PAC-Bayesian Bounds on Rate-Efficient Classifiers

- Alhabib Abbas, Yiannis Andreopoulos
- abstract: We derive analytic bounds on the noise invariance of majority vote classifiers operating on compressed inputs. Specifically, starting from recent bounds on the true risk of majority vote classifiers, we extend the applicability of PAC-Bayesian theory to quantify the resilience of majority votes to input noise stemming from compression. The derived bounds are intuitive in binary classification settings, where they can be measured as expressions of voter differentials and voter pair agreement. By combining measures of input distortion with analytic guarantees on noise invariance, we prescribe rate-efficient machines to compress inputs without affecting subsequent classification. Our validation shows how bounding noise invariance can inform the compression stage for any majority vote classifier such that worst-case implications of bad input reconstructions are known, and inputs can be compressed to the minimum amount of information needed prior to inference.

Sharp-MAML: Sharpness-Aware Model-Agnostic Meta Learning

- Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, Tianyi Chen
- abstract: Model-agnostic meta learning (MAML) is currently one of the dominating approaches for few-shot meta-learning. Albeit its effectiveness, the optimization of MAML can be challenging due to the innate bilevel problem structure. Specifically, the loss landscape of MAML is much more complex with possibly more saddle points and local minimizers than its empirical risk minimization counterpart. To address this challenge, we leverage the recently invented sharpness-aware minimization and develop a sharpness-aware MAML approach that we term Sharp-MAML. We empirically demonstrate that Sharp-MAML and its computation-efficient variant can outperform the plain-vanilla MAML baseline (e.g., +3% accuracy on Mini-Imagenet). We complement the empirical study with the convergence rate analysis and the generalization bound of Sharp-MAML. To the best of our knowledge, this is the first empirical and theoretical study on sharpness-aware minimization in the context of bilevel learning.

An Initial Alignment between Neural Network and Target is Needed for Gradient Descent to Learn

- Emmanuel Abbe, Elisabetta Cornacchia, Jan Hazla, Christopher Marquis
- abstract: This paper introduces the notion of “Initial Alignment” (INAL) between a neural network at initialization and a target function. It is proved that if a network and a Boolean target function do not have a noticeable INAL, then noisy gradient descent with normalized i.i.d. initialization will not learn in polynomial time. Thus a certain amount of knowledge about the target (measured by the INAL) is needed in the architecture design. This also provides an answer to an open problem posed in (AS-NeurIPS’20). The results are based on deriving lower-bounds for descent algorithms on symmetric neural networks without explicit knowledge of the target function beyond its INAL.

Active Sampling for Min-Max Fairness

- Jacob D Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, Jie Zhang
- abstract: We propose simple active sampling and reweighting strategies for optimizing min-max fairness that can be applied to any classification or regression model learned via loss minimization. The key intuition behind our approach is to use at each timestep a datapoint from the group that is worst off under the current model for updating the model. The ease of implementation and the generality of our robust formulation make it an attractive option for improving model performance on disadvantaged groups. For convex learning problems, such as linear or logistic regression, we provide a fine-grained analysis, proving the rate of convergence to a min-max fair solution.

Meaningfully debugging model mistakes using conceptual counterfactual explanations

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

Batched Dueling Bandits

- Arpit Agarwal, Rohan Ghuge, Viswanath Nagarajan
- abstract: The K-armed dueling bandit problem, where the feedback is in the form of noisy pairwise comparisons, has been widely studied. Previous works have only focused on the sequential setting where the policy adapts after every comparison. However, in many applications such as search ranking and recommendation systems, it is preferable to perform comparisons in a limited number of parallel batches. We study the batched K-armed dueling bandit problem under two standard settings: (i) existence of a Condorcet winner, and (ii) strong stochastic transitivity and stochastic triangle inequality. For both settings, we obtain algorithms with a smooth trade-off between the number of batches and regret. Our regret bounds match the best known sequential regret bounds (up to poly-logarithmic factors), using only a logarithmic number of batches. We complement our regret analysis with a nearly-matching lower bound. Finally, we also validate our theoretical results via experiments on synthetic and real data.

Hierarchical Shrinkage: Improving the accuracy and interpretability of tree-based models.

- Abhineet Agarwal, Yan Shuo Tan, Omer Ronen, Chandan Singh, Bin Yu
- abstract: Decision trees and random forests (RF) are a cornerstone of modern machine learning practice. Due to their tendency to overfit, trees are typically regularized by a variety of techniques that modify their structure (e.g. pruning). We introduce Hierarchical Shrinkage (HS), a post-hoc algorithm which regularizes the tree not by altering its structure, but by shrinking the prediction over each leaf toward the sample means over each of its ancestors, with weights depending on a single regularization parameter and the number of samples in each ancestor. Since HS is a post-hoc method, it is extremely fast, compatible with any tree-growing algorithm and can be used synergistically with other regularization techniques. Extensive experiments over a wide variety of real-world datasets show that HS substantially increases the predictive performance of decision trees even when used in conjunction with other regularization techniques. Moreover, we find that applying HS to individual trees in a RF often improves its accuracy and interpretability by simplifying and stabilizing decision boundaries and SHAP values. We further explain HS by showing that it to be equivalent to ridge regression on a basis that is constructed of decision stumps associated to the internal nodes of a tree. All code and models are released in a full-fledged package available on Github

Deep equilibrium networks are sensitive to initialization statistics

- Atish Agarwala, Samuel S Schoenholz

- abstract: Deep equilibrium networks (DEQs) are a promising way to construct models which trade off memory for compute. However, theoretical understanding of these models is still lacking compared to traditional networks, in part because of the repeated application of a single set of weights. We show that DEQs are sensitive to the higher order statistics of the matrix families from which they are initialized. In particular, initializing with orthogonal or symmetric matrices allows for greater stability in training. This gives us a practical prescription for initializations which allow for training with a broader range of initial weight scales.

Learning of Cluster-based Feature Importance for Electronic Health Record Time-series

- Henrique Aguiar, Mauro Santos, Peter Watkinson, Tingting Zhu
- abstract: The recent availability of Electronic Health Records (EHR) has allowed for the development of algorithms predicting inpatient risk of deterioration and trajectory evolution. However, prediction of disease progression with EHR is challenging since these data are sparse, heterogeneous, multi-dimensional, and multi-modal time-series. As such, clustering is regularly used to identify similar groups within the patient cohort to improve prediction. Current models have shown some success in obtaining cluster representations of patient trajectories. However, they i) fail to obtain clinical interpretability for each cluster, and ii) struggle to learn meaningful cluster numbers in the context of imbalanced distribution of disease outcomes. We propose a supervised deep learning model to cluster EHR data based on the identification of clinically understandable phenotypes with regard to both outcome prediction and patient trajectory. We introduce novel loss functions to address the problems of class imbalance and cluster collapse, and furthermore propose a feature-time attention mechanism to identify cluster-based phenotype importance across time and feature dimensions. We tested our model in two datasets corresponding to distinct medical settings. Our model yielded added interpretability to cluster formation and outperformed benchmarks by at least 4% in relevant metrics.

On the Convergence of the Shapley Value in Parametric Bayesian Learning Games

- Lucas Agussurja, Xinyi Xu, Bryan Kian Hsiang Low
- abstract: Measuring contributions is a classical problem in cooperative game theory where the Shapley value is the most well-known solution concept. In this paper, we establish the convergence property of the Shapley value in parametric Bayesian learning games where players perform a Bayesian inference using their combined data, and the posterior-prior KL divergence is used as the characteristic function. We show that for any two players, under some regularity conditions, their difference in Shapley value converges in probability to the difference in Shapley value of a limiting game whose characteristic function is proportional to the log-determinant of the joint Fisher information. As an application, we present an online collaborative learning framework that is asymptotically Shapley-fair. Our result enables this to be achieved without any costly computations of posterior-prior KL divergences. Only a consistent estimator of the Fisher information is needed. The effectiveness of our framework is demonstrated with experiments using real-world data.

Individual Preference Stability for Clustering

- Saba Ahmadi, Pranjal Awasthi, Samir Khuller, Matthäus Kleindessner, Jamie Morgenstern, Pattara Sukprasert, Ali Vakilian
- abstract: In this paper, we propose a natural notion of individual preference (IP) stability for clustering, which asks that every data point, on average, is closer to the points in its own cluster than to the points in any other cluster. Our notion can be motivated from several perspectives, including game theory and algorithmic fairness. We study several questions related to our proposed notion. We first show that deciding whether a given data set allows for an IP-stable clustering in general is NP-hard. As a result, we explore the design of efficient algorithms for finding IP-stable clusterings in some restricted metric spaces. We present a polytime algorithm to find a clustering satisfying exact IP-stability on the real line, and an efficient algorithm to find an IP-stable 2-clustering for a tree metric. We also consider relaxing the stability constraint, i.e., every data point should not be too far from its own cluster compared to any other cluster. For this case, we provide polytime algorithms with different guarantees. We evaluate some of our algorithms and several standard clustering approaches on real data sets.

Understanding the unstable convergence of gradient descent

- Kwangjun Ahn, Jingzhao Zhang, Suvrit Sra
- abstract: Most existing analyses of (stochastic) gradient descent rely on the condition that for L-smooth costs, the step size is less than $2/L$. However, many works have observed that in machine learning applications step sizes often do not fulfill this condition, yet (stochastic) gradient descent still converges, albeit in an unstable manner. We investigate this unstable convergence phenomenon from first principles, and discuss key causes behind it. We also identify its main characteristics, and how they interrelate based on both theory and experiments, offering a principled view toward understanding the phenomenon.

Minimum Cost Intervention Design for Causal Effect Identification

- Sina Akbari, Jalal Etesami, Negar Kiyavash
- abstract: Pearl’s do calculus is a complete axiomatic approach to learn the identifiable causal effects from observational data. When such an effect is not identifiable, it is necessary to perform a collection of often costly interventions in the system to learn the causal effect. In this work, we consider the problem of designing the collection of interventions with the minimum cost to identify the desired effect. First, we prove that this prob-em is NP-complete, and subsequently propose an algorithm that can either find the optimal solution or a logarithmic-factor approximation of it. This is done by establishing a connection between our problem and the minimum hitting set problem. Additionally, we propose several polynomial time heuristic algorithms to tackle the computational complexity of the problem. Although these algorithms could potentially stumble on sub-optimal solutions, our simulations show that they achieve small regrets on random graphs.

How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models

- Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, Mihaela van der Schaar
- abstract: Devising domain- and model-agnostic evaluation metrics for generative models is an important and as yet unresolved problem. Most existing metrics, which were tailored solely to the image synthesis setup, exhibit a limited capacity for diagnosing the different modes of failure of generative models across broader application domains. In this paper, we introduce a 3-dimensional evaluation metric, (α-Precision, β-Recall, Authenticity), that characterizes the fidelity, diversity and generalization performance of any generative model in a domain-agnostic fashion. Our metric unifies statistical divergence measures with precision-recall analysis, enabling sample- and distribution-level diagnoses of model fidelity and diversity. We introduce generalization as an additional, independent dimension (to the fidelity-diversity trade-off) that quantifies the extent to which a model copies training data{—}a crucial performance indicator when modeling sensitive data with requirements on privacy. The three metric components correspond to (interpretable) probabilistic quantities, and are estimated via sample-level binary classification. The sample-level nature of our metric inspires a novel use case which we call model auditing, wherein we judge the quality of individual samples generated by a (black-box) model, discarding low-quality samples and hence improving the overall model performance in a post-hoc manner.

A Natural Actor-Critic Framework for Zero-Sum Markov Games

- Ahmet Alacaoglu, Luca Viano, Niao He, Volkan Cevher

- abstract: We introduce algorithms based on natural actor-critic and analyze their sample complexity for solving two player zero-sum Markov games in the tabular case. Our results improve the best-known sample complexities of policy gradient/actor-critic methods for convergence to Nash equilibrium in the multi-agent setting. We use the error propagation scheme in approximate dynamic programming, recent advances for global convergence of policy gradient methods, temporal difference learning, and techniques from stochastic primal-dual optimization. Our algorithms feature two stages, requiring agents to agree on an etiquette before starting their interactions, which is feasible for instance in self-play. However, the agents only access to joint reward and joint next state and not to each other's actions or policies. Our complexity results match the best-known results for global convergence of policy gradient algorithms for single agent RL. We provide numerical verification of our methods for a two player bandit environment and a two player game, Alesia. We observe improved empirical performance as compared to the recently proposed optimistic gradient descent-ascent variant for Markov games.

Deploying Convolutional Networks on Untrusted Platforms Using 2D Holographic Reduced Representations

- Mohammad Mahmudul Alam, Edward Raff, Tim Oates, James Holt
- abstract: Due to the computational cost of running inference for a neural network, the need to deploy the inferential steps on a third party's compute environment or hardware is common. If the third party is not fully trusted, it is desirable to obfuscate the nature of the inputs and outputs, so that the third party can not easily determine what specific task is being performed. Provably secure protocols for leveraging an untrusted party exist but are too computational demanding to run in practice. We instead explore a different strategy of fast, heuristic security that we call Connectionist Symbolic Pseudo Secrets. By leveraging Holographic Reduced Representations (HRRs), we create a neural network with a pseudo-encryption style defense that empirically shows robustness to attack, even under threat models that unrealistically favor the adversary.

Optimistic Linear Support and Successor Features as a Basis for Optimal Policy Transfer

- Lucas Nunes Alegre, Ana Bazzan, Bruno C. Da Silva
- abstract: In many real-world applications, reinforcement learning (RL) agents might have to solve multiple tasks, each one typically modeled via a reward function. If reward functions are expressed linearly, and the agent has previously learned a set of policies for different tasks, successor features (SFs) can be exploited to combine such policies and identify reasonable solutions for new problems. However, the identified solutions are not guaranteed to be optimal. We introduce a novel algorithm that addresses this limitation. It allows RL agents to combine existing policies and directly identify optimal policies for arbitrary new problems, without requiring any further interactions with the environment. We first show (under mild assumptions) that the transfer learning problem tackled by SFs is equivalent to the problem of learning to optimize multiple objectives in RL. We then introduce an SF-based extension of the Optimistic Linear Support algorithm to learn a set of policies whose SFs form a convex coverage set. We prove that policies in this set can be combined via generalized policy improvement to construct optimal behaviors for any new linearly-expressible tasks, without requiring any additional training samples. We empirically show that our method outperforms state-of-the-art competing algorithms both in discrete and continuous domains under value function approximation.

Structured Stochastic Gradient MCMC

- Antonios Alexos, Alex J Boyd, Stephan Mandt
- abstract: Stochastic gradient Markov Chain Monte Carlo (SGMCMC) is a scalable algorithm for asymptotically exact Bayesian inference in parameter-rich models, such as Bayesian neural networks. However, since mixing can be slow in high dimensions, practitioners often resort to variational inference (VI). Unfortunately, VI makes strong assumptions on both the factorization and functional form of the posterior. To relax these assumptions, this work proposes a new non-parametric variational inference scheme that combines ideas from both SGMCMC and coordinate-ascent VI. The approach relies on a new Langevin-type algorithm that operates on a "self-averaged" posterior energy function, where parts of the latent variables are averaged over samples from earlier iterations of the Markov chain. This way, statistical dependencies between coordinates can be broken in a controlled way, allowing the chain to mix faster. This scheme can be further modified in a "dropout" manner, leading to even more scalability. We test our scheme for ResNet-20 on CIFAR-10, SVHN, and FMNIST. In all cases, we find improvements in convergence speed and/or final accuracy compared to SGMCMC and parametric VI.

XAI for Transformers: Better Explanations through Conservative Propagation

- Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, Lior Wolf
- abstract: Transformers have become an important workhorse of machine learning, with numerous applications. This necessitates the development of reliable methods for increasing their transparency. Multiple interpretability methods, often based on gradient information, have been proposed. We show that the gradient in a Transformer reflects the function only locally, and thus fails to reliably identify the contribution of input features to the prediction. We identify Attention Heads and LayerNorm as main reasons for such unreliable explanations and propose a more stable way for propagation through these layers. Our proposal, which can be seen as a proper extension of the well-established LRP method to Transformers, is shown both theoretically and empirically to overcome the deficiency of a simple gradient-based approach, and achieves state-of-the-art explanation performance on a broad range of Transformer models and datasets.

RUMs from Head-to-Head Contests

- Matteo Almanza, Flavio Chierichetti, Ravi Kumar, Alessandro Panconesi, Andrew Tomkins
- abstract: Random utility models (RUMs) encode the likelihood that a particular item will be selected from a slate of competing items. RUMs are well-studied objects in both discrete choice theory and, more recently, in the machine learning community, as they encode a fairly broad notion of rational user behavior. In this paper, we focus on slates of size two representing head-to-head contests. Given a tournament matrix M such that $M_{i,j}$ is the probability that item j will be selected from ${i, j}$, we consider the problem of finding the RUM that most closely reproduces M. For this problem we obtain a polynomial-time algorithm returning a RUM that approximately minimizes the average error over the pairs. Our experiments show that RUMs can perfectly represent many of the tournament matrices that have been considered in the literature; in fact, the maximum average error induced by RUMs on the matrices we considered is negligible (≈ 0.001). We also show that RUMs are competitive, on prediction tasks, with previous approaches.

Neuro-Symbolic Language Modeling with Automaton-augmented Retrieval

- Uri Alon, Frank Xu, Junxian He, Sudipta Sengupta, Dan Roth, Graham Neubig
- abstract: Retrieval-based language models (R-LM) model the probability of natural language text by combining a standard language model (LM) with examples retrieved from an external datastore at test time. While effective, a major bottleneck of using these models in practice is the computationally costly datastore search, which can be performed as frequently as every time step. In this paper, we present RetoMaton - retrieval automaton - which approximates the datastore search, based on (1) saving pointers between consecutive datastore entries, and (2) clustering of entries into "states". This effectively results in a weighted finite automaton built on top of the datastore, instead of representing the datastore as a flat list. The creation of the automaton is unsupervised, and a RetoMaton can be constructed from any text collection: either the original training corpus or from another domain. Traversing this automaton at inference time, in parallel to the LM inference, reduces its perplexity by up to 1.85, or alternatively saves up to 83% of the nearest neighbor searches over kNN-LM (Khandelwal et al., 2020) without hurting perplexity. Our code and trained models are available at https://github.com/neulab/retomaton .

