusing AST and Source Code information in Graph Networks
Junyan Cheng,Iordanis Fostiropoulos,Barry Boehm
As opposed to natural languages, source code understanding is influenced by grammar relations between tokens regardless of their identifier name. Considering graph representation of source code such as Abstract Syntax Tree (AST) and Control Flow Graph (CFG), can capture a token’s grammatical relationships that are not obvious from the source code. Most existing methods are late fusion and underperform when supplementing the source code text with a graph representation. We propose a novel method called GN-Transformer to fuse representations learned from graph and text modalities under the Graph Networks (GN) framework with attention mechanism. Our method learns the embedding on a constructed graph called Syntax-Code Graph (SCG). We perform experiments on the structure of SCG, an ablation study on the model design and the hyper-paramaters to conclude that the performance advantage is from the fusion method and not the specific details of the model. The proposed method achieved state of the art performance in two code summarization datasets and across three metrics.
****
Learn2Weight: Weights Transfer Defense against Similar-domain Adversarial Attacks
Siddhartha Datta
Recent work in black-box adversarial attacks for NLP systems has attracted attention. Prior black-box attacks assume that attackers can observe output labels from target models based on selected inputs. In this work, inspired by adversarial transferability, we propose a new type of black-box NLP adversarial attack that an attacker can choose a similar domain and transfer the adversarial examples to the target domain and cause poor performance in target model. Based on domain adaptation theory, we then propose a defensive strategy, called Learn2Weight, which trains to predict the weight adjustments for target model in order to defense the attack of similar-domain adversarial examples. Using Amazon multi-domain sentiment classification dataset, we empirically show that Learn2Weight model is effective against the attack compared to standard black-box defense methods such as adversarial training and defense distillation. This work contributes to the growing literature on machine learning safety.
****
Graph Convolution with Low-rank Learnable Local Filters
Xiuyuan Cheng,Zichen Miao,Qiang Qiu
Geometric variations like rotation, scaling, and viewpoint changes pose a significant challenge to visual understanding. One common solution is to directly model certain intrinsic structures, e.g., using landmarks. However, it then becomes

non-trivial to build effective deep models, especially when the underlying non-Euclidean grid is irregular and coarse. Recent deep models using graph convolutions provide an appropriate framework to handle such non-Euclidean data, but many of them, particularly those based on global graph Laplacians, lack expressiveness to capture local features required for representation of signals lying on the non-Euclidean grid. The current paper introduces a new type of graph convolution with learnable low-rank local filters, which is provably more expressive than previous spectral graph convolution methods. The model also provides a unified framework for both spectral and spatial graph convolutions. To improve model robustness, regularization by local graph Laplacians is introduced. The representation stability against input graph data perturbation is theoretically proved, making use of the graph filter locality and the local graph regularization. Experiments on spherical mesh data, real-world facial expression recognition/skeleton-based action recognition data, and data with simulated graph noise show the empirical advantage of the proposed model.

****

Necessary and Sufficient Conditions for Compositional Representations

Yuanpeng Li

Humans leverage compositionality for flexible and efficient learning, but current machine learning algorithms lack such ability. Despite many efforts in specific cases, there is still absence of theories and tools to study it systematically. In this paper, we leverage group theory to mathematically prove necessary and sufficient conditions for two fundamental questions of compositional representations. (1) What are the properties for a set of components to be expressed compositionally. (2) What are the properties for mappings between compositional and entangled representations. We provide examples to better understand the conditions and how to apply them. E.g., we use the theory to give a new explanation of why attention mechanism helps compositionality. We hope this work will help to advance understanding of compositionality and improvement of artificial intelligence towards human level.

****

Stochastic Inverse Reinforcement Learning

Ce Ju

The goal of the inverse reinforcement learning (IRL) problem is to recover the reward functions from expert demonstrations. However, the IRL problem like any ill-posed inverse problem suffers the congenital defect that the policy may be optimal for many reward functions, and expert demonstrations may be optimal for many policies. In this work, we generalize the IRL problem to a well-posed expectation optimization problem stochastic inverse reinforcement learning (SIRL) to recover the probability distribution over reward functions. We adopt the Monte Carlo expectation-maximization (MCEM) method to estimate the parameter of the probability distribution as the first solution to the SIRL problem. The solution is succinct, robust, and transferable for a learning task and can generate alternative solutions to the IRL problem. Through our formulation, it is possible to observe the intrinsic property for the IRL problem from a global viewpoint, and our approach achieves a considerable performance on the objectworld.

****

PSTNet: Point Spatio-Temporal Convolution on Point Cloud Sequences

Hehe Fan,Xin Yu,Yuhang Ding,Yi Yang,Mohan Kankanhalli

Point cloud sequences are irregular and unordered in the spatial dimension while exhibiting regularities and order in the temporal dimension. Therefore, existing grid based convolutions for conventional video processing cannot be directly applied to spatio-temporal modeling of raw point cloud sequences. In this paper, we propose a point spatio-temporal (PST) convolution to achieve informative representations of point cloud sequences. The proposed PST convolution first disentangles space and time in point cloud sequences. Then, a spatial convolution is employed to capture the local structure of points in the 3D space, and a temporal convolution is used to model the dynamics of the spatial regions along the time dimension. Furthermore, we incorporate the proposed PST convolution into a dee

p network, namely PSTNet, to extract features of point cloud sequences in a hier archical manner. Extensive experiments on widely-used 3D action recognition and 4D semantic segmentation datasets demonstrate the effectiveness of PSTNet to mo del point cloud sequences.
****

Numeric Encoding Options with Automunge
Nicholas Teague
Mainstream practice in machine learning with tabular data may take for granted t hat any feature engineering beyond scaling for numeric sets is superfluous in co ntext of deep neural networks. This paper will offer arguments for potential ben efits of extended encodings of numeric streams in deep learning by way of a surv ey of options for numeric transformations as available in the Automunge open sou rce python library platform for tabular data pipelines, where transformations ma y be applied to distinct columns in “family tree” sets with generations and bran ches of derivations. Automunge transformation options include normalization, bin ning, noise injection, derivatives, and more. The aggregation of these methods i nto family tree sets of transformations are demonstrated for use to present nume ric features to machine learning in multiple configurations of varying informati on content, as may be applied to encode numeric sets of unknown interpretation. Experiments demonstrate the realization of a novel generalized solution to data augmentation by noise injection for tabular learning, as may materially benefit model performance in applications with underserved training data.
****

Adversarial Synthetic Datasets for Neural Program Synthesis
Alexander Suh,Yuval Timen
Program synthesis is the task of automatically generating a program consistent w ith a given specification. A natural way to specify programs is to provide examp les of desired input-output behavior, and many current program synthesis approac hes have achieved impressive results after training on randomly generated input- output examples. However, recent work has discovered that some of these approach es generalize poorly to data distributions different from that of the randomly g enerated examples. We show that this problem applies to other state-of-the-art a pproaches as well and that current methods to counteract this problem are insuff icient. We then propose a new, adversarial approach to control the bias of synth etic data distributions and show that it outperforms current approaches.
****

A Simple and General Graph Neural Network with Stochastic Message Passing
Ziwei Zhang,Chenhao Niu,Peng Cui,Bo Zhang,Wei Cui,Wenwu Zhu
Graph neural networks (GNNs) are emerging machine learning models on graphs. One key property behind the expressiveness of existing GNNs is that the learned nod e representations are permutation-equivariant. Though being a desirable property for certain tasks, however, permutation-equivariance prevents GNNs from being p roximity-aware, i.e., preserving the walk-based proximities between pairs of nod es, which is another critical property for graph analytical tasks. On the other hand, some variants of GNNs are proposed to preserve node proximities, but they fail to maintain permutation-equivariance. How to empower GNNs to be proximity a ware while maintaining permutation-equivariance remains an open problem. In this paper, we propose Stochastic Message Passing (SMP), a general and simple GNN to maintain both proximity-awareness and permutation-equivariance properties. Spec ifically, we augment the existing GNNs with stochastic node representations lear ned to preserve node proximities. Though seemingly simple, we prove that such a mechanism can enable GNNs to preserve node proximities in theory while maintaini ng permutation-equivariance with certain parametrization. Extensive experimental results demonstrate the effectiveness and efficiency of SMP for tasks including node classification and link prediction.
****

Network Pruning That Matters: A Case Study on Retraining Variants
Duong Hoang Le,Binh-Son Hua
Network pruning is an effective method to reduce the computational expense of ov er-parameterized neural networks for deployment on low-resource systems. Recent

state-of-the-art techniques for retraining pruned networks such as weight rewinding and learning rate rewinding have been shown to outperform the traditional fine-tuning technique in recovering the lost accuracy (Renda et al., 2020), but so far it is unclear what accounts for such performance. In this work, we conduct extensive experiments to verify and analyze the uncanny effectiveness of learning rate rewinding. We find that the reason behind the success of learning rate rewinding is the usage of a large learning rate. Similar phenomenon can be observed in other learning rate schedules that involve large learning rates, e.g., the 1-cycle learning rate schedule (Smith et al., 2019). By leveraging the right learning rate schedule in retraining, we demonstrate a counter-intuitive phenomenon in that randomly pruned networks could even achieve better performance than methodically pruned networks (fine-tuned with the conventional approach). Our results emphasize the cruciality of the learning rate schedule in pruned network retraining - a detail often overlooked by practitioners during the implementation of network pruning.