Minimax Classification under Concept Drift with Multidimensional Adaptation and Performance Guarantees

- Verónica Álvarez, Santiago Mazuelas, Jose A Lozano
- abstract: The statistical characteristics of instance-label pairs often change with time in practical scenarios of supervised classification. Conventional learning techniques adapt to such concept drift accounting for a scalar rate of change by means of a carefully chosen learning rate, forgetting factor, or window size. However, the time changes in common scenarios are multidimensional, i.e., different statistical characteristics often change in a different manner. This paper presents adaptive minimax risk classifiers (AMRCs) that account for multidimensional time changes by means of a multivariate and high-order tracking of the time-varying underlying distribution. In addition, differently from conventional techniques, AMRCs can provide computable tight performance guarantees. Experiments on multiple benchmark datasets show the classification improvement of AMRCs compared to the state-of-the-art and the reliability of the presented performance guarantees.

Scalable First-Order Bayesian Optimization via Structured Automatic Differentiation

- Sebastian E Ament, Carla P Gomes
- abstract: Bayesian Optimization (BO) has shown great promise for the global optimization of functions that are expensive to evaluate, but despite many successes, standard approaches can struggle in high dimensions. To improve the performance of BO, prior work suggested incorporating gradient information into a Gaussian process surrogate of the objective, giving rise to kernel matrices of size nd {\texttimes} nd for n observations in d dimensions. Naïvely multiplying with (resp. inverting) these matrices requires $O(n^2d^2)$ (resp. $O(n^3d^3)$) operations, which becomes infeasible for moderate dimensions and sample sizes. Here, we observe that a wide range of kernels gives rise to structured matrices, enabling an exact $O(n^2d)$ matrix-vector multiply for gradient observations and $O(n^2d^2)$ for Hessian observations. Beyond canonical kernel classes, we derive a programmatic approach to leveraging this type of structure for transformations and combinations of the discussed kernel classes, which constitutes a structure-aware automatic differentiation algorithm. Our methods apply to virtually all canonical kernels and automatically extend to complex kernels, like the neural network, radial basis function network, and spectral mixture kernels without any additional derivations, enabling flexible, problem-dependent modeling while scaling first-order BO to high d.

Public Data-Assisted Mirror Descent for Private Model Training

- Ehsan Amid, Arun Ganesh, Rajiv Mathews, Swaroop Ramaswamy, Shuang Song, Thomas Steinke, Thomas Steinke, Vinith M Suriyakumar, Om Thakkar, Abhradeep Thakurta
- abstract: In this paper, we revisit the problem of using in-distribution public data to improve the privacy/utility trade-offs for differentially private (DP) model training. (Here, public data refers to auxiliary data sets that have no privacy concerns.) We design a natural variant of DP mirror descent, where the DP gradients of the private/sensitive data act as the linear term, and the loss generated by the public data as the mirror map. We show that, for linear regression with feature vectors drawn from a non-isotropic sub-Gaussian distribution, our algorithm, PDA-DPMD (a variant of mirror descent), provides population risk guarantees that are asymptotically better than the best known guarantees under DP (without having access to public data), when the number of public data samples is sufficiently large. We further show that our algorithm has natural “noise stability” properties that control the variance due to noise added to ensure DP. We demonstrate the efficacy of our algorithm by showing privacy/utility trade-offs on four benchmark datasets (StackOverflow, WikiText-2, CIFAR-10, and EMNIST). We show that our algorithm not only significantly improves over traditional DP-SGD, which does not have access to public data, but to our knowledge is the first to improve over DP-SGD on models that have been pre-trained with public data.

On Last-Iterate Convergence Beyond Zero-Sum Games

- Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, Tuomas Sandholm
- abstract: Most existing results about last-iterate convergence of learning dynamics are limited to two-player zero-sum games, and only apply under rigid assumptions about what dynamics the players follow. In this paper we provide new results and techniques that apply to broader families of games and learning dynamics. First, we show that in a class of games that includes constant-sum polymatrix and strategically zero-sum games, the trajectories of dynamics such as optimistic mirror descent (OMD) exhibit a boundedness property, which holds even when players employ different algorithms and prediction mechanisms. This property enables us to obtain $O(1/\sqrt{T})$ rates and optimal $O(1)$ regret bounds. Our analysis also reveals a surprising property: OMD either reaches arbitrarily close to a Nash equilibrium or it outperforms the robust price of anarchy in efficiency. Moreover, for potential games we establish convergence to an ϵ-equilibrium after $O(1/\epsilon^2)$ iterations for mirror descent under a broad class of regularizers, as well as optimal $O(1)$ regret bounds for OMD variants. Our framework also extends to near-potential games, and unifies known analyses for distributed learning in Fisher’s market model. Finally, we analyze the convergence, efficiency, and robustness of optimistic gradient descent (OGD) in general-sum continuous games.

Online Algorithms with Multiple Predictions

- Keerti Anand, Rong Ge, Amit Kumar, Debmalya Panigrahi
- abstract: This paper studies online algorithms augmented with multiple machine-learned predictions. We give a generic algorithmic framework for online covering problems with multiple predictions that obtains an online solution that is competitive against the performance of the best solution obtained from the predictions. Our algorithm incorporates the use of predictions in the classic potential-based analysis of online algorithms. We apply our algorithmic framework to solve classical problems such as online set cover, (weighted) caching, and online facility location in the multiple predictions setting.

Learning to Hash Robustly, Guaranteed

- Alexandr Andoni, Daniel Beaglehole
- abstract: The indexing algorithms for the high-dimensional nearest neighbor search (NNS) with the best worst-case guarantees are based on the randomized Locality Sensitive Hashing (LSH), and its derivatives. In practice, many heuristic approaches exist to "learn" the best indexing method in order to speed-up NNS, crucially adapting to the structure of the given dataset. Oftentimes, these heuristics outperform the LSH-based algorithms on real datasets, but, almost always, come at the cost of losing the guarantees of either correctness or robust performance on adversarial queries, or apply to datasets with an assumed extra structure/model. In this paper, we design an NNS algorithm for the Hamming space that has worst-case guarantees essentially matching that of theoretical algorithms, while optimizing the hashing to the structure of the dataset (think instance-optimal algorithms) for performance on the minimum-performing query. We evaluate the algorithm’s ability to optimize for a given dataset both theoretically and practically. On the theoretical side, we exhibit a natural setting (dataset model) where our algorithm is much better than the standard theoretical one. On the practical side, we run experiments that show that our algorithm has a 1.8x and 2.1x better recall on the worst-performing queries to the MNIST and ImageNet datasets.

Set Based Stochastic Subsampling

- Bruno Andreis, Seanie Lee, A. Tuan Nguyen, Juho Lee, Eunho Yang, Sung Ju Hwang
- abstract: Deep models are designed to operate on huge volumes of high dimensional data such as images. In order to reduce the volume of data these models must process, we propose a set-based two-stage end-to-end neural subsampling model that is jointly optimized with an arbitrary downstream task network (e.g. classifier). In the first stage, we efficiently subsample candidate elements using conditionally independent Bernoulli random variables by capturing coarse grained global information using set encoding functions, followed by conditionally dependent autoregressive subsampling of the candidate elements using Categorical random variables by modeling pair-wise interactions using set attention networks in the second stage. We apply our method to feature and instance selection and show that it outperforms the relevant baselines under low subsampling rates on a variety of tasks including

image classification, image reconstruction, function reconstruction and few-shot classification. Additionally, for nonparametric models such as Neural Processes that require to leverage the whole training data at inference time, we show that our method enhances the scalability of these models.

Towards Understanding Sharpness-Aware Minimization

- Maksym Andriushchenko, Nicolas Flammarion
- abstract: Sharpness-Aware Minimization (SAM) is a recent training method that relies on worst-case weight perturbations which significantly improves generalization in various settings. We argue that the existing justifications for the success of SAM which are based on a PAC-Bayes generalization bound and the idea of convergence to flat minima are incomplete. Moreover, there are no explanations for the success of using m-sharpness in SAM which has been shown as essential for generalization. To better understand this aspect of SAM, we theoretically analyze its implicit bias for diagonal linear networks. We prove that SAM always chooses a solution that enjoys better generalization properties than standard gradient descent for a certain class of problems, and this effect is amplified by using m-sharpness. We further study the properties of the implicit bias on non-linear networks empirically, where we show that fine-tuning a standard model with SAM can lead to significant generalization improvements. Finally, we provide convergence results of SAM for non-convex objectives when used with stochastic gradients. We illustrate these results empirically for deep networks and discuss their relation to the generalization behavior of SAM. The code of our experiments is available at https://github.com/tml-epfl/understanding-sam.

Fair and Fast k-Center Clustering for Data Summarization

- Haris Angelidakis, Adam Kurpisz, Leon Sering, Rico Zenklusen
- abstract: We consider two key issues faced by many clustering methods when used for data summarization, namely (a) an unfair representation of "demographic groups" and (b) distorted summarizations, where data points in the summary represent subsets of the original data of vastly different sizes. Previous work made important steps towards handling separately each of these two issues in the context of the fundamental k-Center clustering objective through the study of fast algorithms for natural models that address them. We show that it is possible to effectively address both (a) and (b) simultaneously by presenting a clustering procedure that works for a canonical combined model and (i) is fast, both in theory and practice, (ii) exhibits a worst-case constant-factor guarantee, and (iii) gives promising computational results showing that there can be significant benefits in addressing both issues together instead of sequentially.

Interactive Correlation Clustering with Existential Cluster Constraints

- Rico Angell, Nicholas Monath, Nishant Yadav, Andrew Mccallum
- abstract: We consider the problem of clustering with user feedback. Existing methods express constraints about the input data points, most commonly through must-link and cannot-link constraints on data point pairs. In this paper, we introduce existential cluster constraints: a new form of feedback where users indicate the features of desired clusters. Specifically, users make statements about the existence of a cluster having (and not having) particular features. Our approach has multiple advantages: (1) constraints on clusters can express user intent more efficiently than point pairs; (2) in cases where the users' mental model is of the desired clusters, it is more natural for users to express cluster-wise preferences; (3) it functions even when privacy restrictions prohibit users from seeing raw data. In addition to introducing existential cluster constraints, we provide an inference algorithm for incorporating our constraints into the output clustering. Finally, we demonstrate empirically that our proposed framework facilitates more accurate clustering with dramatically fewer user feedback inputs.

Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

- Anastasios N Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, Yaniv Romano
- abstract: Image-to-image regression is an important learning task, used frequently in biological imaging. Current algorithms, however, do not generally offer statistical guarantees that protect against a model's mistakes and hallucinations. To address this, we develop uncertainty quantification techniques with rigorous statistical guarantees for image-to-image regression problems. In particular, we show how to derive uncertainty intervals around each pixel that are guaranteed to contain the true value with a user-specified confidence probability. Our methods work in conjunction with any base machine learning model, such as a neural network, and endow it with formal mathematical guarantees{—}regardless of the true unknown data distribution or choice of model. Furthermore, they are simple to implement and computationally inexpensive. We evaluate our procedure on three image-to-image regression tasks: quantitative phase microscopy, accelerated magnetic resonance imaging, and super-resolution transmission electron microscopy of a Drosophila melanogaster brain.

AdaGrad Avoids Saddle Points

- Kimon Antonakopoulos, Panayotis Mertikopoulos, Georgios Piliouras, Xiao Wang
- abstract: Adaptive first-order methods in optimization have widespread ML applications due to their ability to adapt to non-convex landscapes. However, their convergence guarantees are typically stated in terms of vanishing gradient norms, which leaves open the issue of converging to undesirable saddle points (or even local maxima). In this paper, we focus on the AdaGrad family of algorithms - from scalar to full-matrix preconditioning - and we examine the question of whether the method's trajectories avoid saddle points. A major challenge that arises here is that AdaGrad's step-size (or, more accurately, the method's preconditioner) evolves over time in a filtration-dependent way, i.e., as a function of all gradients observed in earlier iterations; as a result, avoidance results for methods with a constant or vanishing step-size do not apply. We resolve this challenge by combining a series of step-size stabilization arguments with a recursive representation of the AdaGrad preconditioner that allows us to employ center-stable techniques and ultimately show that the induced trajectories avoid saddle points from almost any initial condition.

UnderGrad: A Universal Black-Box Optimization Method with Almost Dimension-Free Convergence Rate Guarantees

- Kimon Antonakopoulos, Dong Quan Vu, Volkan Cevher, Kfir Levy, Panayotis Mertikopoulos
- abstract: Universal methods achieve optimal convergence rate guarantees in convex optimization without any prior knowledge of the problem's regularity parameters or the attributes of the gradient oracle employed by the method. In this regard, existing state-of-the-art algorithms achieve an $O(1/T^2)$ convergence rate in Lipschitz smooth problems with a perfect gradient oracle, and an $O(1/sqrt{T})$ convergence speed when the underlying problem is non-smooth and/or the gradient oracle is stochastic. On the downside, these methods do not take into account the dependence of these guarantees on the problem's dimensionality, and this can have a catastrophic impact on a method's convergence, in both theory and practice. Our paper aims to bridge this gap by providing a scalable universal method - dubbed UnDERGrad - which enjoys an almost dimension-free oracle complexity in problems with a favorable geometry (like the simplex, ℓ_1-ball or trace-constraints), while retaining the order-optimal dependence on T described above. These "best of both worlds" guarantees are achieved via a primal-dual update scheme inspired by the dual exploration method for variational inequalities.

Adapting the Linearised Laplace Model Evidence for Modern Deep Learning

- Javier Antoran, David Janz, James U Allingham, Erik Daxberger, Riccardo Rb Barbano, Eric Nalisnick, Jose Miguel Hernandez-Lobato
- abstract: The linearised Laplace method for estimating model uncertainty has received renewed attention in the Bayesian deep learning community. The method provides reliable error bars and admits a closed-form expression for the model evidence, allowing for scalable selection of model hyperparameters. In this work, we examine the assumptions behind this method, particularly in conjunction with model selection. We show that these

interact poorly with some now-standard tools of deep learning–stochastic approximation methods and normalisation layers–and make recommendations for how to better adapt this classic method to the modern setting. We provide theoretical support for our recommendations and validate them empirically on MLPs, classic CNNs, residual networks with and without normalisation layers, generative autoencoders and transformers.

EAT-C: Environment-Adversarial sub-Task Curriculum for Efficient Reinforcement Learning

- Shuang Ao, Tianyi Zhou, Jing Jiang, Guodong Long, Xuan Song, Chengqi Zhang
- abstract: Reinforcement learning (RL) is inefficient on long-horizon tasks due to sparse rewards and its policy can be fragile to slightly perturbed environments. We address these challenges via a curriculum of tasks with coupled environments, generated by two policies trained jointly with RL: (1) a co-operative planning policy recursively decomposing a hard task into a coarse-to-fine sub-task tree; and (2) an adversarial policy modifying the environment in each sub-task. They are complementary to acquire more informative feedback for RL: (1) provides dense reward of easier sub-tasks while (2) modifies sub-tasks' environments to be more challenging and diverse. Conversely, they are trained by RL's dense feedback on sub-tasks so their generated curriculum keeps adaptive to RL's progress. The sub-task tree enables an easy-to-hard curriculum for every policy: its top-down construction gradually increases sub-tasks the planner needs to generate, while the adversarial training between the environment and RL follows a bottom-up traversal that starts from a dense sequence of easier sub-tasks allowing more frequent environment changes. We compare EAT-C with RL/planning targeting similar problems and methods with environment generators or adversarial agents. Extensive experiments on diverse tasks demonstrate the advantages of our method on improving RL's efficiency and generalization.

Online Balanced Experimental Design

- David Arbour, Drew Dimmery, Tung Mai, Anup Rao
- abstract: We consider the experimental design problem in an online environment, an important practical task for reducing the variance of estimates in randomized experiments which allows for greater precision, and in turn, improved decision making. In this work, we present algorithms that build on recent advances in online discrepancy minimization which accommodate both arbitrary treatment probabilities and multiple treatments. The proposed algorithms are computational efficient, minimize covariate imbalance, and include randomization which enables robustness to misspecification. We provide worst case bounds on the expected mean squared error of the causal estimate and show that the proposed estimator is no worse than an implicit ridge regression, which are within a logarithmic factor of the best known results for offline experimental design. We conclude with a detailed simulation study showing favorable results relative to complete randomization as well as to offline methods for experimental design with time complexities exceeding our algorithm, which has a linear dependence on the number of observations, by polynomial factors.

VariGrow: Variational Architecture Growing for Task-Agnostic Continual Learning based on Bayesian Novelty

- Randy Ardywibowo, Zepeng Huo, Zhangyang Wang, Bobak J Mortazavi, Shuai Huang, Xiaoning Qian
- abstract: Continual Learning (CL) is the problem of sequentially learning a set of tasks and preserving all the knowledge acquired. Many existing methods assume that the data stream is explicitly divided into a sequence of known contexts (tasks), and use this information to know when to transfer knowledge from one context to another. Unfortunately, many real-world CL scenarios have no clear task nor context boundaries, motivating the study of task-agnostic CL, where neither the specific tasks nor their switches are known both in training and testing. This paper proposes a variational architecture growing framework dubbed VariGrow. By interpreting dynamically growing neural networks as a Bayesian approximation, and defining flexible implicit variational distributions, VariGrow detects if a new task is arriving through an energy-based novelty score. If the novelty score is high and the sample is "detected" as a new task, VariGrow will grow a new expert module to be responsible for it. Otherwise, the sample will be assigned to one of the existing experts who is most "familiar" with it (i.e., one with the lowest novelty score). We have tested VariGrow on several CIFAR and ImageNet-based benchmarks for the strict task-agnostic CL setting and demonstrate its consistent superior performance. Perhaps surprisingly, its performance can even be competitive compared to task-aware methods.

Thresholded Lasso Bandit

- Kaito Ariu, Kenshi Abe, Alexandre Proutiere
- abstract: In this paper, we revisit the regret minimization problem in sparse stochastic contextual linear bandits, where feature vectors may be of large dimension d, but where the reward function depends on a few, say $s_0\ll d$, of these features only. We present Thresholded Lasso bandit, an algorithm that (i) estimates the vector defining the reward function as well as its sparse support, i.e., significant feature elements, using the Lasso framework with thresholding, and (ii) selects an arm greedily according to this estimate projected on its support. The algorithm does not require prior knowledge of the sparsity index s_0 and can be parameter-free under some symmetric assumptions. For this simple algorithm, we establish non-asymptotic regret upper bounds scaling as $\mathcal{O}(\log d + \sqrt{T})$ in general, and as $\mathcal{O}(\log d + \log T)$ under the so-called margin condition (a probabilistic condition on the separation of the arm rewards). The regret of previous algorithms scales as $\mathcal{O}(\log d + \sqrt{T \log (d T)})$ and $\mathcal{O}(\log T \log d)$ in the two settings, respectively. Through numerical experiments, we confirm that our algorithm outperforms existing methods.