****

EEC: Learning to Encode and Regenerate Images for Continual Learning

Ali Ayub,Alan Wagner

The two main impediments to continual learning are catastrophic forgetting and memory limitations on the storage of data. To cope with these challenges, we propose a novel, cognitively-inspired approach which trains autoencoders with Neural Style Transfer to encode and store images. Reconstructed images from encoded episodes are replayed when training the classifier model on a new task to avoid catastrophic forgetting. The loss function for the reconstructed images is weighted to reduce its effect during classifier training to cope with image degradation. When the system runs out of memory the encoded episodes are converted into centroids and covariance matrices, which are used to generate pseudo-images during classifier training, keeping classifier performance stable with less memory. Our approach increases classification accuracy by 13-17% over state-of-the-art methods on benchmark datasets, while requiring 78% less storage space.

****

Mind the Pad -- CNNs Can Develop Blind Spots

Bilal Alsallakh,Narine Kokhlikyan,Vivek Miglani,Jun Yuan,Orion Reblitz-Richardson

We show how feature maps in convolutional networks are susceptible to spatial bias. Due to a combination of architectural choices, the activation at certain locations is systematically elevated or weakened. The major source of this bias is the padding mechanism. Depending on several aspects of convolution arithmetic, this mechanism can apply the padding unevenly, leading to asymmetries in the learned weights. We demonstrate how such bias can be detrimental to certain tasks such as small object detection: the activation is suppressed if the stimulus lies in the impacted area, leading to blind spots and misdetection. We explore alternative padding methods and propose solutions for analyzing and mitigating spatial bias.

****

Revisiting Point Cloud Classification with a Simple and Effective Baseline

Ankit Goyal,Hei Law,Bowei Liu,Alejandro Newell,Jia Deng

Processing point cloud data is an important component of many real-world systems. As such, a wide variety of point-based approaches have been proposed, reporting steady benchmark improvements over time. We study the key ingredients of this progress and uncover two critical results. First, we find that auxiliary factors like different evaluation schemes, data augmentation strategies, and loss functions, which are independent of the model architecture, make a large difference in performance. The differences are large enough that they obscure the effect of architecture. When these factors are controlled for, PointNet++, a relatively older network, performs competitively with recent methods. Second, a very simple projection-based method, which we refer to as SimpleView, performs surprisingly well. It achieves on par or better results than sophisticated state-of-the-art methods on ModelNet40, while being half the size of PointNet++. It also outperform

s state-of-the-art methods on ScanObjectNN, a real-world point cloud benchmark, and demonstrates better cross-dataset generalization.

****
DARTS-: Robustly Stepping out of Performance Collapse Without Indicators
Xiangxiang Chu,Xiaoxing Wang,Bo Zhang,Shun Lu,Xiaolin Wei,Junchi Yan
Despite the fast development of differentiable architecture search (DARTS), it suffers from a standing instability issue regarding searching performance, which extremely limits its application. Existing robustifying methods draw clues from the outcome instead of finding out the causing factor. Various indicators such as Hessian eigenvalues are proposed as a signal of performance collapse, and the searching should be stopped once an indicator reaches a preset threshold. However, these methods tend to easily reject good architectures if thresholds are inappropriately set, let alone the searching is intrinsically noisy. In this paper, we undertake a more subtle and direct approach to resolve the collapse. We first demonstrate that skip connections with a learnable architectural coefficient can easily recover from a disadvantageous state and become dominant. We conjecture that skip connections profit too much from this privilege, hence causing the collapse for the derived model. Therefore, we propose to factor out this benefit with an auxiliary skip connection, ensuring a fairer competition for all operations. Extensive experiments on various datasets verify that our approach can substantially improve the robustness of DARTS. Our code is available at https://github.com/Meituan-AutoML/DARTS-
****
Transformers satisfy
Feng Shi,CHEN LI,Shijie Bian,Yiqiao Jin,Ziheng Xu,Tian Han,Song-Chun Zhu
The Propositional Satisfiability Problem (SAT), and more generally, the Constraint Satisfaction Problem (CSP), are mathematical questions defined as finding an assignment to a set of objects that satisfies a series of constraints. The modern approach is trending to solve CSP through neural symbolic methods. Most recent works are sequential model-based and adopt neural embedding, i.e., reinforcement learning with neural graph networks, and graph recurrent neural networks. This work proposes a one-shot model derived from the eminent Transformer architecture for factor graph structure to solve the CSP problem. We define the heterogeneous attention mechanism based on meta-paths for the self-attention between literals, the cross-attention based on the bipartite graph links from literal to clauses, or vice versa. This model takes advantage of parallelism. Our model achieves high speed and very high accuracy on the factor graph for CSPs with arbitrary size.
****
Meta-k: Towards Unsupervised Prediction of Number of Clusters
Azade Farshad,Samin Hamidi,Nassir Navab
Data clustering is a well-known unsupervised learning approach. Despite the recent advances in clustering using deep neural networks, determining the number of clusters without any information about the given dataset remains an existing problem. There have been classical approaches based on data statistics that require the manual analysis of a data scientist to calculate the probable number of clusters in a dataset. In this work, we propose a new method for unsupervised prediction of the number of clusters in a dataset given only the data without any labels. We evaluate our method extensively on randomly generated datasets using the scikit-learn package and multiple computer vision datasets and show that our method is able to determine the number of classes in a dataset effectively without any supervision.
****
Ranking Cost: One-Stage Circuit Routing by Directly Optimizing Global Objective Function
Shiyu Huang,Bin Wang,Dong Li,Jianye Hao,Jun Zhu,Ting Chen
Circuit routing has been a historically challenging problem in designing electronic systems such as very large-scale integration (VLSI) and printed circuit boards (PCBs). The main challenge is that connecting a large number of electronic co

mponents under specific design rules and constraints involves a very large searc
h space, which is proved to be NP-complete.
Early solutions are typically designed with hard-coded heuristics, which suffer
from problems of non-optimum solutions and lack of flexibility for new design ne
eds. Although a few learning-based methods have been proposed recently, their me
thods are cumbersome and hard to extend to large-scale applications. In this wor
k, we propose a new algorithm for circuit routing, named as Ranking Cost (RC),
which innovatively combines search-based methods (i.e., A* algorithm) and learni
ng-based methods (i.e., Evolution Strategies) to form an efficient and trainable
 router under a proper parameterization. Different from two-stage routing method
s (i.e., first global routing and then detailed routing), our method involves a
 one-stage procedure that directly optimizes the global objective function, thus
 it can be easy to adapt to new routing rules and constraints. In our method, we
 introduce a new set of variables called cost maps, which can help the A* router
 to find out proper paths to achieve the global object. We also train a ranking
parameter, which can produce the ranking order and further improve the performan
ce of our method. Our algorithm is trained in an end-to-end manner and does not
use any artificial data or human demonstration. In the experiments, we compare o
ur method with the sequential A* algorithm and a canonical reinforcement learnin
g approach, and results show that our method outperforms these baselines with hi
gher connectivity rates and better scalability. Our ablation study shows that ou
r trained cost maps can capture the global information and guide the routing res
ult to approach global optimum.
****

Stabilized Medical Image Attacks
Gege Qi,Lijun GONG,Yibing Song,Kai Ma,Yefeng Zheng
Convolutional Neural Networks (CNNs) have advanced existing medical systems for
automatic disease diagnosis. However, a threat to these systems arises that adve
rsarial attacks make CNNs vulnerable. Inaccurate diagnosis results make a negati
ve influence on human healthcare. There is a need to investigate potential adver
sarial attacks to robustify deep medical diagnosis systems. On the other side, t
here are several modalities of medical images (e.g., CT, fundus, and endoscopic
image) of which each type is significantly different from others. It is more cha
llenging to generate adversarial perturbations for different types of medical im
ages. In this paper, we propose an image-based medical adversarial attack method
 to consistently produce adversarial perturbations on medical images. The object
ive function of our method consists of a loss deviation term and a loss stabiliz
ation term. The loss deviation term increases the divergence between the CNN pre
diction of an adversarial example and its ground truth label. Meanwhile, the los
s stabilization term ensures similar CNN predictions of this example and its smo
othed input. From the perspective of the whole iterations for perturbation gener
ation, the proposed loss stabilization term exhaustively searches the perturbati
on space to smooth the single spot for local optimum escape. We further analyze
the KL-divergence of the proposed loss function and find that the loss stabiliza
tion term makes the perturbations updated towards a fixed objective spot while d
eviating from the ground truth. This stabilization ensures the proposed medical
attack effective for different types of medical images while producing perturbat
ions in small variance. Experiments on several medical image analysis benchmarks
 including the recent COVID-19 dataset show the stability of the proposed method
.
****

Forward Prediction for Physical Reasoning
Rohit Girdhar,Laura Gustafson,Aaron B. Adcock,Laurens van der Maaten
Physical reasoning requires forward prediction: the ability to forecast what wil
l happen next given some initial world state. We study the performance of state-
of-the-art forward-prediction models in the complex physical-reasoning tasks of
the PHYRE benchmark (Bakhtin et al., 2019). We do so by incorporating models tha
t operate on object or pixel-based representations of the world into simple phys
ical-reasoning agents. We find that forward-prediction models can improve physic
al-reasoning performance, particularly on complex tasks that involve many object