Gradient Based Clustering

- Aleksandar Armacki, Dragana Bajovic, Dusan Jakovetic, Soummya Kar
- abstract: We propose a general approach for distance based clustering, using the gradient of the cost function that measures clustering quality with respect to cluster assignments and cluster center positions. The approach is an iterative two step procedure (alternating between cluster assignment and cluster center updates) and is applicable to a wide range of functions, satisfying some mild assumptions. The main advantage of the proposed approach is a simple and computationally cheap update rule. Unlike previous methods that specialize to a specific formulation of the clustering problem, our approach is applicable to a wide range of costs, including non-Bregman clustering methods based on the Huber loss. We analyze the convergence of the proposed algorithm, and show that it converges to the set of appropriately defined fixed points, under arbitrary center initialization. In the special case of Bregman cost functions, the algorithm converges to the set of centroidal Voronoi partitions, which is consistent with prior works. Numerical experiments on real data demonstrate the effectiveness of the proposed method.

Understanding Gradient Descent on the Edge of Stability in Deep Learning

- Sanjeev Arora, Zhiyuan Li, Abhishek Panigrahi
- abstract: Deep learning experiments by \citet{cohen2021gradient} using deterministic Gradient Descent (GD) revealed an Edge of Stability (EoS) phase when learning rate (LR) and sharpness (i.e., the largest eigenvalue of Hessian) no longer behave as in traditional optimization. Sharpness stabilizes around $2/$LR and loss goes up and down across iterations, yet still with an overall downward trend. The current paper mathematically analyzes a new mechanism of implicit regularization in the EoS phase, whereby GD updates due to non-smooth loss landscape turn out to evolve along some deterministic flow on the manifold of minimum loss. This is in contrast to many previous results about implicit bias either relying on infinitesimal updates or noise in gradient. Formally, for any smooth function L with certain regularity condition, this effect is demonstrated for (1) Normalized GD, i.e., GD with a varying LR $\eta_t =\frac{\eta}{\norm{\nabla L(x(t))}}$ and loss L; (2) GD with constant LR and loss $\sqrt{L- \min_x L(x)}$. Both provably enter the Edge of Stability, with the associated flow on the manifold minimizing $\lambda_{1}(\nabla^2 L)$. The above theoretical results have been corroborated by an experimental study.

Private optimization in the interpolation regime: faster rates and hardness results

- Hilal Asi, Karan Chadha, Gary Cheng, John Duchi
- abstract: In non-private stochastic convex optimization, stochastic gradient methods converge much faster on interpolation problems—namely, problems where there exists a solution that simultaneously minimizes all of the sample losses—than on non-interpolating ones; similar improvements are not known in the private setting. In this paper, we investigate differentially private stochastic optimization in the interpolation regime. First, we show that without additional assumptions, interpolation problems do not exhibit an improved convergence rates with differential privacy. However, when the functions exhibit quadratic growth around the optimum, we show (near) exponential improvements in the private sample complexity. In particular, we propose an adaptive algorithm that improves the sample complexity to achieve expected error α from $\frac{d}{\diffp \sqrt{\alpha}}$ to $\frac{1}{\alpha^\rho} + \frac{d}{\diffp} \log\paren{\frac{1}{\alpha}}$ for any fixed $\rho >0$, while retaining the standard minimax-optimal sample complexity for non-interpolation problems. We prove a lower bound that shows the dimension-dependent term in the expression above is tight. Furthermore, we provide a superefficiency result which demonstrates the necessity of the polynomial term for adaptive algorithms: any algorithm that has a polylogarithmic sample complexity for interpolation problems cannot achieve the minimax-optimal rates for the family of non-interpolation problems.

Optimal Algorithms for Mean Estimation under Local Differential Privacy

- Hilal Asi, Vitaly Feldman, Kunal Talwar
- abstract: We study the problem of mean estimation of ℓ_2-bounded vectors under the constraint of local differential privacy. While the literature has a variety of algorithms that achieve the (asymptotic) optimal rates for this problem, the performance of these algorithms in practice can vary significantly due to varying (and often large) hidden constants. In this work, we investigate the question of designing the randomizer with the smallest variance. We show that PrivUnit (Bhowmick et al. 2018) with optimized parameters achieves the optimal variance among a large family of natural randomizers. To prove this result, we establish some properties of local randomizers, and use symmetrization arguments that allow us to write the optimal randomizer as the optimizer of a certain linear program. These structural results, which should extend to other problems, then allow us to show that the optimal randomizer belongs to the PrivUnit family. We also develop a new variant of PrivUnit based on the Gaussian distribution which is more amenable to mathematical analysis and enjoys the same optimality guarantees. This allows us to establish several useful properties on the exact constants of the optimal error as well as to numerically estimate these constants.

Asymptotically-Optimal Gaussian Bandits with Side Observations

- Alexia Atsidakou, Orestis Papadigenopoulos, Constantine Caramanis, Sujay Sanghavi, Sanjay Shakkottai
- abstract: We study the problem of Gaussian bandits with general side information, as first introduced by Wu, Szepesvári, and György. In this setting, the play of an arm reveals information about other arms, according to an arbitrary a priori known side information matrix: each element of this matrix encodes the fidelity of the information that the “row" arm reveals about the “column" arm. In the case of Gaussian noise, this model subsumes standard bandits, full-feedback, and graph-structured feedback as special cases. In this work, we first construct an LP-based asymptotic instance-dependent lower bound on the regret. The LP optimizes the cost (regret) required to reliably estimate the suboptimality gap of each arm. This LP lower bound motivates our main contribution: the first known asymptotically optimal algorithm for this general setting.

Congested Bandits: Optimal Routing via Short-term Resets

- Pranjal Awasthi, Kush Bhatia, Sreenivas Gollapudi, Kostas Kollias
- abstract: For traffic routing platforms, the choice of which route to recommend to a user depends on the congestion on these routes – indeed, an individual’s utility depends on the number of people using the recommended route at that instance. Motivated by this, we introduce the problem of Congested Bandits where each arm’s reward is allowed to depend on the number of times it was played in the past Δ timesteps. This dependence on past history of actions leads to a dynamical system where an algorithm’s present choices also affect its future pay-offs, and requires an algorithm to plan for this. We study the congestion aware formulation in the multi-armed bandit (MAB) setup and in the contextual bandit setup with linear rewards. For the multi-armed setup, we propose a UCB style algorithm and show that its policy regret scales as $\tilde{O}(\sqrt{K \Delta T})$. For the linear contextual bandit setup, our algorithm, based on an iterative least squares planner, achieves policy regret $\tilde{O}(\sqrt{dT} + \Delta)$. From an experimental standpoint, we corroborate the no-regret properties of our algorithms via a simulation study.

Do More Negative Samples Necessarily Hurt In Contrastive Learning?

- Pranjal Awasthi, Nishanth Dikkala, Pritish Kamath
- abstract: Recent investigations in noise contrastive estimation suggest, both empirically as well as theoretically, that while having more “negative samples” in the contrastive loss improves downstream classification performance initially, beyond a threshold, it hurts downstream performance due to a “collision-coverage” trade-off. But is such a phenomenon inherent in contrastive learning? We show in a simple theoretical setting, where positive pairs are generated by sampling from the underlying latent class (introduced by Saunshi et al. (ICML 2019)), that the downstream performance of the representation optimizing the (population) contrastive loss in fact does not degrade with the number of negative samples. Along the way, we give a structural characterization of the optimal representation in our framework, for noise contrastive estimation. We also provide empirical support for our theoretical results on CIFAR-10 and CIFAR-100 datasets.

H-Consistency Bounds for Surrogate Loss Minimizers

- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, Yutao Zhong
- abstract: We present a detailed study of estimation errors in terms of surrogate loss estimation errors. We refer to such guarantees as H-consistency bounds, since they account for the hypothesis set H adopted. These guarantees are significantly stronger than H-calibration or H-consistency. They are also more informative than similar excess error bounds derived in the literature, when H is the family of all measurable functions. We prove general theorems providing such guarantees, for both the distribution-dependent and distribution-independent settings. We show that our bounds are tight, modulo a convexity assumption. We also show that previous excess error bounds can be recovered as special cases of our general results. We then present a series of explicit bounds in the case of the zero-one loss, with multiple choices of the surrogate loss and for both the family of linear functions and neural networks with one hidden-layer. We further prove more favorable distribution-dependent guarantees in that case. We also present a series of explicit bounds in the case of the adversarial loss, with surrogate losses based on the supremum of the ρ-margin, hinge or sigmoid loss and for the same two general hypothesis sets. Here too, we prove several enhancements of these guarantees under natural distributional assumptions. Finally, we report the results of simulations illustrating our bounds and their tightness.

Iterative Hard Thresholding with Adaptive Regularization: Sparser Solutions Without Sacrificing Runtime

- Kyriakos Axiotis, Maxim Sviridenko
- abstract: We propose a simple modification to the iterative hard thresholding (IHT) algorithm, which recovers asymptotically sparser solutions as a function of the condition number. When aiming to minimize a convex function f(x) with condition number κ subject to x being an s-sparse vector, the standard IHT guarantee is a solution with relaxed sparsity $O(s\kappa^2)$, while our proposed algorithm, regularized IHT, returns a solution with sparsity $O(s\kappa)$. Our algorithm significantly improves over ARHT [Axiotis & Sviridenko, 2021] which also achieves $O(s\kappa)$, as it does not

require re-optimization in each iteration (and so is much faster), is deterministic, and does not require knowledge of the optimal solution value f(x*) or the optimal sparsity level s. Our main technical tool is an adaptive regularization framework, in which the algorithm progressively learns the weights of an l_2 regularization term that will allow convergence to sparser solutions. We also apply this framework to low rank optimization, where we achieve a similar improvement of the best known condition number dependence from κ^2 to κ.

Proving Theorems using Incremental Learning and Hindsight Experience Replay

- Eser Aygün, Ankit Anand, Laurent Orseau, Xavier Glorot, Stephen M Mcaleer, Vlad Firoiu, Lei M Zhang, Doina Precup, Shibl Mourad
- abstract: Traditional automated theorem proving systems for first-order logic depend on speed-optimized search and many handcrafted heuristics designed to work over a wide range of domains. Machine learning approaches in the literature either depend on these traditional provers to bootstrap themselves, by leveraging these heuristics, or can struggle due to limited existing proof data. The latter issue can be explained by the lack of a smooth difficulty gradient in theorem proving datasets; large gaps in difficulty between different theorems can make training harder or even impossible. In this paper, we adapt the idea of hindsight experience replay from reinforcement learning to the automated theorem proving domain, so as to use the intermediate data generated during unsuccessful proof attempts. We build a first-order logic prover by disabling all the smart clause-scoring heuristics of the state-of-the-art E prover and replacing them with a clause-scoring neural network learned by using hindsight experience replay in an incremental learning setting. Clauses are represented as graphs and presented to transformer networks with spectral features. We show that provers trained in this way can outperform previous machine learning approaches and compete with the state of the art heuristic-based theorem prover E in its best configuration, on the popular benchmarks MPTP2078, M2k and Mizar40. The proofs generated by our algorithm are also almost always significantly shorter than E’s proofs.

Near-optimal rate of consistency for linear models with missing values

- Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, Erwan Scornet
- abstract: Missing values arise in most real-world data sets due to the aggregation of multiple sources and intrinsically missing information (sensor failure, unanswered questions in surveys...). In fact, the very nature of missing values usually prevents us from running standard learning algorithms. In this paper, we focus on the extensively-studied linear models, but in presence of missing values, which turns out to be quite a challenging task. Indeed, the Bayes predictor can be decomposed as a sum of predictors corresponding to each missing pattern. This eventually requires to solve a number of learning tasks, exponential in the number of input features, which makes predictions impossible for current real-world datasets. First, we propose a rigorous setting to analyze a least-square type estimator and establish a bound on the excess risk which increases exponentially in the dimension. Consequently, we leverage the missing data distribution to propose a new algorithm, and derive associated adaptive risk bounds that turn out to be minimax optimal. Numerical experiments highlight the benefits of our method compared to state-of-the-art algorithms used for predictions with missing values.

How Tempering Fixes Data Augmentation in Bayesian Neural Networks

- Gregor Bachmann, Lorenzo Noci, Thomas Hofmann
- abstract: While Bayesian neural networks (BNNs) provide a sound and principled alternative to standard neural networks, an artificial sharpening of the posterior usually needs to be applied to reach comparable performance. This is in stark contrast to theory, dictating that given an adequate prior and a well-specified model, the untempered Bayesian posterior should achieve optimal performance. Despite the community’s extensive efforts, the observed gains in performance still remain disputed with several plausible causes pointing at its origin. While data augmentation has been empirically recognized as one of the main drivers of this effect, a theoretical account of its role, on the other hand, is largely missing. In this work we identify two interlaced factors concurrently influencing the strength of the cold posterior effect, namely the correlated nature of augmentations and the degree of invariance of the employed model to such transformations. By theoretically analyzing simplified settings, we prove that tempering implicitly reduces the misspecification arising from modeling augmentations as i.i.d. data. The temperature mimics the role of the effective sample size, reflecting the gain in information provided by the augmentations. We corroborate our theoretical findings with extensive empirical evaluations, scaling to realistic BNNs. By relying on the framework of group convolutions, we experiment with models of varying inherent degree of invariance, confirming its hypothesized relationship with the optimal temperature.

ASAP.SGD: Instance-based Adaptiveness to Staleness in Asynchronous SGD

- Karl Bäckström, Marina Papatriantafilou, Philippas Tsigas
- abstract: Concurrent algorithmic implementations of Stochastic Gradient Descent (SGD) give rise to critical questions for compute-intensive Machine Learning (ML). Asynchrony implies speedup in some contexts, and challenges in others, as stale updates may lead to slower, or non-converging executions. While previous works showed asynchrony-adaptiveness can improve stability and speedup by reducing the step size for stale updates according to static rules, there is no one-size-fits-all adaptation rule, since the optimal strategy depends on several factors. We introduce (i) $\mathtt{ASAP.SGD}$, an analytical framework capturing necessary and desired properties of staleness-adaptive step size functions and (ii) \textsc{tail}-τ, a method for utilizing key properties of the execution instance, generating a tailored strategy that not only dampens the impact of stale updates, but also leverages fresh ones. We recover convergence bounds for adaptiveness functions satisfying the $\mathtt{ASAP.SGD}$ conditions for general, convex and non-convex problems, and establish novel bounds for ones satisfying the Polyak-Lojasiewicz property. We evaluate \textsc{tail}-τ with representative AsyncSGD concurrent algorithms, for Deep Learning problems, showing \textsc{tail}-τ is a vital complement to AsyncSGD, with (i) persistent speedup in wall-clock convergence time in the parallelism spectrum, (ii) considerably lower risk of non-convergence, as well as (iii) precision levels for which original SGD implementations fail.

From Noisy Prediction to True Label: Noisy Prediction Calibration via Generative Model

- Heesun Bae, Seungjae Shin, Byeonghu Na, Joonho Jang, Kyungwoo Song, Il-Chul Moon
- abstract: Noisy labels are inevitable yet problematic in machine learning society. It ruins the generalization of a classifier by making the classifier over-fitted to noisy labels. Existing methods on noisy label have focused on modifying the classifier during the training procedure. It has two potential problems. First, these methods are not applicable to a pre-trained classifier without further access to training. Second, it is not easy to train a classifier and regularize all negative effects from noisy labels, simultaneously. We suggest a new branch of method, Noisy Prediction Calibration (NPC) in learning with noisy labels. Through the introduction and estimation of a new type of transition matrix via generative model, NPC corrects the noisy prediction from the pre-trained classifier to the true label as a post-processing scheme. We prove that NPC theoretically aligns with the transition matrix based methods. Yet, NPC empirically provides more accurate pathway to estimate true label, even without involvement in classifier learning. Also, NPC is applicable to any classifier trained with noisy label methods, if training instances and its predictions are available. Our method, NPC, boosts the classification performances of all baseline models on both synthetic and real-world datasets. The implemented code is available at https://github.com/BaeHeeSun/NPC.

data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language

- Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, Michael Auli
- abstract: While the general idea of self-supervised learning is identical across modalities, the actual algorithms and objectives differ widely because they were developed with a single modality in mind. To get us closer to general self-supervised learning, we present data2vec, a framework that uses the same learning method for either speech, NLP or computer vision. The core idea is to predict latent representations of the full input data based on a masked view of the input in a self-distillation setup using a standard Transformer architecture. Instead of predicting modality-specific targets such as words, visual tokens or units of human speech which are local in nature, data2vec predicts contextualized latent representations that contain information from the entire

input. Experiments on the major benchmarks of speech recognition, image classification, and natural language understanding demonstrate a new state of the art or competitive performance to predominant approaches.

End-to-End Balancing for Causal Continuous Treatment-Effect Estimation

- Taha Bahadori, Eric Tchetgen Tchetgen, David Heckerman
- abstract: We study the problem of observational causal inference with continuous treatment. We focus on the challenge of estimating the causal response curve for infrequently-observed treatment values. We design a new algorithm based on the framework of entropy balancing which learns weights that directly maximize causal inference accuracy using end-to-end optimization. Our weights can be customized for different datasets and causal inference algorithms. We propose a new theory for consistency of entropy balancing for continuous treatments. Using synthetic and real-world data, we show that our proposed algorithm outperforms the entropy balancing in terms of causal inference accuracy.

A Hierarchical Transitive-Aligned Graph Kernel for Un-attributed Graphs

- Lu Bai, Lixin Cui, Hancock Edwin
- abstract: In this paper, we develop a new graph kernel, namely the Hierarchical Transitive-Aligned Kernel, by transitively aligning the vertices between graphs through a family of hierarchical prototype graphs. Comparing to most existing state-of-the-art graph kernels, the proposed kernel has three theoretical advantages. First, it incorporates the locational correspondence information between graphs into the kernel computation, and thus overcomes the shortcoming of ignoring structural correspondences arising in most R-convolution kernels. Second, it guarantees the transitivity between the correspondence information that is not available for most existing matching kernels. Third, it incorporates the information of all graphs under comparisons into the kernel computation process, and thus encapsulates richer characteristics. Experimental evaluations demonstrate the effectiveness of the new transitive-aligned kernel.