s. However, we also find that these improvements are contingent on the test task s being small variations of train tasks, and that generalization to completely n ew task templates is challenging. Surprisingly, we observe that forward predicto rs with better pixel accuracy do not necessarily lead to better physical-reasoni ng performance. Nevertheless, our best models set a new state-of-the-art on the PHYRE benchmark.
****
Proper Measure for Adversarial Robustness
Hyeongji Kim,Ketil Malde
This paper analyzes the problems of adversarial accuracy and adversarial trainin g. We argue that standard adversarial accuracy fails to properly measure the rob ustness of classifiers. Its definition has a tradeoff with standard accuracy eve n when we neglect generalization. In order to handle the problems of the standar d adversarial accuracy, we introduce a new measure for the robustness of classif iers called genuine adversarial accuracy. It can measure the adversarial robustn ess of classifiers without trading off accuracy on clean data and accuracy on th e adversarially perturbed samples. In addition, it does not favor a model with i nvariance-based adversarial examples, samples whose predicted classes are unchan ged even if the perceptual classes are changed. We prove that a single nearest n eighbor (1-NN) classifier is the most robust classifier according to genuine adv ersarial accuracy for given data and a norm-based distance metric when the class for each data point is unique. Based on this result, we suggest that using poor distance metrics might be one factor for the tradeoff between test accuracy and l_p norm-based test adversarial robustness.
****
THE EFFICACY OF L1 REGULARIZATION IN NEURAL NETWORKS
Gen Li,Yuantao Gu,Jie Ding
A crucial problem in neural networks is to select the most appropriate number of hidden neurons and obtain tight statistical risk bounds. In this work, we prese nt a new perspective towards the bias-variance tradeoff in neural networks. As a n alternative to selecting the number of neurons, we theoretically show that $L_ 1$ regularization can control the generalization error and sparsify the input di mension. In particular, with an appropriate L_1 regularization on the output l ayer, the network can produce a statistical risk that is near minimax optimal. M oreover, an appropriate L_1 regularization on the input layer leads to a risk bound that does not involve the input data dimension. Our analysis is based on a new amalgamation of dimension-based and norm-based complexity analysis to bound the generalization error. A consequent observation from our results is that an excessively large number of neurons do not necessarily inflate generalization er rors under a suitable regularization.

****
BiPointNet: Binary Neural Network for Point Clouds
Haotong Qin,Zhongang Cai,Mingyuan Zhang,Yifu Ding,Haiyu Zhao,Shuai Yi,Xianglong Liu,Hao Su
To alleviate the resource constraint for real-time point cloud applications that run on edge devices, in this paper we present BiPointNet, the first model binar ization approach for efficient deep learning on point clouds. We discover that t he immense performance drop of binarized models for point clouds mainly stems fr om two challenges: aggregation-induced feature homogenization that leads to a de gradation of information entropy, and scale distortion that hinders optimization and invalidates scale-sensitive structures. With theoretical justifications and in-depth analysis, our BiPointNet introduces Entropy-Maximizing Aggregation (EM A) to modulate the distribution before aggregation for the maximum information e ntropy, and Layer-wise Scale Recovery (LSR) to efficiently restore feature repre sentation capacity. Extensive experiments show that BiPointNet outperforms exist ing binarization methods by convincing margins, at the level even comparable wit h the full precision counterpart. We highlight that our techniques are generic, guaranteeing significant improvements on various fundamental tasks and mainstrea m backbones. Moreover, BiPointNet gives an impressive 14.7× speedup and 18.9× st

orage saving on real-world resource-constrained devices.
****
Center-wise Local Image Mixture For Contrastive Representation Learning
Hao Li,XIAOPENG ZHANG,Ruoyu Sun,Hongkai Xiong,Qi Tian
Recent advances in unsupervised representation learning have experienced remarka ble progress, especially with the achievements of contrastive learning, which re gards each image as well its augmentations as a separate class, while does not c onsider the semantic similarity among images. This paper proposes a new kind of data augmentation, named Center-wise Local Image Mixture, to expand the neighbor hood space of an image. CLIM encourages both local similarity and global aggrega tion while pulling similar images. This is achieved by searching local similar s amples of an image, and only selecting images that are closer to the correspondi ng cluster center, which we denote as center-wise local selection. As a result, similar representations are progressively approaching the clusters, while do not break the local similarity. Furthermore, image mixture is used as a smoothing r egularization to avoid overconfident the selected samples. Besides, we introduce multi-resolution augmentation, which enables the representation to be scale inv ariant. Integrating the two augmentations produces better feature representation on several unsupervised benchmarks. Notably, we reach 75.5% top-1 accuracy with linear evaluation over ResNet-50, and 59.3% top-1 accuracy when fine-tuned with only 1% labels, as well as consistently outperforming supervised pretraining on several downstream transfer tasks.
****
Mixture of Step Returns in Bootstrapped DQN
PoHan Chiang,Hsuan-Kung Yang,Zhang-Wei Hong,Chun-Yi Lee
The concept of utilizing multi-step returns for updating value functions has bee n adopted in deep reinforcement learning (DRL) for a number of years. Updating v alue functions with different backup lengths provides advantages in different as pects, including bias and variance of value estimates, convergence speed, and ex ploration behavior of the agent. Conventional methods such as TD-lambda leverage these advantages by using a target value equivalent to an exponential average o f different step returns. Nevertheless, integrating step returns into a single t arget sacrifices the diversity of the advantages offered by different step retur n targets. To address this issue, we propose Mixture Bootstrapped DQN (MB-DQN) b uilt on top of bootstrapped DQN, and uses different backup lengths for different bootstrapped heads. MB-DQN enables heterogeneity of the target values that is u navailable in approaches relying only on a single target value. As a result, it is able to maintain the advantages offered by different backup lengths. In this paper, we first discuss the motivational insights through a simple maze environm ent. In order to validate the effectiveness of MB-DQN, we perform experiments on the Atari 2600 benchmark environments and demonstrate the performance improveme nt of MB-DQN over a number of baseline methods. We further provide a set of abla tion studies to examine the impacts of different design configurations of MB-DQN .
****
Multilayer Dense Connections for Hierarchical Concept Classification
Toufiq Parag,Hongcheng Wang
Classification is a pivotal function for many computer vision tasks such as imag e recognition, object detection, scene segmentation. Multinomial logistic regre ssion with a single final layer of dense connections has become the ubiquitous t echnique for CNN-based classification. While these classifiers project a mapping between the input and a set of output category classes, they do not typically y ield a comprehensive description of the category. In particular, when a CNN bas ed image classifier correctly identifies the image of a Chimpanzee, its output d oes not clarify that Chimpanzee is a member of Primate, Mammal, Chordate familie s and a living thing. We propose a multilayer dense connectivity for a CNN to si multaneously predict the category \emph{and} its conceptual superclasses in hier archical order. We experimentally demonstrate that our proposed dense connection s, in conjunction with popular convolutional feature layers, can learn to predi ct the conceptual classes with minimal increase in network size while maintaini

ng the categorical classification accuracy.
****
Segmenting Natural Language Sentences via Lexical Unit Analysis
Yangming Li,lemao liu,Shuming Shi
In this work, we present Lexical Unit Analysis (LUA), a framework for general sequence segmentation tasks. Given a natural language sentence, LUA scores all the valid segmentation candidates and utilizes dynamic programming (DP) to extract the maximum scoring one. LUA enjoys a number of appealing properties such as inherently guaranteeing the predicted segmentation to be valid and facilitating globally optimal training and inference. Besides, the practical time complexity of LUA can be reduced to linear time, which is very efficient. We have conducted extensive experiments on 5 tasks, including syntactic chunking, named entity recognition (NER), slot filling, Chinese word segmentation, and Chinese part-of-speech (POS) tagging, across 15 datasets. Our models have achieved the state-of-the-art performances on 13 of them. The results also show that the F1 score of identifying long-length segments is notably improved.
****
AdaFuse: Adaptive Temporal Fusion Network for Efficient Action Recognition
Yue Meng,Rameswar Panda,Chung-Ching Lin,Prasanna Sattigeri,Leonid Karlinsky,Kate Saenko,Aude Oliva,Rogerio Feris
Temporal modelling is the key for efficient video action recognition. While understanding temporal information can improve recognition accuracy for dynamic actions, removing temporal redundancy and reusing past features can significantly save computation leading to efficient action recognition. In this paper, we introduce an adaptive temporal fusion network, called AdaFuse, that dynamically fuses channels from current and past feature maps for strong temporal modelling. Specifically, the necessary information from the historical convolution feature maps is fused with current pruned feature maps with the goal of improving both recognition accuracy and efficiency. In addition, we use a skipping operation to further reduce the computation cost of action recognition. Extensive experiments on SomethingV1 & V2, Jester and Mini-Kinetics show that our approach can achieve about 40% computation savings with comparable accuracy to state-of-the-art methods. The project page can be found at https://mengyuest.github.io/AdaFuse/
****
Truthful Self-Play
Shohei Ohsawa
We present a general framework for evolutionary learning to emergent unbiased state representation without any supervision. Evolutionary frameworks such as self-play converge to bad local optima in case of multi-agent reinforcement learning in non-cooperative partially observable environments with communication due to information asymmetry. Our proposed framework is a simple modification of self-play inspired by mechanism design, also known as {\em reverse game theory}, to elicit truthful signals and make the agents cooperative. The key idea is to add imaginary rewards using the peer prediction method, i.e., a mechanism for evaluating the validity of information exchanged between agents in a decentralized environment. Numerical experiments with predator prey, traffic junction and StarCraft tasks demonstrate that the state-of-the-art performance of our framework.

****
Transforming Recurrent Neural Networks with Attention and Fixed-point Equations
Zhaobin Xu,Baotian Hu,Buzhou Tang
Transformer has achieved state of the art performance in multiple Natural Language Processing tasks recently. Yet the Feed Forward Network(FFN) in a Transformer block is computationally expensive. In this paper, we present a framework to transform Recurrent Neural Networks(RNNs) and their variants into self-attention-style models, with an approximation of Banach Fixed-point Theorem. Within this framework, we propose a new model, StarSaber, by solving a set of equations obtained from RNN with Fixed-point Theorem and further approximate it with a Multi-layer Perceptron. It provides a view of stacking layers. StarSaber achieves better performance than both the vanilla Transformer and an improved version called ReZ

ero on three datasets and is more computationally efficient, due to the reductio n of Transformer's FFN layer. It has two major parts. One is a way to encode pos ition information with two different matrices. For every position in a sequence, we have a matrix operating on positions before it and another matrix operating on positions after it. The other is the introduction of direct paths from the in put layer to the rest of layers. Ablation studies show the effectiveness of thes e two parts. We additionally show that other RNN variants such as RNNs with gate s can also be transformed in the same way, outperforming the two kinds of Transf ormers as well.
****