Near-Optimal Learning of Extensive-Form Games with Imperfect Information

- Yu Bai, Chi Jin, Song Mei, Tiancheng Yu
- abstract: This paper resolves the open question of designing near-optimal algorithms for learning imperfect-information extensive-form games from bandit feedback. We present the first line of algorithms that require only $\widetilde{\mathcal{O}}((XA+YB)/\varepsilon^2)$ episodes of play to find an ε-approximate Nash equilibrium in two-player zero-sum games, where X,Y are the number of information sets and A,B are the number of actions for the two players. This improves upon the best known sample complexity of $\widetilde{\mathcal{O}}((X^2A+Y^2B)/\varepsilon^2)$ by a factor of $\widetilde{\mathcal{O}}(\max{X, Y})$, and matches the information-theoretic lower bound up to logarithmic factors. We achieve this sample complexity by two new algorithms: Balanced Online Mirror Descent, and Balanced Counterfactual Regret Minimization. Both algorithms rely on novel approaches of integrating balanced exploration policies into their classical counterparts. We also extend our results to learning Coarse Correlated Equilibria in multi-player general-sum games.

Gaussian Mixture Variational Autoencoder with Contrastive Learning for Multi-Label Classification

- Junwen Bai, Shufeng Kong, Carla P Gomes
- abstract: Multi-label classification (MLC) is a prediction task where each sample can have more than one label. We propose a novel contrastive learning boosted multi-label prediction model based on a Gaussian mixture variational autoencoder (C-GMVAE), which learns a multimodal prior space and employs a contrastive loss. Many existing methods introduce extra complex neural modules like graph neural networks to capture the label correlations, in addition to the prediction modules. We find that by using contrastive learning in the supervised setting, we can exploit label information effectively in a data-driven manner, and learn meaningful feature and label embeddings which capture the label correlations and enhance the predictive power. Our method also adopts the idea of learning and aligning latent spaces for both features and labels. In contrast to previous works based on a unimodal prior, C-GMVAE imposes a Gaussian mixture structure on the latent space, to alleviate the posterior collapse and over-regularization issues. C-GMVAE outperforms existing methods on multiple public datasets and can often match other models' full performance with only 50% of the training data. Furthermore, we show that the learnt embeddings provide insights into the interpretation of label-label interactions.

A^{3}T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing

- He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, Liang Huang
- abstract: Recently, speech representation learning has improved many speech-related tasks such as speech recognition, speech classification, and speech-to-text translation. However, all the above tasks are in the direction of speech understanding, but for the inverse direction, speech synthesis, the potential of representation learning is yet to be realized, due to the challenging nature of generating high-quality speech. To address this problem, we propose our framework, Alignment-Aware Acoustic-Text Pretraining (A^{3}T), which reconstructs masked acoustic signals with text input and acoustic-text alignment during training. In this way, the pretrained model can generate high quality reconstructed spectrogram, which can be applied to the speech editing and unseen speaker TTS directly. Experiments show A^{3}T outperforms SOTA models on speech editing, and improves multi-speaker speech synthesis without the external speaker verification model.

Stability Based Generalization Bounds for Exponential Family Langevin Dynamics

- Arindam Banerjee, Tiancong Chen, Xinyan Li, Yingxue Zhou
- abstract: Recent years have seen advances in generalization bounds for noisy stochastic algorithms, especially stochastic gradient Langevin dynamics (SGLD) based on stability (Mou et al., 2018; Li et al., 2020) and information theoretic approaches (Xu & Raginsky, 2017; Negrea et al., 2019; Steinke & Zakynthinou, 2020). In this paper, we unify and substantially generalize stability based generalization bounds and make three technical contributions. First, we bound the generalization error in terms of expected (not uniform) stability which arguably leads to quantitatively sharper bounds. Second, as our main contribution, we introduce Exponential Family Langevin Dynamics (EFLD), a substantial generalization of SGLD, which includes noisy versions of Sign-SGD and quantized SGD as special cases. We establish data dependent expected stability based generalization bounds for any EFLD algorithm with a O(1/n) sample dependence and dependence on gradient discrepancy rather than the norm of gradients, yielding significantly sharper bounds. Third, we establish optimization guarantees for special cases of EFLD. Further, empirical results on benchmarks illustrate that our bounds are non-vacuous, quantitatively sharper than existing bounds, and behave correctly under noisy labels.

Certified Neural Network Watermarks with Randomized Smoothing

- Arpit Bansal, Ping-Yeh Chiang, Michael J Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John P Dickerson, Tom Goldstein
- abstract: Watermarking is a commonly used strategy to protect creators' rights to digital images, videos and audio. Recently, watermarking methods have been extended to deep learning models – in principle, the watermark should be preserved when an adversary tries to copy the model. However, in practice, watermarks can often be removed by an intelligent adversary. Several papers have proposed watermarking methods that claim to be empirically resistant to different types of removal attacks, but these new techniques often fail in the face of new or better-tuned adversaries. In this paper, we propose the first certifiable watermarking method. Using the randomized smoothing technique, we show that our watermark is guaranteed to be unremovable unless the

model parameters are changed by more than a certain ℓ_2 threshold. In addition to being certifiable, our watermark is also empirically more robust compared to previous watermarking methods.

Data Scaling Laws in NMT: The Effect of Noise and Architecture

- Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, Orhan Firat
- abstract: In this work, we study the effect of varying the architecture and training data quality on the data scaling properties of Neural Machine Translation (NMT). First, we establish that the test loss of encoder-decoder transformer models scales as a power law in the number of training samples, with a dependence on the model size. Then, we systematically vary aspects of the training setup to understand how they impact the data scaling laws. In particular, we change the following (1) Architecture and task setup: We compare to a transformer-LSTM hybrid, and a decoder-only transformer with a language modeling loss (2) Noise level in the training distribution: We experiment with filtering, and adding iid synthetic noise. In all the above cases, we find that the data scaling exponents are minimally impacted, suggesting that marginally worse architectures or training data can be compensated for by adding more data. Lastly, we find that using back-translated data instead of parallel data, can significantly degrade the scaling exponent.

Learning Stable Classifiers by Transferring Unstable Features

- Yujia Bao, Shiyu Chang, Dr.Regina Barzilay
- abstract: While unbiased machine learning models are essential for many applications, bias is a human-defined concept that can vary across tasks. Given only input-label pairs, algorithms may lack sufficient information to distinguish stable (causal) features from unstable (spurious) features. However, related tasks often share similar biases – an observation we may leverage to develop stable classifiers in the transfer setting. In this work, we explicitly inform the target classifier about unstable features in the source tasks. Specifically, we derive a representation that encodes the unstable features by contrasting different data environments in the source task. We achieve robustness by clustering data of the target task according to this representation and minimizing the worst-case risk across these clusters. We evaluate our method on both text and image classifications. Empirical results demonstrate that our algorithm is able to maintain robustness on the target task for both synthetically generated environments and real-world environments. Our code is available at https://github.com/YujiaBao/Tofu.

Fast Composite Optimization and Statistical Recovery in Federated Learning

- Yajie Bao, Michael Crawshaw, Shan Luo, Mingrui Liu
- abstract: As a prevalent distributed learning paradigm, Federated Learning (FL) trains a global model on a massive amount of devices with infrequent communication. This paper investigates a class of composite optimization and statistical recovery problems in the FL setting, whose loss function consists of a data-dependent smooth loss and a non-smooth regularizer. Examples include sparse linear regression using Lasso, low-rank matrix recovery using nuclear norm regularization, etc. In the existing literature, federated composite optimization algorithms are designed only from an optimization perspective without any statistical guarantees. In addition, they do not consider commonly used (restricted) strong convexity in statistical recovery problems. We advance the frontiers of this problem from both optimization and statistical perspectives. From optimization upfront, we propose a new algorithm named Fast Federated Dual Averaging for strongly convex and smooth loss and establish state-of-the-art iteration and communication complexity in the composite setting. In particular, we prove that it enjoys a fast rate, linear speedup, and reduced communication rounds. From statistical upfront, for restricted strongly convex and smooth loss, we design another algorithm, namely Multi-stage Federated Dual Averaging, and prove a high probability complexity bound with linear speedup up to optimal statistical precision. Numerical experiments in both synthetic and real data demonstrate that our methods perform better than other baselines. To the best of our knowledge, this is the first work providing fast optimization algorithms and statistical recovery guarantees for composite problems in FL.

Generative Modeling for Multi-task Visual Learning

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

Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models

- Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, Bo Zhang
- abstract: Diffusion probabilistic models (DPMs) are a class of powerful deep generative models (DGMs). Despite their success, the iterative generation process over the full timesteps is much less efficient than other DGMs such as GANs. Thus, the generation performance on a subset of timesteps is crucial, which is greatly influenced by the covariance design in DPMs. In this work, we consider diagonal and full covariances to improve the expressive power of DPMs. We derive the optimal result for such covariances, and then correct it when the mean of DPMs is imperfect. Both the optimal and the corrected ones can be decomposed into terms of conditional expectations over functions of noise. Building upon it, we propose to estimate the optimal covariance and its correction given imperfect mean by learning these conditional expectations. Our method can be applied to DPMs with both discrete and continuous timesteps. We consider the diagonal covariance in our implementation for computational efficiency. For an efficient practical implementation, we adopt a parameter sharing scheme and a two-stage training process. Empirically, our method outperforms a wide variety of covariance design on likelihood results, and improves the sample quality especially on a small number of timesteps.

On the Surrogate Gap between Contrastive and Supervised Losses

- Han Bao, Yoshihiro Nagano, Kento Nozawa
- abstract: Contrastive representation learning encourages data representation to make semantically similar pairs closer than randomly drawn negative samples, which has been successful in various domains such as vision, language, and graphs. Recent theoretical studies have attempted to explain the benefit of the large negative sample size by upper-bounding the downstream classification loss with the contrastive loss. However, the previous surrogate bounds have two drawbacks: they are only legitimate for a limited range of negative sample sizes and prohibitively large even within that range. Due to these drawbacks, there still does not exist a consensus on how negative sample size theoretically correlates with downstream classification performance. Following the simplified setting where positive pairs are drawn from the true distribution (not generated by data augmentation; as supposed in previous studies), this study establishes surrogate upper and lower bounds for the downstream classification loss for all negative sample sizes that best explain the empirical observations on the negative sample size in the earlier studies. Our bounds suggest that the contrastive loss can be viewed as a surrogate objective of the downstream loss and larger negative sample sizes improve downstream classification because the surrogate gap between contrastive and supervised losses decays. We verify that our theory is consistent with experiments on synthetic, vision, and language datasets.

Representation Topology Divergence: A Method for Comparing Neural Network Representations.

- Serguei Barannikov, Ilya Trofimov, Nikita Balabin, Evgeny Burnaev
- abstract: Comparison of data representations is a complex multi-aspect problem. We propose a method for comparing two data representations. We introduce the Representation Topology Divergence (RTD) score measuring the dissimilarity in multi-scale topology between two point clouds of equal size with a one-to-one correspondence between points. The two data point clouds can lie in different ambient spaces. The RTD score is one of the few topological data analysis based practical methods applicable to real machine learning datasets. Experiments show the agreement of RTD with the intuitive assessment of data representation similarity. The proposed RTD score is sensitive to the data representation's fine topological structure. We use the RTD score to gain insights on neural networks representations in computer vision and NLP domains for various problems: training dynamics analysis, data distribution shift, transfer learning, ensemble learning, disentanglement assessment.

Sparse Mixed Linear Regression with Guarantees: Taming an Intractable Problem with Invex Relaxation

- Adarsh Barik, Jean Honorio
- abstract: In this paper, we study the problem of sparse mixed linear regression on an unlabeled dataset that is generated from linear measurements from two different regression parameter vectors. Since the data is unlabeled, our task is to not only figure out a good approximation of regression parameter vectors but also label the dataset correctly. In its original form, this problem is NP-hard. The most popular algorithms to solve this problem (such as Expectation-Maximization) have a tendency to stuck at local minima. We provide a novel invex relaxation for this intractable problem which leads to a solution with provable theoretical guarantees. This relaxation enables exact recovery of data labels. Furthermore, we recover close approximation of regression parameter vectors which match the true parameter vectors in support and sign. Our formulation uses a carefully constructed primal dual witnesses framework for the invex problem. Furthermore, we show that the sample complexity of our method is only logarithmic in terms of the dimension of the regression parameter vectors.

Neural Fisher Discriminant Analysis: Optimal Neural Network Embeddings in Polynomial Time

- Burak Bartan, Mert Pilanci
- abstract: Fisher's Linear Discriminant Analysis (FLDA) is a statistical analysis method that linearly embeds data points to a lower dimensional space to maximize a discrimination criterion such that the variance between classes is maximized while the variance within classes is minimized. We introduce a natural extension of FLDA that employs neural networks, called Neural Fisher Discriminant Analysis (NFDA). This method finds the optimal two-layer neural network that embeds data points to optimize the same discrimination criterion. We use tools from convex optimization to transform the optimal neural network embedding problem into a convex problem. The resulting problem is easy to interpret and solve to global optimality. We evaluate the method's performance on synthetic and real datasets.

Fictitious Play and Best-Response Dynamics in Identical Interest and Zero-Sum Stochastic Games

- Lucas Baudin, Rida Laraki
- abstract: This paper proposes an extension of a popular decentralized discrete-time learning procedure when repeating a static game called fictitious play (FP) (Brown, 1951; Robinson, 1951) to a dynamic model called discounted stochastic game (Shapley, 1953). Our family of discrete-time FP procedures is proven to converge to the set of stationary Nash equilibria in identical interest discounted stochastic games. This extends similar convergence results for static games (Monderer & Shapley, 1996a). We then analyze the continuous-time counterpart of our FP procedures, which include as a particular case the best-response dynamic introduced and studied by Leslie et al. (2020) in the context of zero-sum stochastic games. We prove the converge of this dynamics to stationary Nash equilibria in identical-interest and zero-sum discounted stochastic games. Thanks to stochastic approximations, we can infer from the continuous-time convergence some discrete time results such as the convergence to stationary equilibria in zero-sum and team stochastic games (Holler, 2020).

Information Discrepancy in Strategic Learning

- Yahav Bechavod, Chara Podimata, Steven Wu, Juba Ziani
- abstract: We initiate the study of the effects of non-transparency in decision rules on individuals' ability to improve in strategic learning settings. Inspired by real-life settings, such as loan approvals and college admissions, we remove the assumption typically made in the strategic learning literature, that the decision rule is fully known to individuals, and focus instead on settings where it is inaccessible. In their lack of knowledge, individuals try to infer this rule by learning from their peers (e.g., friends and acquaintances who previously applied for a loan), naturally forming groups in the population, each with possibly different type and level of information regarding the decision rule. We show that, in equilibrium, the principal's decision rule optimizing welfare across sub-populations may cause a strong negative externality: the true quality of some of the groups can actually deteriorate. On the positive side, we show that, in many natural cases, optimal improvement can be guaranteed simultaneously for all sub-populations. We further introduce a measure we term information overlap proxy, and demonstrate its usefulness in characterizing the disparity in improvements across sub-populations. Finally, we identify a natural condition under which improvement can be guaranteed for all sub-populations while maintaining high predictive accuracy. We complement our theoretical analysis with experiments on real-world datasets.

On the Hidden Biases of Policy Mirror Ascent in Continuous Action Spaces

- Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M Sadler, Pratap Tokekar, Alec Koppel
- abstract: We focus on parameterized policy search for reinforcement learning over continuous action spaces. Typically, one assumes the score function associated with a policy is bounded, which {fails to hold even for Gaussian policies. } To properly address this issue, one must introduce an exploration tolerance parameter to quantify the region in which it is bounded. Doing so incurs a persistent bias that appears in the attenuation rate of the expected policy gradient norm, which is inversely proportional to the radius of the action space. To mitigate this hidden bias, heavy-tailed policy parameterizations may be used, which exhibit a bounded score function, but doing so can cause instability in algorithmic updates. To address these issues, in this work, we study the convergence of policy gradient algorithms under heavy-tailed parameterizations, which we propose to stabilize with a combination of mirror ascent-type updates and gradient tracking. Our main theoretical contribution is the establishment that this scheme converges with constant batch sizes, whereas prior works require these parameters to respectively shrink to null or grow to infinity. Experimentally, this scheme under a heavy-tailed policy parameterization yields improved reward accumulation across a variety of settings as compared with standard benchmarks.

Imitation Learning by Estimating Expertise of Demonstrators

- Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, Ramtin Pedarsani
- abstract: Many existing imitation learning datasets are collected from multiple demonstrators, each with different expertise at different parts of the environment. Yet, standard imitation learning algorithms typically treat all demonstrators as homogeneous, regardless of their expertise, absorbing the weaknesses of any suboptimal demonstrators. In this work, we show that unsupervised learning over demonstrator expertise can lead to a consistent boost in the performance of imitation learning algorithms. We develop and optimize a joint model over a learned policy and expertise levels of the demonstrators. This enables our model to learn from the optimal behavior and filter out the suboptimal behavior of each demonstrator. Our model learns a single policy that can outperform even the best demonstrator, and can be used to estimate the expertise of any demonstrator at any state. We illustrate our

findings on real-robotic continuous control tasks from Robomimic and discrete environments such as MiniGrid and chess, out-performing competing methods in 21 out of 23 settings, with an average of 7% and up to 60% improvement in terms of the final reward.

Matching Normalizing Flows and Probability Paths on Manifolds

- Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Maximillian Nickel, Aditya Grover, Ricky T. Q. Chen, Yaron Lipman
- abstract: Continuous Normalizing Flows (CNFs) are a class of generative models that transform a prior distribution to a model distribution by solving an ordinary differential equation (ODE). We propose to train CNFs on manifolds by minimizing probability path divergence (PPD), a novel family of divergences between the probability density path generated by the CNF and a target probability density path. PPD is formulated using a logarithmic mass conservation formula which is a linear first order partial differential equation relating the log target probabilities and the CNF's defining vector field. PPD has several key benefits over existing methods: it sidesteps the need to solve an ODE per iteration, readily applies to manifold data, scales to high dimensions, and is compatible with a large family of target paths interpolating pure noise and data in finite time. Theoretically, PPD is shown to bound classical probability divergences. Empirically, we show that CNFs learned by minimizing PPD achieve state-of-the-art results in likelihoods and sample quality on existing low-dimensional manifold benchmarks, and is the first example of a generative model to scale to moderately high dimensional manifolds.

Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models

- Viktor Bengs, Aadirupa Saha, Eyke Hüllermeier
- abstract: We consider the regret minimization task in a dueling bandits problem with context information. In every round of the sequential decision problem, the learner makes a context-dependent selection of two choice alternatives (arms) to be compared with each other and receives feedback in the form of noisy preference information. We assume that the feedback process is determined by a linear stochastic transitivity model with contextualized utilities (CoLST), and the learner's task is to include the best arm (with highest latent context-dependent utility) in the duel. We propose a computationally efficient algorithm, \Algo{CoLSTIM}, which makes its choice based on imitating the feedback process using perturbed context-dependent utility estimates of the underlying CoLST model. If each arm is associated with a d-dimensional feature vector, we show that \Algo{CoLSTIM} achieves a regret of order $\tilde O(\sqrt{dT})$ after T learning rounds. Additionally, we also establish the optimality of \Algo{CoLSTIM} by showing a lower bound for the weak regret that refines the existing average regret analysis. Our experiments demonstrate its superiority over state-of-art algorithms for special cases of CoLST models.