LAYER SPARSITY IN NEURAL NETWORKS
Mohamed Hebiri,Johannes Lederer
Sparsity has become popular in machine learning, because it can save computation al resources, facilitate interpretations, and prevent overfitting. In this paper , we discuss sparsity in the framework of neural networks. In particular, we for mulate a new notion of sparsity that concerns the networks’ layers and, therefor e, aligns particularly well with the current trend toward deep networks. We call this notion layer sparsity. We then introduce corresponding regularization and refitting schemes that can complement standard deep-learning pipelines to genera te more compact and accurate networks.
****

Generating Furry Cars: Disentangling Object Shape and Appearance across Multiple Domains
Utkarsh Ojha,Krishna Kumar Singh,Yong Jae Lee
We consider the novel task of learning disentangled representations of object sh ape and appearance across multiple domains (e.g., dogs and cars). The goal is t o learn a generative model that learns an intermediate distribution, which borro ws a subset of properties from each domain, enabling the generation of images th at did not exist in any domain exclusively. This challenging problem requires a n accurate disentanglement of object shape, appearance, and background from each domain, so that the appearance and shape factors from the two domains can be in terchanged. We augment an existing approach that can disentangle factors within a single domain but struggles to do so across domains. Our key technical contri bution is to represent object appearance with a differentiable histogram of visu al features, and to optimize the generator so that two images with the same late nt appearance factor but different latent shape factors produce similar histogra ms. On multiple multi-domain datasets, we demonstrate our method leads to accura te and consistent appearance and shape transfer across domains.
****

Dynamic Graph: Learning Instance-aware Connectivity for Neural Networks
Kun Yuan,Quanquan Li,Dapeng Chen,Aojun Zhou,Junjie Yan
One practice of employing deep neural networks is to apply the same architecture to all the input instances. However, a fixed architecture may not be representa tive enough for data with high diversity. To promote the model capacity, existin g approaches usually employ larger convolutional kernels or deeper network struc ture, which may increase the computational cost. In this paper, we address this issue by raising the Dynamic Graph Network (DG-Net). The network learns the inst ance-aware connectivity, which creates different forward paths for different ins tances. Specifically, the network is initialized as a complete directed acyclic graph, where the nodes represent convolutional blocks and the edges represent th e connection paths. We generate edge weights by a learnable module \textit{route r} and select the edges whose weights are larger than a threshold, to adjust the connectivity of the neural network structure. Instead of using the same path of the network, DG-Net aggregates features dynamically in each node, which allows the network to have more representation ability. To facilitate the training, we represent the network connectivity of each sample in an adjacency matrix. The ma trix is updated to aggregate features in the forward pass, cached in the memory, and used for gradient computing in the backward pass. We verify the effectiven ess of our method with several static architectures, including MobileNetV2, ResN et, ResNeXt, and RegNet. Extensive experiments are performed on ImageNet classif

ication and COCO object detection, which shows the effectiveness and generalizat
ion ability of our approach.
****
Towards Defending Multiple Adversarial Perturbations via Gated Batch Normalizati
on
Aishan Liu,Shiyu Tang,Xianglong Liu,Xinyun Chen,Lei Huang,Zhuozhuo Tu,Dawn Song,
Dacheng Tao
There is now extensive evidence demonstrating that deep neural networks are vuln
erable to adversarial examples, motivating the development of defenses against a
dversarial attacks. However, existing adversarial defenses typically improve mod
el robustness against individual specific perturbation types. Some recent method
s improve model robustness against adversarial attacks in multiple ℓ_p ball
s, but their performance against each perturbation type is still far from satisf
actory. To better understand this phenomenon, we propose the \emph{multi-domain}
 hypothesis, stating that different types of adversarial perturbations are drawn
 from different domains. Guided by the multi-domain hypothesis, we propose~\emph
{Gated Batch Normalization (GBN)}, a novel building block for deep neural networ
ks that improves robustness against multiple perturbation types. GBN consists of
 a gated sub-network and a multi-branch batch normalization (BN) layer, where th
e gated sub-network separates different perturbation types, and each BN branch i
s in charge of a single perturbation type and learns domain-specific statistics
for input transformation. Then, features from different branches are aligned as
domain-invariant representations for the subsequent layers. We perform extensive
 evaluations of our approach on MNIST, CIFAR-10, and Tiny-ImageNet, and in doing
 so demonstrate that GBN outperforms previous defense proposals against multiple
 perturbation types, \ie, ℓ_1, ℓ_2, and ℓ_{∞} perturbations,
 by large margins of 10-20\%.
****
Parsed Categoric Encodings with Automunge
Nicholas Teague
The Automunge open source python library platform for tabular data pre-processin
g automates feature engineering data transformations of numerical encoding and m
issing data infill to received tidy data on bases fit to properties of columns i
n a designated train set for consistent and efficient application to subsequent
data pipelines such as for inference, where transformations may be applied to di
stinct columns in “family tree” sets with generations and branches of derivation
s. Included in the library of transformations are methods to extract structure f
rom bounded categorical string sets by way of automated string parsing, in which
 comparisons between entries in the set of unique values are parsed to identify
character subset overlaps which may be encoded by appended columns of boolean ov
erlap detection activations or by replacing string entries with identified overl
ap partitions. Further string parsing options, which may also be applied to unbo
unded categoric sets, include extraction of numeric substring partitions from en
tries or search functions to identify presence of specified substring partitions
. The aggregation of these methods into “family tree” sets of transformations ar
e demonstrated for use to automatically extract structure from categoric string
compositions in relation to the set of entries in a column, such as may be appli
ed to prepare categoric string set encodings for machine learning without human
intervention.
****
Pruning Neural Networks at Initialization: Why Are We Missing the Mark?
Jonathan Frankle,Gintare Karolina Dziugaite,Daniel Roy,Michael Carbin
Recent work has explored the possibility of pruning neural networks at initializ
ation. We assess proposals for doing so: SNIP (Lee et al., 2019), GraSP (Wang et
 al., 2020), SynFlow (Tanaka et al., 2020), and magnitude pruning. Although thes
e methods surpass the trivial baseline of random pruning, they remain below the
accuracy of magnitude pruning after training, and we endeavor to understand why.
 We show that, unlike pruning after training, randomly shuffling the weights the
se methods prune within each layer or sampling new initial values preserves or i
mproves accuracy. As such, the per-weight pruning decisions made by these method