Neural Inverse Kinematic

- Raphael Bensadoun, Shir Gur, Nitsan Blau, Lior Wolf
- abstract: Inverse kinematic (IK) methods recover the parameters of the joints, given the desired position of selected elements in the kinematic chain. While the problem is well-defined and low-dimensional, it has to be solved rapidly, accounting for multiple possible solutions. In this work, we propose a neural IK method that employs the hierarchical structure of the problem to sequentially sample valid joint angles conditioned on the desired position and on the preceding joints along the chain. In our solution, a hypernetwork f recovers the parameters of multiple primary networks {$g_1,g_2,…,g_N$, where N is the number of joints}, such that each g_i outputs a distribution of possible joint angles, and is conditioned on the sampled values obtained from the previous primary networks $g_j, j Cite this Paper

BibTeX

@InProceedings{pmlr-v162-bensadoun22a, title = {Neural Inverse Kinematic}, author = {Bensadoun, Raphael and Gur, Shir and Blau, Nitsan and Wolf, Lior}, booktitle = {Proceedings of the 39th International Conference on Machine Learning}, pages = {1787--1797}, year = {2022}, editor = {Chaudhuri, Kamalika and Jegelka, Stefanie and Song, Le and Szepesvari, Csaba and Niu, Gang and Sabato, Sivan}, volume = {162}, series = {Proceedings of Machine Learning Research}, month = {17--23 Jul}, publisher = {PMLR}, pdf = {https://proceedings.mlr.press/v162/bensadoun22a/bensadoun22a.pdf}, url = {https://proceedings.mlr.press/v162/bensadoun22a.html}, abstract = {Inverse kinematic (IK) methods recover the parameters of the joints, given the desired position of selected elements in the kinematic chain. While the problem is well-defined and low-dimensional, it has to be solved rapidly, accounting for multiple possible solutions. In this work, we propose a neural IK method that employs the hierarchical structure of the problem to sequentially sample valid joint angles conditioned on the desired position and on the preceding joints along the chain. In our solution, a hypernetwork f recovers the parameters of multiple primary networks {$g_1,g_2,…,g_N$, where N is the number of joints}, such that each g_i outputs a distribution of possible joint angles, and is conditioned on the sampled values obtained from the previous primary networks $g_j, j Copy to ClipboardDownload Endnote %0 Conference Paper %T Neural Inverse Kinematic %A Raphael Bensadoun %A Shir Gur %A Nitsan Blau %A Lior Wolf %B Proceedings of the 39th International Conference on Machine Learning %C Proceedings of Machine Learning Research %D 2022 %E Kamalika Chaudhuri %E Stefanie Jegelka %E Le Song %E Csaba Szepesvari %E Gang Niu %E Sivan Sabato %F pmlr-v162-bensadoun22a %I PMLR %P 1787--1797 %U https://proceedings.mlr.press/v162/bensadoun22a.html %V 162 %X Inverse kinematic (IK) methods recover the parameters of the joints, given the desired position of selected elements in the kinematic chain. While the problem is well-defined and low-dimensional, it has to be solved rapidly, accounting for multiple possible solutions. In this work, we propose a neural IK method that employs the hierarchical structure of the problem to sequentially sample valid joint angles conditioned on the desired position and on the preceding joints along the chain. In our solution, a hypernetwork f recovers the parameters of multiple primary networks {$g_1,g_2,…,g_N$, where N is the number of joints}, such that each g_i outputs a distribution of possible joint angles, and is conditioned on the sampled values obtained from the previous primary networks $g_j, j Copy to ClipboardDownload APA

Bensadoun, R., Gur, S., Blau, N. & Wolf, L.. (2022). Neural Inverse Kinematic. Proceedings of the 39th International Conference on Machine Learning, in Proceedings of Machine Learning Research 162:1787-1797 Available from https://proceedings.mlr.press/v162/bensadoun22a.html.

Copy to ClipboardDownload

Related Material

Download PDF

Volatility Based Kernels and Moving Average Means for Accurate Forecasting with Gaussian Processes

- Gregory Benton, Wesley Maddox, Andrew Gordon Wilson
- abstract: A broad class of stochastic volatility models are defined by systems of stochastic differential equations, and while these models have seen widespread success in domains such as finance and statistical climatology, they typically lack an ability to condition on historical data to produce a true posterior distribution. To address this fundamental limitation, we show how to re-cast a class of stochastic volatility models as a hierarchical Gaussian process (GP) model with specialized covariance functions. This GP model retains the inductive biases of the stochastic volatility model while providing the posterior predictive distribution given by GP inference. Within this framework, we take inspiration from well studied domains to introduce a new class of models, Volt and Magpie, that significantly outperform baselines in stock and wind speed forecasting, and naturally extend to the multitask setting.

Gradient Descent on Neurons and its Link to Approximate Second-order Optimization

- Frederik Benzing

- abstract: Second-order optimizers are thought to hold the potential to speed up neural network training, but due to the enormous size of the curvature matrix, they typically require approximations to be computationally tractable. The most successful family of approximations are Kronecker-Factored, block-diagonal curvature estimates (KFAC). Here, we combine tools from prior work to evaluate exact second-order updates with careful ablations to establish a surprising result: Due to its approximations, KFAC is not closely related to second-order updates, and in particular, it significantly outperforms true second-order updates. This challenges widely held believes and immediately raises the question why KFAC performs so well. Towards answering this question we present evidence strongly suggesting that KFAC approximates a first-order algorithm, which performs gradient descent on neurons rather than weights. Finally, we show that this optimizer often improves over KFAC in terms of computational cost and data-efficiency.

Safe Learning in Tree-Form Sequential Decision Making: Handling Hard and Soft Constraints

- Martino Bernasconi, Federico Cacciamani, Matteo Castiglioni, Alberto Marchesi, Nicola Gatti, Francesco Trovò
- abstract: We study decision making problems in which an agent sequentially interacts with a stochastic environment defined by means of a tree structure. The agent repeatedly faces the environment over time, and, after each round, it perceives a utility and a cost, which are both stochastic. The goal of the agent is to learn an optimal strategy in an online fashion, while, at the same time, keeping costs below a given safety threshold. Our model naturally fits many real-world scenarios, such as, e.g., opponent exploitation in games and web link selection. We study the hard-threshold problem of achieving sublinear regret while guaranteeing that the threshold constraint is satisfied at every iteration with high probability. First, we show that, in general, any algorithm with such a guarantee incurs in a linear regret. This motivates the introduction of a relaxed problem, namely the soft-threshold problem, in which we only require that the cumulative violation of the threshold constraint grows sublinearly, and, thus, we can provide an algorithm with sublinear regret. Next, we show how, in the hard-threshold problem, a sublinear regret algorithm can be designed under the additional assumption that there exists a known strategy strictly satisfying the threshold constraint. We also show that our regret bounds are tight. Finally, we cast the opponent exploitation problem to our model, and we experimentally evaluate our algorithms on a standard testbed of games.

Skin Deep Unlearning: Artefact and Instrument Debiasing in the Context of Melanoma Classification

- Peter Bevan, Amir Atapour-Abarghouei
- abstract: Convolutional Neural Networks have demonstrated dermatologist-level performance in the classification of melanoma from skin lesion images, but prediction irregularities due to biases seen within the training data are an issue that should be addressed before widespread deployment is possible. In this work, we robustly remove bias and spurious variation from an automated melanoma classification pipeline using two leading bias unlearning techniques. We show that the biases introduced by surgical markings and rulers presented in previous studies can be reasonably mitigated using these bias removal methods. We also demonstrate the generalisation benefits of unlearning spurious variation relating to the imaging instrument used to capture lesion images. Our experimental results provide evidence that the effects of each of the aforementioned biases are notably reduced, with different debiasing techniques excelling at different tasks.

Approximate Bayesian Computation with Domain Expert in the Loop

- Ayush Bharti, Louis Filstroff, Samuel Kaski
- abstract: Approximate Bayesian computation (ABC) is a popular likelihood-free inference method for models with intractable likelihood functions. As ABC methods usually rely on comparing summary statistics of observed and simulated data, the choice of the statistics is crucial. This choice involves a trade-off between loss of information and dimensionality reduction, and is often determined based on domain knowledge. However, handcrafting and selecting suitable statistics is a laborious task involving multiple trial-and-error steps. In this work, we introduce an active learning method for ABC statistics selection which reduces the domain expert's work considerably. By involving the experts, we are able to handle misspecified models, unlike the existing dimension reduction methods. Moreover, empirical results show better posterior estimates than with existing methods, when the simulation budget is limited.

Minimax M-estimation under Adversarial Contamination

- Sujay Bhatt, Guanhua Fang, Ping Li, Gennady Samorodnitsky
- abstract: We present a new finite-sample analysis of Catoni's M-estimator under adversarial contamination, where an adversary is allowed to corrupt a fraction of the samples arbitrarily. We make minimal assumptions on the distribution of the uncontaminated random variables, namely, we only assume the existence of a known upper bound $\upsilon_{\varepsilon} > 0$ on the $(1+\varepsilon)^{th}$ central moment of the random variables, namely, for $\varepsilon \in (0,1]$ [\mathbb{E}*{X_1 \sim \mathcal{D}} \Big| X_1 - \mu \Big|^{1+\varepsilon} \leq \upsilon*{\varepsilon}.]{We} provide a lower bound on the minimax error rate for the mean estimation problem under adversarial corruption under this weak assumption, and establish that the proposed M-estimator achieves this lower bound (up to multiplicative constants). When the variance is infinite, the tolerance to contamination of any estimator reduces as $\varepsilon \downarrow 0$. We establish a tight upper bound that characterizes this bargain. To illustrate the usefulness of the derived robust M-estimator in an online setting, we present a bandit algorithm for the partially identifiable best arm identification problem that improves upon the sample complexity of the state of the art algorithms.

Nearly Optimal Catoni's M-estimator for Infinite Variance

- Sujay Bhatt, Guanhua Fang, Ping Li, Gennady Samorodnitsky
- abstract: In this paper, we extend the remarkable M-estimator of Catoni \citep{Cat12} to situations where the variance is infinite. In particular, given a sequence of i.i.d random variables ${X_i}*{i=1}^n$ *from distribution $\mathcal{D}$ over $\mathbb{R}$ with mean μ, we only assume the existence of a known upper bound $\upsilon*{\varepsilon} > 0$* on the $(1+\varepsilon)^{th}$ central moment of the random variables, namely, for $\varepsilon \in (0,1]$ [\mathbb{E}*{X_1 \sim \mathcal{D}} \Big| X_1 - \mu \Big|^{1+\varepsilon} \leq \upsilon*{\varepsilon}.]{The} extension is non-trivial owing to the difficulty in characterizing the roots of certain polynomials of degree smaller than 2. The proposed estimator has the same order of magnitude and the same asymptotic constant as in \citet{Cat12}, but for the case of bounded moments. We further propose a version of the estimator that does not require even the knowledge of υ_{ε}, but adapts the moment bound in a data-driven manner. Finally, to illustrate the usefulness of the derived non-asymptotic confidence bounds, we consider an application in multi-armed bandits and propose best arm identification algorithms, in the fixed confidence setting, that outperform the state of the art.

Personalization Improves Privacy-Accuracy Tradeoffs in Federated Learning

- Alberto Bietti, Chen-Yu Wei, Miroslav Dudik, John Langford, Steven Wu
- abstract: Large-scale machine learning systems often involve data distributed across a collection of users. Federated learning algorithms leverage this structure by communicating model updates to a central server, rather than entire datasets. In this paper, we study stochastic optimization algorithms for a personalized federated learning setting involving local and global models subject to user-level (joint) differential privacy. While learning a private global model induces a cost of privacy, local learning is perfectly private. We provide generalization guarantees showing that coordinating local learning with private centralized learning yields a generically useful and improved tradeoff between accuracy and privacy. We illustrate our theoretical results with experiments on synthetic and real-world datasets.

Non-Vacuous Generalisation Bounds for Shallow Neural Networks

- Felix Biggs, Benjamin Guedj
- abstract: We focus on a specific class of shallow neural networks with a single hidden layer, namely those with L_2-normalised data and either a sigmoid-shaped Gaussian error function (“erf”) activation or a Gaussian Error Linear Unit (GELU) activation. For these networks, we derive new generalisation bounds through the PAC-Bayesian theory; unlike most existing such bounds they apply to neural networks with deterministic rather than randomised parameters. Our bounds are empirically non-vacuous when the network is trained with vanilla stochastic gradient descent on MNIST and Fashion-MNIST.

Structure-preserving GANs

- Jeremiah Birrell, Markos Katsoulakis, Luc Rey-Bellet, Wei Zhu
- abstract: Generative adversarial networks (GANs), a class of distribution-learning methods based on a two-player game between a generator and a discriminator, can generally be formulated as a minmax problem based on the variational representation of a divergence between the unknown and the generated distributions. We introduce structure-preserving GANs as a data-efficient framework for learning distributions with additional structure such as group symmetry, by developing new variational representations for divergences. Our theory shows that we can reduce the discriminator space to its projection on the invariant discriminator space, using the conditional expectation with respect to the sigma-algebra associated to the underlying structure. In addition, we prove that the discriminator space reduction must be accompanied by a careful design of structured generators, as flawed designs may easily lead to a catastrophic “mode collapse” of the learned distribution. We contextualize our framework by building symmetry-preserving GANs for distributions with intrinsic group symmetry, and demonstrate that both players, namely the equivariant generator and invariant discriminator, play important but distinct roles in the learning process. Empirical experiments and ablation studies across a broad range of data sets, including real-world medical imaging, validate our theory, and show our proposed methods achieve significantly improved sample fidelity and diversity—almost an order of magnitude measured in Frechet Inception Distance—especially in the small data regime.

Scalable Spike-and-Slab

- Niloy Biswas, Lester Mackey, Xiao-Li Meng
- abstract: Spike-and-slab priors are commonly used for Bayesian variable selection, due to their interpretability and favorable statistical properties. However, existing samplers for spike-and-slab posteriors incur prohibitive computational costs when the number of variables is large. In this article, we propose Scalable Spike-and-Slab (S^3), a scalable Gibbs sampling implementation for high-dimensional Bayesian regression with the continuous spike-and-slab prior of George & McCulloch (1993). For a dataset with n observations and p covariates, S^3 has order max{n^2 p_t, np} computational cost at iteration t where p_t never exceeds the number of covariates switching spike-and-slab states between iterations t and t-1 of the Markov chain. This improves upon the order n^2 p per-iteration cost of state-of-the-art implementations as, typically, p_t is substantially smaller than p. We apply S^3 on synthetic and real-world datasets, demonstrating orders of magnitude speed-ups over existing exact samplers and significant gains in inferential quality over approximate samplers with comparable cost.

Breaking Down Out-of-Distribution Detection: Many Methods Based on OOD Training Data Estimate a Combination of the Same Core Quantities

- Julian Bitterwolf, Alexander Meinke, Maximilian Augustin, Matthias Hein
- abstract: It is an important problem in trustworthy machine learning to recognize out-of-distribution (OOD) inputs which are inputs unrelated to the in-distribution task. Many out-of-distribution detection methods have been suggested in recent years. The goal of this paper is to recognize common objectives as well as to identify the implicit scoring functions of different OOD detection methods. We focus on the sub-class of methods that use surrogate OOD data during training in order to learn an OOD detection score that generalizes to new unseen out-distributions at test time. We show that binary discrimination between in- and (different) out-distributions is equivalent to several distinct formulations of the OOD detection problem. When trained in a shared fashion with a standard classifier, this binary discriminator reaches an OOD detection performance similar to that of Outlier Exposure. Moreover, we show that the confidence loss which is used by Outlier Exposure has an implicit scoring function which differs in a non-trivial fashion from the theoretically optimal scoring function in the case where training and test out-distribution are the same, which again is similar to the one used when training an Energy-Based OOD detector or when adding a background class. In practice, when trained in exactly the same way, all these methods perform similarly.

A query-optimal algorithm for finding counterfactuals

- Guy Blanc, Caleb Koch, Jane Lange, Li-Yang Tan
- abstract: We design an algorithm for finding counterfactuals with strong theoretical guarantees on its performance. For any monotone model $f : X^d \to {0,1}$ and instance $x^\star$, our algorithm makes [{S}(f)^{O(\Delta_f(x^\star))}\cdot \log d]{queries} to f and returns an {\sl optimal} counterfactual for $x^\star$: a nearest instance $x’$ to $x^\star$ for which $f(x’)\ne f(x^\star)$. Here $S(f)$ is the sensitivity of f, a discrete analogue of the Lipschitz constant, and $\Delta_f(x^\star)$ is the distance from $x^\star$ to its nearest counterfactuals. The previous best known query complexity was $d^{\,O(\Delta_f(x^\star))}$, achievable by brute-force local search. We further prove a lower bound of $S(f)^{\Omega(\Delta_f(x^\star))} + \Omega(\log d)$ on the query complexity of any algorithm, thereby showing that the guarantees of our algorithm are essentially optimal.

Popular decision tree algorithms are provably noise tolerant

- Guy Blanc, Jane Lange, Ali Malik, Li-Yang Tan
- abstract: Using the framework of boosting, we prove that all impurity-based decision tree learning algorithms, including the classic ID3, C4.5, and CART, are highly noise tolerant. Our guarantees hold under the strongest noise model of nasty noise, and we provide near-matching upper and lower bounds on the allowable noise rate. We further show that these algorithms, which are simple and have long been central to everyday machine learning, enjoy provable guarantees in the noisy setting that are unmatched by existing algorithms in the theoretical literature on decision tree learning. Taken together, our results add to an ongoing line of research that seeks to place the empirical success of these practical decision tree algorithms on firm theoretical footing.

Optimizing Sequential Experimental Design with Deep Reinforcement Learning

- Tom Blau, Edwin V. Bonilla, Iadine Chades, Amir Dezfouli
- abstract: Bayesian approaches developed to solve the optimal design of sequential experiments are mathematically elegant but computationally challenging. Recently, techniques using amortization have been proposed to make these Bayesian approaches practical, by training a parameterized policy that proposes designs efficiently at deployment time. However, these methods may not sufficiently explore the design space, require access to a differentiable probabilistic model and can only optimize over continuous design spaces. Here, we address these limitations by showing that the problem of optimizing policies can be reduced to solving a Markov decision process (MDP). We solve the equivalent MDP with modern deep reinforcement learning techniques. Our experiments show that our approach is also computationally efficient at deployment time and exhibits state-of-the-art performance on both continuous and discrete design spaces, even when the probabilistic model is a black box.

Lagrangian Method for Q-Function Learning (with Applications to Machine Translation)

- Huang Bojun
- abstract: This paper discusses a new approach to the fundamental problem of learning optimal Q-functions. In this approach, optimal Q-functions are formulated as saddle points of a nonlinear Lagrangian function derived from the classic Bellman optimality equation. The paper shows that the Lagrangian enjoys strong duality, in spite of its nonlinearity, which paves the way to a general Lagrangian method to Q-function learning. As a demonstration, the paper develops an imitation learning algorithm based on the duality theory, and applies the algorithm to a state-of-the-art machine translation benchmark. The paper then turns to demonstrate a symmetry breaking phenomenon regarding the optimality of the Lagrangian saddle points, which justifies a largely overlooked direction in developing the Lagrangian method.

Generalized Results for the Existence and Consistency of the MLE in the Bradley-Terry-Luce Model

- Heejong Bong, Alessandro Rinaldo
- abstract: Ranking problems based on pairwise comparisons, such as those arising in online gaming, often involve a large pool of items to order. In these situations, the gap in performance between any two items can be significant, and the smallest and largest winning probabilities can be very close to zero or one. Furthermore, each item may be compared only to a subset of all the items, so that not all pairwise comparisons are observed. In this paper, we study the performance of the Bradley-Terry-Luce model for ranking from pairwise comparison data under more realistic settings than those considered in the literature so far. In particular, we allow for near-degenerate winning probabilities and arbitrary comparison designs. We obtain novel results about the existence of the maximum likelihood estimator (MLE) and the corresponding ℓ_2 estimation error without the bounded winning probability assumption commonly used in the literature and for arbitrary comparison graph topologies. Central to our approach is the reliance on the Fisher information matrix to express the dependence on the graph topologies and the impact of the values of the winning probabilities on the estimation risk and on the conditions for the existence of the MLE. Our bounds recover existing results as special cases but are more broadly applicable.

How to Train Your Wide Neural Network Without Backprop: An Input-Weight Alignment Perspective

- Akhilan Boopathy, Ila Fiete
- abstract: Recent works have examined theoretical and empirical properties of wide neural networks trained in the Neural Tangent Kernel (NTK) regime. Given that biological neural networks are much wider than their artificial counterparts, we consider NTK regime wide neural networks as a possible model of biological neural networks. Leveraging NTK theory, we show theoretically that gradient descent drives layerwise weight updates that are aligned with their input activity correlations weighted by error, and demonstrate empirically that the result also holds in finite-width wide networks. The alignment result allows us to formulate a family of biologically-motivated, backpropagation-free learning rules that are theoretically equivalent to backpropagation in infinite-width networks. We test these learning rules on benchmark problems in feedforward and recurrent neural networks and demonstrate, in wide networks, comparable performance to backpropagation. The proposed rules are particularly effective in low data regimes, which are common in biological learning settings.

Improving Language Models by Retrieving from Trillions of Tokens

- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, Laurent Sifre
- abstract: We enhance auto-regressive language models by conditioning on document chunks retrieved from a large corpus, based on local similarity with preceding tokens. With a 2 trillion token database, our Retrieval-Enhanced Transformer (RETRO) obtains comparable performance to GPT-3 and Jurassic-1 on the Pile, despite using 25{\texttimes} fewer parameters. After fine-tuning, RETRO performance translates to downstream knowledge-intensive tasks such as question answering. RETRO combines a frozen Bert retriever, a differentiable encoder and a chunked cross-attention mechanism to predict tokens based on an order of magnitude more data than what is typically consumed during training. We typically train RETRO from scratch, yet can also rapidly RETROfit pre-trained transformers with retrieval and still achieve good performance. Our work opens up new avenues for improving language models through explicit memory at unprecedented scale.

Lie Point Symmetry Data Augmentation for Neural PDE Solvers

- Johannes Brandstetter, Max Welling, Daniel E Worrall
- abstract: Neural networks are increasingly being used to solve partial differential equations (PDEs), replacing slower numerical solvers. However, a critical issue is that neural PDE solvers require high-quality ground truth data, which usually must come from the very solvers they are designed to replace. Thus, we are presented with a proverbial chicken-and-egg problem. In this paper, we present a method, which can partially alleviate this problem, by improving neural PDE solver sample complexity—Lie point symmetry data augmentation (LPSDA). In the context of PDEs, it turns out we are able to quantitatively derive an exhaustive list of data transformations, based on the Lie point symmetry group of the PDEs in question, something not possible in other application areas. We present this framework and demonstrate how it can easily be deployed to improve neural PDE solver sample complexity by an order of magnitude.

An iterative clustering algorithm for the Contextual Stochastic Block Model with optimality guarantees

- Guillaume Braun, Hemant Tyagi, Christophe Biernacki
- abstract: Real-world networks often come with side information that can help to improve the performance of network analysis tasks such as clustering. Despite a large number of empirical and theoretical studies conducted on network clustering methods during the past decade, the added value of side information and the methods used to incorporate it optimally in clustering algorithms are relatively less understood. We propose a new iterative algorithm to cluster networks with side information for nodes (in the form of covariates) and show that our algorithm is optimal under the Contextual Symmetric Stochastic Block Model. Our algorithm can be applied to general Contextual Stochastic Block Models and avoids hyperparameter tuning in contrast to previously proposed methods. We confirm our theoretical results on synthetic data experiments where our algorithm significantly outperforms other methods, and show that it can also be applied to signed graphs. Finally we demonstrate the practical interest of our method on real data.

Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems

- Manuel Brenner, Florian Hess, Jonas M Mikhaeil, Leonard F Bereska, Zahra Monfared, Po-Chen Kuo, Daniel Durstewitz
- abstract: In many scientific disciplines, we are interested in inferring the nonlinear dynamical system underlying a set of observed time series, a challenging task in the face of chaotic behavior and noise. Previous deep learning approaches toward this goal often suffered from a lack of interpretability and tractability. In particular, the high-dimensional latent spaces often required for a faithful embedding, even when the underlying dynamics lives on a lower-dimensional manifold, can hamper theoretical analysis. Motivated by the emerging principles of dendritic computation, we augment a dynamically interpretable and mathematically tractable piecewise-linear (PL) recurrent neural network (RNN) by a linear spline basis expansion. We show that this approach retains all the theoretically appealing properties of the simple PLRNN, yet boosts its capacity for approximating arbitrary nonlinear dynamical systems in comparatively low dimensions. We employ two frameworks for training the system, one combining BPTT with teacher forcing, and another based on fast and scalable variational inference. We show that the dendritically expanded PLRNN achieves better reconstructions with fewer parameters and dimensions on various dynamical systems benchmarks and compares favorably to other methods, while retaining a tractable and interpretable structure.

Learning to Predict Graphs with Fused Gromov-Wasserstein Barycenters

- Luc Brogat-Motte, Rémi Flamary, Celine Brouard, Juho Rousu, Florence D'Alché-Buc
- abstract: This paper introduces a novel and generic framework to solve the flagship task of supervised labeled graph prediction by leveraging Optimal Transport tools. We formulate the problem as regression with the Fused Gromov-Wasserstein (FGW) loss and propose a predictive model relying on a FGW barycenter whose weights depend on inputs. First we introduce a non-parametric estimator based on kernel ridge regression for which theoretical results such as consistency and excess risk bound are proved. Next we propose an interpretable parametric model where the barycenter weights are modeled with a neural network and the graphs on which the FGW barycenter is calculated are additionally learned. Numerical experiments show the strength of the method and its ability to interpolate in the labeled graph space on simulated data and on a difficult metabolic identification problem where it can reach very good performance with very little engineering.

Efficient Learning of CNNs using Patch Based Features

- Alon Brutzkus, Amir Globerson, Eran Malach, Alon Regev Netser, Shai Shalev-Schwartz
- abstract: Recent work has demonstrated the effectiveness of using patch based representations when learning from image data. Here we provide theoretical support for this observation, by showing that a simple semi-supervised algorithm that uses patch statistics can efficiently learn labels produced by a one-hidden-layer Convolutional Neural Network (CNN). Since CNNs are known to be computationally hard to learn in the worst case, our analysis holds under some distributional assumptions. We show that these assumptions are necessary and sufficient for our results to hold. We verify that the distributional assumptions hold on real-world data by experimenting on the CIFAR-10 dataset, and find that the analyzed algorithm outperforms a vanilla one-hidden-layer CNN. Finally, we demonstrate that by running the algorithm in a layer-by-layer fashion we can build a deep model which gives further improvements, hinting that this method provides insights about the behavior of deep CNNs.

Causal structure-based root cause analysis of outliers

- Kailash Budhathoki, Lenon Minorics, Patrick Bloebaum, Dominik Janzing
- abstract: Current techniques for explaining outliers cannot tell what caused the outliers. We present a formal method to identify "root causes" of outliers, amongst variables. The method requires a causal graph of the variables along with the functional causal model. It quantifies the contribution of each variable to the target outlier score, which explains to what extent each variable is a "root cause" of the target outlier. We study the empirical performance of the method through simulations and present a real-world case study identifying "root causes" of extreme river flows.

IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages

- Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, Ivan Vulić
- abstract: Reliable evaluation benchmarks designed for replicability and comprehensiveness have driven progress in machine learning. Due to the lack of a multilingual benchmark, however, vision-and-language research has mostly focused on English language tasks. To fill this gap, we introduce the Image-Grounded Language Understanding Evaluation benchmark. IGLUE brings together{—}by both aggregating pre-existing datasets and creating new ones{—}visual question answering, cross-modal retrieval, grounded reasoning, and grounded entailment tasks across 20 diverse languages. Our benchmark enables the evaluation of multilingual multimodal models for transfer learning, not only in a zero-shot setting, but also in newly defined few-shot learning setups. Based on the evaluation of the available state-of-the-art models, we find that translate-test transfer is superior to zero-shot transfer and that few-shot learning is hard to harness for many tasks. Moreover, downstream performance is partially explained by the amount of available unlabelled textual data for pretraining, and only weakly by the typological distance of target{–}source languages. We hope to encourage future research efforts in this area by releasing the benchmark to the community.

Interactive Inverse Reinforcement Learning for Cooperative Games

- Thomas Kleine Büning, Anne-Marie George, Christos Dimitrakakis
- abstract: We study the problem of designing autonomous agents that can learn to cooperate effectively with a potentially suboptimal partner while having no access to the joint reward function. This problem is modeled as a cooperative episodic two-agent Markov decision process. We assume control over only the first of the two agents in a Stackelberg formulation of the game, where the second agent is acting so as to maximise expected utility given the first agent's policy. How should the first agent act in order to learn the joint reward function as quickly as possible and so that the joint policy is as close to optimal as possible? We analyse how knowledge about the reward function can be gained in this interactive two-agent scenario. We show that when the learning agent's policies have a significant effect on the transition function, the reward function can be learned efficiently.

Convolutional and Residual Networks Provably Contain Lottery Tickets

- Rebekka Burkholz
- abstract: The Lottery Ticket Hypothesis continues to have a profound practical impact on the quest for small scale deep neural networks that solve modern deep learning tasks at competitive performance. These lottery tickets are identified by pruning large randomly initialized neural networks with architectures that are as diverse as their applications. Yet, theoretical insights that attest their existence have been mostly focused on deed fully-connected feed forward networks with ReLU activation functions. We prove that also modern architectures consisting of convolutional and residual layers that can be equipped with almost arbitrary activation functions can contain lottery tickets with high probability.

Near-Optimal Algorithms for Autonomous Exploration and Multi-Goal Stochastic Shortest Path

- Haoyuan Cai, Tengyu Ma, Simon Du
- abstract: We revisit the incremental autonomous exploration problem proposed by Lim and Auer (2012). In this setting, the agent aims to learn a set of near-optimal goal-conditioned policies to reach the L-controllable states: states that are incrementally reachable from an initial state s_0 within L steps in expectation. We introduce a new algorithm with stronger sample complexity bounds than existing ones. Furthermore, we also prove the first lower bound for the autonomous exploration problem. In particular, the lower bound implies that our proposed algorithm, Value-Aware Autonomous Exploration, is nearly minimax-optimal when the number of L-controllable states grows polynomially with respect to L. Key in our algorithm design is a connection between autonomous exploration and multi-goal stochastic shortest path, a new problem that naturally generalizes the classical stochastic shortest path problem. This new problem and its connection to autonomous exploration can be of independent interest.

Convergence of Invariant Graph Networks

- Chen Cai, Yusu Wang
- abstract: Although theoretical properties such as expressive power and over-smoothing of graph neural networks (GNN) have been extensively studied recently, its convergence property is a relatively new direction. In this paper, we investigate the convergence of one powerful GNN, Invariant Graph Network (IGN) over graphs sampled from graphons. We first prove the stability of linear layers for general k-IGN (of order k) based on a novel interpretation of linear equivariant layers. Building upon this result, we prove the convergence of k-IGN under the model of \citet{ruiz2020graphon}, where we access the edge weight but the convergence error is measured for graphon inputs. Under the more natural (and more challenging) setting of

\citet{keriven2020convergence} where one can only access 0-1 adjacency matrix sampled according to edge probability, we first show a negative result that the convergence of any IGN is not possible. We then obtain the convergence of a subset of IGNs, denoted as IGN-small, after the edge probability estimation. We show that IGN-small still contains function class rich enough that can approximate spectral GNNs arbitrarily well. Lastly, we perform experiments on various graphon models to verify our statements.

Reinforcement Learning from Partial Observation: Linear Function Approximation with Provable Sample Efficiency

- Qi Cai, Zhuoran Yang, Zhaoran Wang
- abstract: We study reinforcement learning for partially observed Markov decision processes (POMDPs) with infinite observation and state spaces, which remains less investigated theoretically. To this end, we make the first attempt at bridging partial observability and function approximation for a class of POMDPs with a linear structure. In detail, we propose a reinforcement learning algorithm (Optimistic Exploration via Adversarial Integral Equation or OP-TENET) that attains an ϵ-optimal policy within $O(1/\epsilon^2)$ episodes. In particular, the sample complexity scales polynomially in the intrinsic dimension of the linear structure and is independent of the size of the observation and state spaces. The sample efficiency of OP-TENET is enabled by a sequence of ingredients: (i) a Bellman operator with finite memory, which represents the value function in a recursive manner, (ii) the identification and estimation of such an operator via an adversarial integral equation, which features a smoothed discriminator tailored to the linear structure, and (iii) the exploration of the observation and state spaces via optimism, which is based on quantifying the uncertainty in the adversarial integral equation.

Scaling Gaussian Process Optimization by Evaluating a Few Unique Candidates Multiple Times

- Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, Lorenzo Rosasco
- abstract: Computing a Gaussian process (GP) posterior has a computational cost cubical in the number of historical points. A reformulation of the same GP posterior highlights that this complexity mainly depends on how many unique historical points are considered. This can have important implication in active learning settings, where the set of historical points is constructed sequentially by the learner. We show that sequential black-box optimization based on GPs (GP-Opt) can be made efficient by sticking to a candidate solution for multiple evaluation steps and switch only when necessary. Limiting the number of switches also limits the number of unique points in the history of the GP. Thus, the efficient GP reformulation can be used to exactly and cheaply compute the posteriors required to run the GP-Opt algorithms. This approach is especially useful in real-world applications of GP-Opt with high switch costs (e.g. switching chemicals in wet labs, data/model loading in hyperparameter optimization). As examples of this meta-approach, we modify two well-established GP-Opt algorithms, GP-UCB and GP-EI, to switch candidates as infrequently as possible adapting rules from batched GP-Opt. These versions preserve all the theoretical no-regret guarantees while improving practical aspects of the algorithms such as runtime, memory complexity, and the ability of batching candidates and evaluating them in parallel.

Adaptive Gaussian Process Change Point Detection

- Edoardo Caldarelli, Philippe Wenk, Stefan Bauer, Andreas Krause
- abstract: Detecting change points in time series, i.e., points in time at which some observed process suddenly changes, is a fundamental task that arises in many real-world applications, with consequences for safety and reliability. In this work, we propose ADAGA, a novel Gaussian process-based solution to this problem, that leverages a powerful heuristics we developed based on statistical hypothesis testing. In contrast to prior approaches, ADAGA adapts to changes both in mean and covariance structure of the temporal process. In extensive experiments, we show its versatility and applicability to different classes of change points, demonstrating that it is significantly more accurate than current state-of-the-art alternatives.

Measuring dissimilarity with diffeomorphism invariance

- Théophile Cantelobre, Carlo Ciliberto, Benjamin Guedj, Alessandro Rudi
- abstract: Measures of similarity (or dissimilarity) are a key ingredient to many machine learning algorithms. We introduce DID, a pairwise dissimilarity measure applicable to a wide range of data spaces, which leverages the data's internal structure to be invariant to diffeomorphisms. We prove that DID enjoys properties which make it relevant for theoretical study and practical use. By representing each datum as a function, DID is defined as the solution to an optimization problem in a Reproducing Kernel Hilbert Space and can be expressed in closed-form. In practice, it can be efficiently approximated via Nystr{ö}m sampling. Empirical experiments support the merits of DID.

A Model-Agnostic Randomized Learning Framework based on Random Hypothesis Subspace Sampling

- Yiting Cao, Chao Lan
- abstract: We propose a model-agnostic randomized learning framework based on Random Hypothesis Subspace Sampling (RHSS). Given any hypothesis class, it randomly samples k hypotheses and learns a near-optimal model from their span by simply solving a linear least square problem in $O(n k^2)$ time, where n is the number of training instances. On the theory side, we derive the performance guarantee of RHSS from a generic subspace approximation perspective, leveraging properties of metric entropy and random matrices. On the practical side, we apply the RHSS framework to learn kernel, network and tree based models. Experimental results show they converge efficiently as k increases and outperform their model-specific counterparts including random Fourier feature, random vector functional link and extra tree on real-world data sets.