s can be replaced by a per-layer choice of the fraction of weights to prune. Thi s property suggests broader challenges with the underlying pruning heuristics, t he desire to prune at initialization, or both.
****

Provable More Data Hurt in High Dimensional Least Squares Estimator
Zeng Li,Chuanlong Xie,QINWEN WANG
This paper investigates the finite-sample prediction risk of the high-dimensiona l least squares estimator. We derive the central limit theorem for the predictio n risk when both the sample size and the number of features tend to infinity. Fu rthermore, the finite-sample distribution and the confidence interval of the pre diction risk are provided. Our theoretical results demonstrate the sample-wise n on-monotonicity of the prediction risk and confirm ''more data hurt'' phenomenon .
****

ABS: Automatic Bit Sharing for Model Compression
Jing Liu,Bohan Zhuang,Peng Chen,Yong Guo,Chunhua Shen,Jianfei Cai,Mingkui Tan
We present Automatic Bit Sharing (ABS) to automatically search for optimal model compression configurations (e.g., pruning ratio and bitwidth). Unlike previous works that consider model pruning and quantization separately, we seek to optimi ze them jointly. To deal with the resultant large designing space, we propose a novel super-bit model, a single-path method, to encode all candidate compression configurations, rather than maintaining separate paths for each configuration. Specifically, we first propose a novel decomposition of quantization that encaps ulates all the candidate bitwidths in the search space. Starting from a low bitw idth, we sequentially consider higher bitwidths by recursively adding re-assignm ent offsets. We then introduce learnable binary gates to encode the choice of bi twidth, including filter-wise 0-bit for pruning. By jointly training the binary gates in conjunction with network parameters, the compression configurations of each layer can be automatically determined. Our ABS brings two benefits for mode l compression: 1) It avoids the combinatorially large design space, with a reduc ed number of trainable parameters and search costs. 2) It also averts directly f itting an extremely low bit quantizer to the data, hence greatly reducing the op timization difficulty due to the non-differentiable quantization. Experiments on CIFAR-100 and ImageNet show that our methods achieve significant computational cost reduction while preserving promising performance.
****

BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction
Yuhang Li,Ruihao Gong,Xu Tan,Yang Yang,Peng Hu,Qi Zhang,Fengwei Yu,Wei Wang,Shi Gu
We study the challenging task of neural network quantization without end-to-end retraining, called Post-training Quantization (PTQ). PTQ usually requires a smal l subset of training data but produces less powerful quantized models than Quant ization-Aware Training (QAT). In this work, we propose a novel PTQ framework, du bbed BRECQ, which pushes the limits of bitwidth in PTQ down to INT2 for the firs t time. BRECQ leverages the basic building blocks in neural networks and reconst ructs them one-by-one. In a comprehensive theoretical study of the second-order error, we show that BRECQ achieves a good balance between cross-layer dependency and generalization error. To further employ the power of quantization, the mixe d precision technique is incorporated in our framework by approximating the inte r-layer and intra-layer sensitivity. Extensive experiments on various handcrafte d and searched neural architectures are conducted for both image classification and object detection tasks. And for the first time we prove that, without bells and whistles, PTQ can attain 4-bit ResNet and MobileNetV2 comparable with QAT an d enjoy 240 times faster production of quantized models. Codes are available at https://github.com/yhhhli/BRECQ.
****