Gaussian Process Uniform Error Bounds with Unknown Hyperparameters for Safety-Critical Applications

- Alexandre Capone, Armin Lederer, Sandra Hirche
- abstract: Gaussian processes have become a promising tool for various safety-critical settings, since the posterior variance can be used to directly estimate the model error and quantify risk. However, state-of-the-art techniques for safety-critical settings hinge on the assumption that the kernel hyperparameters are known, which does not apply in general. To mitigate this, we introduce robust Gaussian process uniform error bounds in settings with unknown hyperparameters. Our approach computes a confidence region in the space of hyperparameters, which enables us to obtain a probabilistic upper bound for the model error of a Gaussian process with arbitrary hyperparameters. We do not require to know any bounds for the hyperparameters a priori, which is an assumption commonly found in related work. Instead, we are able to derive bounds from data in an intuitive fashion. We additionally employ the proposed technique to derive performance guarantees for a class of learning-based control problems. Experiments show that the bound performs significantly better than vanilla and fully Bayesian Gaussian processes.

Burst-Dependent Plasticity and Dendritic Amplification Support Target-Based Learning and Hierarchical Imitation Learning

- Cristiano Capone, Cosimo Lupo, Paolo Muratore, Pier Stanislao Paolucci
- abstract: The brain can learn to solve a wide range of tasks with high temporal and energetic efficiency. However, most biological models are composed of simple single-compartment neurons and cannot achieve the state-of-the-art performances of artificial intelligence. We propose a multi-compartment model of pyramidal neuron, in which bursts and dendritic input segregation give the possibility to plausibly support a biological target-based learning. In target-based learning, the internal solution of a problem (a spatio-temporal pattern of bursts in our case) is suggested to the network, bypassing the problems of error backpropagation and credit assignment. Finally, we show that this neuronal architecture naturally supports the orchestration of "hierarchical imitation learning", enabling the decomposition of challenging long-horizon decision-making tasks into simpler subtasks.

A Marriage between Adversarial Team Games and 2-player Games: Enabling Abstractions, No-regret Learning, and Subgame Solving

- Luca Carminati, Federico Cacciamani, Marco Ciccone, Nicola Gatti
- abstract: Ex ante correlation is becoming the mainstream approach for sequential adversarial team games, where a team of players faces another team in a zero-sum game. It is known that team members' asymmetric information makes both equilibrium computation \textsf{APX}-hard and team's strategies not directly representable on the game tree. This latter issue prevents the adoption of successful tools for huge 2-player zero-sum games such as, e.g., abstractions, no-regret learning, and subgame solving. This work shows that we can recover from this weakness by bridging the gap between sequential adversarial team games and 2-player games. In particular, we propose a new, suitable game representation that we call team-public-information, in which a team is represented as a single coordinator who only knows information common to the whole team and prescribes to each member an action for any possible private state. The resulting representation is highly explainable, being a 2-player tree in which the team's strategies are behavioral with a direct interpretation and more expressive than the original extensive form when designing abstractions. Furthermore, we prove payoff equivalence of our representation, and we provide techniques that, starting directly from the extensive form, generate dramatically more compact representations without information loss. Finally, we experimentally evaluate our techniques when applied to a standard testbed, comparing their performance with the current state of the art.

RECAPP: Crafting a More Efficient Catalyst for Convex Optimization

- Yair Carmon, Arun Jambulapati, Yujia Jin, Aaron Sidford
- abstract: The accelerated proximal point method (APPA), also known as "Catalyst", is a well-established reduction from convex optimization to approximate proximal point computation (i.e., regularized minimization). This reduction is conceptually elegant and yields strong convergence rate guarantees. However, these rates feature an extraneous logarithmic term arising from the need to compute each proximal point to high accuracy. In this work, we propose a novel Relaxed Error Criterion for Accelerated Proximal Point (RECAPP) that eliminates the need for high accuracy subproblem solutions. We apply RECAPP to two canonical problems: finite-sum and max-structured minimization. For finite-sum problems, we match the best known complexity, previously obtained by carefully-designed problem-specific algorithms. For minimizing max_y f(x,y) where f is convex in x and strongly-concave in y, we improve on the best known (Catalyst-based) bound by a logarithmic factor.

Estimating and Penalizing Induced Preference Shifts in Recommender Systems

- Micah D Carroll, Anca Dragan, Stuart Russell, Dylan Hadfield-Menell
- abstract: The content that a recommender system (RS) shows to users influences them. Therefore, when choosing a recommender to deploy, one is implicitly also choosing to induce specific internal states in users. Even more, systems trained via long-horizon optimization will have direct incentives to manipulate users, e.g. shift their preferences so they are easier to satisfy. We focus on induced preference shifts in users. We argue that {–} before deployment {–} system designers should: estimate the shifts a recommender would induce; evaluate whether such shifts would be undesirable; and perhaps even actively optimize to avoid problematic shifts. These steps involve two challenging ingredients: estimation requires anticipating how hypothetical policies would influence user preferences if deployed {–} we do this by using historical user interaction data to train a predictive user model which implicitly contains their preference dynamics; evaluation and optimization additionally require metrics to assess whether such influences are manipulative or otherwise unwanted {–} we use the notion of "safe shifts", that define a trust region within which behavior is safe: for instance, the natural way in which users would shift without interference from the system could be deemed "safe". In simulated experiments, we show that our learned preference dynamics model is effective in estimating user preferences and how they would respond to new recommenders. Additionally, we show that recommenders that optimize for staying in the trust region can avoid manipulative behaviors while still generating engagement.

YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone

- Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, Moacir A Ponti
- abstract: YourTTS brings the power of a multilingual approach to the task of zero-shot multi-speaker TTS. Our method builds upon the VITS model and adds several novel modifications for zero-shot multi-speaker and multilingual training. We achieved state-of-the-art (SOTA) results in zero-shot multi-speaker TTS and results comparable to SOTA in zero-shot voice conversion on the VCTK dataset. Additionally, our approach achieves promising results in a target language with a single-speaker dataset, opening possibilities for zero-shot multi-speaker TTS and zero-shot voice conversion systems in low-resource languages. Finally, it is possible to fine-tune the YourTTS model with less than 1 minute of speech and achieve state-of-the-art results in voice similarity and with reasonable quality. This is important to allow synthesis for speakers with a very different voice or recording characteristics from those seen during training.

The Infinite Contextual Graph Markov Model

- Daniele Castellana, Federico Errica, Davide Bacciu, Alessio Micheli
- abstract: The Contextual Graph Markov Model (CGMM) is a deep, unsupervised, and probabilistic model for graphs that is trained incrementally on a layer-by-layer basis. As with most Deep Graph Networks, an inherent limitation is the need to perform an extensive model selection to choose the proper size of each layer's latent representation. In this paper, we address this problem by introducing the Infinite Contextual Graph Markov Model (iCGMM), the first deep Bayesian nonparametric model for graph learning. During training, iCGMM can adapt the complexity of each layer to better fit the underlying data distribution. On 8 graph classification tasks, we show that iCGMM: i) successfully recovers or improves CGMM's performances while reducing the hyper-parameters' search space; ii) performs comparably to most end-to-end supervised methods. The results include studies on the importance of depth, hyper-parameters, and compression of the graph embeddings. We also introduce a novel approximated inference procedure that better deals with larger graph topologies.

Compressed-VFL: Communication-Efficient Learning with Vertically Partitioned Data

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

Online Learning with Knapsacks: the Best of Both Worlds

- Matteo Castiglioni, Andrea Celli, Christian Kroer
- abstract: We study online learning problems in which a decision maker wants to maximize their expected reward without violating a finite set of m resource constraints. By casting the learning process over a suitably defined space of strategy mixtures, we recover strong duality on a Lagrangian

relaxation of the underlying optimization problem, even for general settings with non-convex reward and resource-consumption functions. Then, we provide the first best-of-both-worlds type framework for this setting, with no-regret guarantees both under stochastic and adversarial inputs. Our framework yields the same regret guarantees of prior work in the stochastic case. On the other hand, when budgets grow at least linearly in the time horizon, it allows us to provide a constant competitive ratio in the adversarial case, which improves over the $O(m \log T)$ competitive ratio of Immorlica et al. [FOCS'19]. Moreover, our framework allows the decision maker to handle non-convex reward and cost functions. We provide two game-theoretic applications of our framework to give further evidence of its flexibility.

Stabilizing Off-Policy Deep Reinforcement Learning from Pixels

- Edoardo Cetin, Philip J Ball, Stephen Roberts, Oya Celiktutan
- abstract: Off-policy reinforcement learning (RL) from pixel observations is notoriously unstable. As a result, many successful algorithms must combine different domain-specific practices and auxiliary losses to learn meaningful behaviors in complex environments. In this work, we provide novel analysis demonstrating that these instabilities arise from performing temporal-difference learning with a convolutional encoder and low-magnitude rewards. We show that this new visual deadly triad causes unstable training and premature convergence to degenerate solutions, a phenomenon we name catastrophic self-overfitting. Based on our analysis, we propose A-LIX, a method providing adaptive regularization to the encoder's gradients that explicitly prevents the occurrence of catastrophic self-overfitting using a dual objective. By applying A-LIX, we significantly outperform the prior state-of-the-art on the DeepMind Control and Atari benchmarks without any data augmentation or auxiliary losses.

Accelerated, Optimal and Parallel: Some results on model-based stochastic optimization

- Karan Chadha, Gary Cheng, John Duchi
- abstract: The Approximate-Proximal Point (APROX) family of model-based stochastic optimization algorithms improve over standard stochastic gradient methods, as they are robust to step size choices, adaptive to problem difficulty, converge on a broader range of problems than stochastic gradient methods, and converge very fast on interpolation problems, all while retaining nice minibatching properties \cite{AsiDu19siopt,AsiChChDu20}. In this paper, we propose an acceleration scheme for the APROX family and provide non-asymptotic convergence guarantees, which are order-optimal in all problem-dependent constants and provide even larger minibatching speedups. For interpolation problems where the objective satisfies additional growth conditions, we show that our algorithm achieves linear convergence rates for a wide range of stepsizes. In this setting, we also prove matching lower bounds, identifying new fundamental constants and showing the optimality of the APROX family. We corroborate our theoretical results with empirical testing to demonstrate the gains accurate modeling, acceleration, and minibatching provide.

Robust Imitation Learning against Variations in Environment Dynamics

- Jongseong Chae, Seungyul Han, Whiyoung Jung, Myungsik Cho, Sungho Choi, Youngchul Sung
- abstract: In this paper, we propose a robust imitation learning (IL) framework that improves the robustness of IL when environment dynamics are perturbed. The existing IL framework trained in a single environment can catastrophically fail with perturbations in environment dynamics because it does not capture the situation that underlying environment dynamics can be changed. Our framework effectively deals with environments with varying dynamics by imitating multiple experts in sampled environment dynamics to enhance the robustness in general variations in environment dynamics. In order to robustly imitate the multiple sample experts, we minimize the risk with respect to the Jensen-Shannon divergence between the agent's policy and each of the sample experts. Numerical results show that our algorithm significantly improves robustness against dynamics perturbations compared to conventional IL baselines.

Fairness with Adaptive Weights

- Junyi Chai, Xiaoqian Wang
- abstract: Fairness is now an important issue in machine learning. There are arising concerns that automated decision-making systems reflect real-world biases. Although a wide range of fairness-related methods have been proposed in recent years, the under-representation problem has been less studied. Due to the uneven distribution of samples from different populations, machine learning models tend to be biased against minority groups when trained by minimizing the average empirical risk across all samples. In this paper, we propose a novel adaptive reweighing method to address representation bias. The goal of our method is to achieve group-level balance among different demographic groups by learning adaptive weights for each sample. Our approach emphasizes more on error-prone samples in prediction and enhances adequate representation of minority groups for fairness. We derive a closed-form solution for adaptive weight assignment and propose an efficient algorithm with theoretical convergence guarantees. We theoretically analyze the fairness of our model and empirically verify that our method strikes a balance between fairness and accuracy. In experiments, our method achieves comparable or better performance than state-of-the-art methods in both classification and regression tasks. Furthermore, our method exhibits robustness to label noise on various benchmark datasets.

UNIREX: A Unified Learning Framework for Language Model Rationale Extraction

- Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, Hamed Firooz
- abstract: An extractive rationale explains a language model's (LM's) prediction on a given task instance by highlighting the text inputs that most influenced the prediction. Ideally, rationale extraction should be faithful (reflective of LM's actual behavior) and plausible (convincing to humans), without compromising the LM's (i.e., task model's) task performance. Although attribution algorithms and select-predict pipelines are commonly used in rationale extraction, they both rely on certain heuristics that hinder them from satisfying all three desiderata. In light of this, we propose UNIREX, a flexible learning framework which generalizes rationale extractor optimization as follows: (1) specify architecture for a learned rationale extractor; (2) select explainability objectives (\ie faithfulness and plausibility criteria); and (3) jointly train the task model and rationale extractor on the task using selected objectives. UNIREX enables replacing prior works' heuristic design choices with a generic learned rationale extractor in (1) and optimizing it for all three desiderata in (2)-(3). To facilitate comparison between methods w.r.t. multiple desiderata, we introduce the Normalized Relative Gain (NRG) metric. On five English text classification datasets, our best UNIREX configuration outperforms baselines by an average of 32.9% NRG. Plus, UNIREX rationale extractors' faithfulness can even generalize to unseen datasets and tasks.

Revisiting Label Smoothing and Knowledge Distillation Compatibility: What was Missing?

- Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Yunqing Zhao, Ngai-Man Cheung
- abstract: This work investigates the compatibility between label smoothing (LS) and knowledge distillation (KD). Contemporary findings addressing this thesis statement take dichotomous standpoints: Muller et al. (2019) and Shen et al. (2021b). Critically, there is no effort to understand and resolve these contradictory findings, leaving the primal question \text{-} to smooth or not to smooth a teacher network? \text{-} unanswered. The main contributions of our work are the discovery, analysis and validation of systematic diffusion as the missing concept which is instrumental in understanding and resolving these contradictory findings. This systematic diffusion essentially curtails the benefits of distilling from an LS-trained teacher, thereby rendering KD at increased temperatures ineffective. Our discovery is comprehensively supported by large-scale experiments, analyses and case studies including image classification, neural machine translation and compact student distillation tasks spanning across multiple datasets and teacher-student architectures. Based on our analysis, we suggest practitioners to use an LS-trained teacher with a low-temperature transfer to achieve high performance students. Code and models are available at https://keshik6.github.io/revisiting-ls-kd-compatibility/

Style Equalization: Unsupervised Learning of Controllable Generative Sequence Models

- Jen-Hao Rick Chang, Ashish Shrivastava, Hema Koppula, Xiaoshuai Zhang, Oncel Tuzel
- abstract: Controllable generative sequence models with the capability to extract and replicate the style of specific examples enable many applications, including narrating audiobooks in different voices, auto-completing and auto-correcting written handwriting, and generating missing training samples for downstream recognition tasks. However, under an unsupervised-style setting, typical training algorithms for controllable sequence generative models suffer from the training-inference mismatch, where the same sample is used as content and style input during training but unpaired samples are given during inference. In this paper, we tackle the training-inference mismatch encountered during unsupervised learning of controllable generative sequence models. The proposed method is simple yet effective, where we use a style transformation module to transfer target style information into an unrelated style input. This method enables training using unpaired content and style samples and thereby mitigate the training-inference mismatch. We apply style equalization to text-to-speech and text-to-handwriting synthesis on three datasets. We conduct thorough evaluation, including both quantitative and qualitative user studies. Our results show that by mitigating the training-inference mismatch with the proposed style equalization, we achieve style replication scores comparable to real data in our user studies.

Learning Bellman Complete Representations for Offline Policy Evaluation

- Jonathan Chang, Kaiwen Wang, Nathan Kallus, Wen Sun
- abstract: We study representation learning for Offline Reinforcement Learning (RL), focusing on the important task of Offline Policy Evaluation (OPE). Recent work shows that, in contrast to supervised learning, realizability of the Q-function is not enough for learning it. Two sufficient conditions for sample-efficient OPE are Bellman completeness and coverage. Prior work often assumes that representations satisfying these conditions are given, with results being mostly theoretical in nature. In this work, we propose BCRL, which directly learns from data an approximately linear Bellman complete representation with good coverage. With this learned representation, we perform OPE using Least Square Policy Evaluation (LSPE) with linear functions in our learned representation. We present an end-to-end theoretical analysis, showing that our two-stage algorithm enjoys polynomial sample complexity provided some representation in the rich class considered is linear Bellman complete. Empirically, we extensively evaluate our algorithm on challenging, image-based continuous control tasks from the Deepmind Control Suite. We show our representation enables better OPE compared to previous representation learning methods developed for off-policy RL (e.g., CURL, SPR). BCRL achieve competitive OPE error with the state-of-the-art method Fitted Q-Evaluation (FQE), and beats FQE when evaluating beyond the initial state distribution. Our ablations show that both linear Bellman complete and coverage components of our method are crucial.

Sample Efficient Learning of Predictors that Complement Humans

- Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, Samira Samadi
- abstract: One of the goals of learning algorithms is to complement and reduce the burden on human decision makers. The expert deferral setting wherein an algorithm can either predict on its own or defer the decision to a downstream expert helps accomplish this goal. A fundamental aspect of this setting is the need to learn complementary predictors that improve on the human’s weaknesses rather than learning predictors optimized for average error. In this work, we provide the first theoretical analysis of the benefit of learning complementary predictors in expert deferral. To enable efficiently learning such predictors, we consider a family of consistent surrogate loss functions for expert deferral and analyze their theoretical properties. Finally, we design active learning schemes that require minimal amount of data of human expert predictions in order to learn accurate deferral systems.

Nyström Kernel Mean Embeddings

- Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, Alessandro Rudi
- abstract: Kernel mean embeddings are a powerful tool to represent probability distributions over arbitrary spaces as single points in a Hilbert space. Yet, the cost of computing and storing such embeddings prohibits their direct use in large-scale settings. We propose an efficient approximation procedure based on the Nystr{ö}m method, which exploits a small random subset of the dataset. Our main result is an upper bound on the approximation error of this procedure. It yields sufficient conditions on the subsample size to obtain the standard (1/sqrt(n)) rate while reducing computational costs. We discuss applications of this result for the approximation of the maximum mean discrepancy and quadrature rules, and we illustrate our theoretical findings with numerical experiments.

Coarsening the Granularity: Towards Structurally Sparse Lottery Tickets

- Tianlong Chen, Xuxi Chen, Xiaolong Ma, Yanzhi Wang, Zhangyang Wang
- abstract: The lottery ticket hypothesis (LTH) has shown that dense models contain highly sparse subnetworks (i.e., winning tickets) that can be trained in isolation to match full accuracy. Despite many exciting efforts being made, there is one "commonsense" rarely challenged: a winning ticket is found by iterative magnitude pruning (IMP) and hence the resultant pruned subnetworks have only unstructured sparsity. That gap limits the appeal of winning tickets in practice, since the highly irregular sparse patterns are challenging to accelerate on hardware. Meanwhile, directly substituting structured pruning for unstructured pruning in IMP damages performance more severely and is usually unable to locate winning tickets. In this paper, we demonstrate the first positive result that a structurally sparse winning ticket can be effectively found in general. The core idea is to append "post-processing techniques" after each round of (unstructured) IMP, to enforce the formation of structural sparsity. Specifically, we first "re-fill" pruned elements back in some channels deemed to be important, and then "re-group" non-zero elements to create flexible group-wise structural patterns. Both our identified channel- and group-wise structural subnetworks win the lottery, with substantial inference speedups readily supported by existing hardware. Extensive experiments, conducted on diverse datasets across multiple network backbones, consistently validate our proposal, showing that the hardware acceleration roadblock of LTH is now removed. Specifically, the structural winning tickets obtain up to {64.93%, 64.84%, 60.23%} running time savings at {36% 80%, 74%, 58%} sparsity on {CIFAR, Tiny-ImageNet, ImageNet}, while maintaining comparable accuracy. Code is at https://github.com/VITA-Group/Structure-LTH.

Learning Domain Adaptive Object Detection with Probabilistic Teacher

- Meilin Chen, Weijie Chen, Shicai Yang, Jie Song, Xinchao Wang, Lei Zhang, Yunfeng Yan, Donglian Qi, Yueting Zhuang, Di Xie, Shiliang Pu
- abstract: Self-training for unsupervised domain adaptive object detection is a challenging task, of which the performance depends heavily on the quality of pseudo boxes. Despite the promising results, prior works have largely overlooked the uncertainty of pseudo boxes during self-training. In this paper, we present a simple yet effective framework, termed as Probabilistic Teacher (PT), which aims to capture the uncertainty of unlabeled target data from a gradually evolving teacher and guides the learning of a student in a mutually beneficial manner. Specifically, we propose to leverage the uncertainty-guided consistency training to promote classification adaptation and localization adaptation, rather than filtering pseudo boxes via an elaborate confidence threshold. In addition, we conduct anchor adaptation in parallel with localization adaptation, since anchor can be regarded as a learnable parameter. Together with this framework, we also present a novel Entropy Focal Loss (EFL) to further facilitate the uncertainty-guided self-training. Equipped with EFL, PT outperforms all previous baselines by a large margin and achieve new state-of-the-arts.

The Fundamental Price of Secure Aggregation in Differentially Private Federated Learning

- Wei-Ning Chen, Christopher A Choquette Choo, Peter Kairouz, Ananda Theertha Suresh

- abstract: We consider the problem of training a d dimensional model with distributed differential privacy (DP) where secure aggregation (SecAgg) is used to ensure that the server only sees the noisy sum of n model updates in every training round. Taking into account the constraints imposed by SecAgg, we characterize the fundamental communication cost required to obtain the best accuracy achievable under ε central DP (i.e. under a fully trusted server and no communication constraints). Our results show that $\tilde{O}\lp \min(n^2\varepsilon^2, d) \rp$ bits per client are both sufficient and necessary, and this fundamental limit can be achieved by a linear scheme based on sparse random projections. This provides a significant improvement relative to state-of-the-art SecAgg distributed DP schemes which use $\tilde{O}(d\log(d/\varepsilon^2))$ bits per client. Empirically, we evaluate our proposed scheme on real-world federated learning tasks. We find that our theoretical analysis is well matched in practice. In particular, we show that we can reduce the communication cost to under 1.78 bits per parameter in realistic privacy settings without decreasing test-time performance. Our work hence theoretically and empirically specifies the fundamental price of using SecAgg.

Perfectly Balanced: Improving Transfer and Robustness of Supervised Contrastive Learning

- Mayee Chen, Daniel Y Fu, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, Christopher Re
- abstract: An ideal learned representation should display transferability and robustness. Supervised contrastive learning (SupCon) is a promising method for training accurate models, but produces representations that do not capture these properties due to class collapse—when all points in a class map to the same representation. Recent work suggests that "spreading out" these representations improves them, but the precise mechanism is poorly understood. We argue that creating spread alone is insufficient for better representations, since spread is invariant to permutations within classes. Instead, both the correct degree of spread and a mechanism for breaking this invariance are necessary. We first prove that adding a weighted class-conditional InfoNCE loss to SupCon controls the degree of spread. Next, we study three mechanisms to break permutation invariance: using a constrained encoder, adding a class-conditional autoencoder, and using data augmentation. We show that the latter two encourage clustering of latent subclasses under more realistic conditions than the former. Using these insights, we show that adding a properly-weighted class-conditional InfoNCE loss and a class-conditional autoencoder to SupCon achieves 11.1 points of lift on coarse-to-fine transfer across 5 standard datasets and 4.7 points on worst-group robustness on 3 datasets, setting state-of-the-art on CelebA by 11.5 points.

Strategies for Safe Multi-Armed Bandits with Logarithmic Regret and Risk

- Tianrui Chen, Aditya Gangrade, Venkatesh Saligrama
- abstract: We investigate a natural but surprisingly unstudied approach to the multi-armed bandit problem under safety risk constraints. Each arm is associated with an unknown law on safety risks and rewards, and the learner’s goal is to maximise reward whilst not playing unsafe arms, as determined by a given threshold on the mean risk. We formulate a pseudo-regret for this setting that enforces this safety constraint in a per-round way by softly penalising any violation, regardless of the gain in reward due to the same. This has practical relevance to scenarios such as clinical trials, where one must maintain safety for each round rather than in an aggregated sense. We describe doubly optimistic strategies for this scenario, which maintain optimistic indices for both safety risk and reward. We show that schema based on both frequentist and Bayesian indices satisfy tight gap-dependent logarithmic regret bounds, and further that these play unsafe arms only logarithmically many times in total. This theoretical analysis is complemented by simulation studies demonstrating the effectiveness of the proposed schema, and probing the domains in which their use is appropriate.

On the Sample Complexity of Learning Infinite-horizon Discounted Linear Kernel MDPs

- Yuanzhou Chen, Jiafan He, Quanquan Gu
- abstract: We study reinforcement learning for infinite-horizon discounted linear kernel MDPs, where the transition probability function is linear in a predefined feature mapping. Existing UCLK \citep{zhou2020provably} algorithm for this setting only has a regret guarantee, which cannot lead to a tight sample complexity bound. In this paper, we extend the uniform-PAC sample complexity from episodic setting to the infinite-horizon discounted setting, and propose a novel algorithm dubbed UPAC-UCLK that achieves an $\Tilde{O}\big(d^2/((1-\gamma)^4\epsilon^2)+1/((1-\gamma)^6\epsilon^2)\big)$ uniform-PAC sample complexity, where d is the dimension of the feature mapping, $\gamma \in(0,1)$ is the discount factor of the MDP and ϵ is the accuracy parameter. To the best of our knowledge, this is the first $\tilde{O}(1/\epsilon^2)$ sample complexity bound for learning infinite-horizon discounted MDPs with linear function approximation (without access to the generative model).

Streaming Algorithms for Support-Aware Histograms

- Justin Chen, Piotr Indyk, Tal Wagner
- abstract: Histograms, i.e., piece-wise constant approximations, are a popular tool used to represent data distributions. Traditionally, the difference between the histogram and the underlying distribution (i.e., the approximation error) is measured using the L_p norm, which sums the differences between the two functions over all items in the domain. Although useful in many applications, the drawback of this error measure is that it treats approximation errors of all items in the same way, irrespective of whether the mass of an item is important for the downstream application that uses the approximation. As a result, even relatively simple distributions cannot be approximated by succinct histograms without incurring large error. In this paper, we address this issue by adapting the definition of approximation so that only the errors of the items that belong to the support of the distribution are considered. Under this definition, we develop efficient 1-pass and 2-pass streaming algorithms that compute near-optimal histograms in sub-linear space. We also present lower bounds on the space complexity of this problem. Surprisingly, under this notion of error, there is an exponential gap in the space complexity of 1-pass and 2-pass streaming algorithms. Finally, we demonstrate the utility of our algorithms on a collection of real and synthetic data sets.

Improved No-Regret Algorithms for Stochastic Shortest Path with Linear MDP

- Liyu Chen, Rahul Jain, Haipeng Luo
- abstract: We introduce two new no-regret algorithms for the stochastic shortest path (SSP) problem with a linear MDP that significantly improve over the only existing results of (Vial et al., 2021). Our first algorithm is computationally efficient and achieves a regret bound $O(\sqrt{d^3B_{\star}^2T_{\star} K})$, where d is the dimension of the feature space, $B_{\star}$ and $T_{\star}$ are upper bounds of the expected costs and hitting time of the optimal policy respectively, and K is the number of episodes. The same algorithm with a slight modification also achieves logarithmic regret of order $O(\frac{d^3B_{\star}^4}{c_{\min}^2\text{\rm gap}{\min} }\ln^5\frac{dB{\star} K}{c_{\min}})$, where $\text{\rm gap}{\min}$ *is the minimum sub-optimality gap and* $c{\min}$ is the minimum cost over all state-action pairs. Our result is obtained by developing a simpler and improved analysis for the finite-horizon approximation of (Cohen et al., 2021) with a smaller approximation error, which might be of independent interest. On the other hand, using variance-aware confidence sets in a global optimization problem, our second algorithm is computationally inefficient but achieves the first “horizon-free” regret bound $O(d^{3.5}B_{\star}\sqrt{K})$ with no polynomial dependency on $T_{\star}$ or $1/c_{\min}$, almost matching the $\Omega(dB_{\star}\sqrt{K})$ lower bound from (Min et al., 2021).

Learning Infinite-horizon Average-reward Markov Decision Process with Constraints

- Liyu Chen, Rahul Jain, Haipeng Luo
- abstract: We study regret minimization for infinite-horizon average-reward Markov Decision Processes (MDPs) under cost constraints. We start by designing a policy optimization algorithm with carefully designed action-value estimator and bonus term, and show that for ergodic MDPs, our algorithm ensures $O(\sqrt{T})$ regret and constant constraint violation, where T is the total number of time steps. This strictly improves over the algorithm of (Singh et al., 2020), whose regret and constraint violation are both $O(T^{2/3})$. Next, we consider the most general class of weakly communicating MDPs. Through a finite-horizon approximation, we develop another algorithm with $O(T^{2/3})$ regret and constraint violation, which can be further

improved to $O(\sqrt{T})$ via a simple modification, albeit making the algorithm computationally inefficient. As far as we know, these are the first set of provable algorithms for weakly communicating MDPs with cost constraints.

Active Multi-Task Representation Learning

- Yifang Chen, Kevin Jamieson, Simon Du
- abstract: To leverage the power of big data from source domains and overcome the scarcity of target domain samples, representation learning based on multi-task pretraining has become a standard approach in many applications. However, large-scale pretraining is often computationally expensive and not affordable for small organizations. When there is only one target task, most source tasks can be irrelevant, and we can actively sample a subset of source data from the most To leverage the power of big data from source tasks and overcome the scarcity of the target task samples, representation learning based on multi-task pretraining has become a standard approach in many applications. However, up until now, choosing which source tasks to include in the multi-task learning has been more art than science. In this paper, we give the first formal study on resource task sampling by leveraging the techniques from active learning. We propose an algorithm that iteratively estimates the relevance of each source task to the target task and samples from each source task based on the estimated relevance. Theoretically, we show that for the linear representation class, to achieve the same error rate, our algorithm can save up to a textit{number of source tasks} factor in the source task sample complexity, compared with the naive uniform sampling from all source tasks. We also provide experiments on real-world computer vision datasets to illustrate the effectiveness of our proposed method on both linear and convolutional neural network representation classes. We believe our paper serves as an important initial step to bring techniques from active learning to representation learning.

On Collective Robustness of Bagging Against Data Poisoning

- Ruoxin Chen, Zenan Li, Jie Li, Junchi Yan, Chentao Wu
- abstract: Bootstrap aggregating (bagging) is an effective ensemble protocol, which is believed can enhance robustness by its majority voting mechanism. Recent works further prove the sample-wise robustness certificates for certain forms of bagging (e.g. partition aggregation). Beyond these particular forms, in this paper, we propose the first collective certification for general bagging to compute the tight robustness against the global poisoning attack. Specifically, we compute the maximum number of simultaneously changed predictions via solving a binary integer linear programming (BILP) problem. Then we analyze the robustness of vanilla bagging and give the upper bound of the tolerable poison budget. Based on this analysis, we propose hash bagging to improve the robustness of vanilla bagging almost for free. This is achieved by modifying the random subsampling in vanilla bagging to a hash-based deterministic subsampling, as a way of controlling the influence scope for each poisoning sample universally. Our extensive experiments show the notable advantage in terms of applicability and robustness. Our code is available at https://github.com/Emiyalzn/ICML22-CRB.

Online Active Regression

- Cheng Chen, Yi Li, Yiming Sun
- abstract: Active regression considers a linear regression problem where the learner receives a large number of data points but can only observe a small number of labels. Since online algorithms can deal with incremental training data and take advantage of low computational cost, we consider an online extension of the active regression problem: the learner receives data points one by one and immediately decides whether it should collect the corresponding labels. The goal is to efficiently maintain the regression of received data points with a small budget of label queries. We propose novel algorithms for this problem under ℓ_p loss where $p\in[1,2]$. To achieve a $(1+\epsilon)$-approximate solution, our proposed algorithms only requires $\tilde{\mathcal{O}}(d/poly(\epsilon))$ queries of labels. The numerical results verify our theoretical results and show that our methods have comparable performance with offline active regression algorithms.

Selling Data To a Machine Learner: Pricing via Costly Signaling

- Junjie Chen, Minming Li, Haifeng Xu
- abstract: We consider a new problem of selling data to a machine learner who looks to purchase data to train his machine learning model. A key challenge in this setup is that neither the seller nor the machine learner knows the true quality of data. When designing a revenue-maximizing mechanism, a data seller faces the tradeoff between the cost and precision of data quality estimation. To address this challenge, we study a natural class of mechanisms that price data via costly signaling. Motivated by the assumption of i.i.d. data points as in classic machine learning models, we first consider selling homogeneous data and derive an optimal selling mechanism. We then turn to the sale of heterogeneous data, motivated by the sale of multiple data sets, and show that 1) on the negative side, it is NP-hard to approximate the optimal mechanism within a constant ratio e/(e+1) + o(1); while 2) on the positive side, there is a 1/k-approximate algorithm, where k is the number of the machine learner's private types.

ME-GAN: Learning Panoptic Electrocardio Representations for Multi-view ECG Synthesis Conditioned on Heart Diseases

- Jintai Chen, Kuanlun Liao, Kun Wei, Haochao Ying, Danny Z Chen, Jian Wu
- abstract: Electrocardiogram (ECG) is a widely used non-invasive diagnostic tool for heart diseases. Many studies have devised ECG analysis models (e.g., classifiers) to assist diagnosis. As an upstream task, researches have built generative models to synthesize ECG data, which are beneficial to providing training samples, privacy protection, and annotation reduction. However, previous generative methods for ECG often neither synthesized multi-view data, nor dealt with heart disease conditions. In this paper, we propose a novel disease-aware generative adversarial network for multi-view ECG synthesis called ME-GAN, which attains panoptic electrocardio representations conditioned on heart diseases and projects the representations onto multiple standard views to yield ECG signals. Since ECG manifestations of heart diseases are often localized in specific waveforms, we propose a new "mixup normalization" to inject disease information precisely into suitable locations. In addition, we propose a "view discriminator" to revert disordered ECG views into a pre-determined order, supervising the generator to obtain ECG representing correct view characteristics. Besides, a new metric, rFID, is presented to assess the quality of the synthesized ECG signals. Comprehensive experiments verify that our ME-GAN performs well on multi-view ECG signal synthesis with trusty morbid manifestations.

Weisfeiler-Lehman Meets Gromov-Wasserstein

- Samantha Chen, Sunhyuk Lim, Facundo Memoli, Zhengchao Wan, Yusu Wang
- abstract: The Weisfeiler-Lehman (WL) test is a classical procedure for graph isomorphism testing. The WL test has also been widely used both for designing graph kernels and for analyzing graph neural networks. In this paper, we propose the Weisfeiler-Lehman (WL) distance, a notion of distance between labeled measure Markov chains (LMMCs), of which labeled graphs are special cases. The WL distance is polynomial time computable and is also compatible with the WL test in the sense that the former is positive if and only if the WL test can distinguish the two involved graphs. The WL distance captures and compares subtle structures of the underlying LMMCs and, as a consequence of this, it is more discriminating than the distance between graphs used for defining the state-of-the-art Wasserstein Weisfeiler-Lehman graph kernel. Inspired by the structure of the WL distance we identify a neural network architecture on LMMCs which turns out to be universal w.r.t. continuous functions defined on the space of all LMMCs (which includes all graphs) endowed with the WL distance. Finally, the WL distance turns out to be stable w.r.t. a natural variant of the Gromov-Wasserstein (GW) distance for comparing metric Markov chains that we identify. Hence, the WL distance can also be construed as a polynomial time lower bound for the GW distance which is in general NP-hard to compute.

On Non-local Convergence Analysis of Deep Linear Networks

- Kun Chen, Dachao Lin, Zhihua Zhang
- abstract: In this paper, we study the non-local convergence properties of deep linear networks. Specifically, under the quadratic loss, we consider optimizing deep linear networks in which there is at least a layer with only one neuron. We describe the convergent point of trajectories with an arbitrary balanced starting point under gradient flow, including the paths which converge to one of the saddle points. We also show specific convergence rates of trajectories that converge to the global minimizers by stages. We conclude that the rates vary from polynomial to linear. As far as we know, our results are the first to give a non-local analysis of deep linear neural networks with arbitrary balanced initialization, rather than the lazy training regime which has dominated the literature on neural networks or the restricted benign initialization.

Flow-based Recurrent Belief State Learning for POMDPs

- Xiaoyu Chen, Yao Mark Mu, Ping Luo, Shengbo Li, Jianyu Chen
- abstract: Partially Observable Markov Decision Process (POMDP) provides a principled and generic framework to model real world sequential decision making processes but yet remains unsolved, especially for high dimensional continuous space and unknown models. The main challenge lies in how to accurately obtain the belief state, which is the probability distribution over the unobservable environment states given historical information. Accurately calculating this belief state is a precondition for obtaining an optimal policy of POMDPs. Recent advances in deep learning techniques show great potential to learn good belief states. However, existing methods can only learn approximated distribution with limited flexibility. In this paper, we introduce the \textbf{F}l\textbf{O}w-based \textbf{R}ecurrent \textbf{BE}lief \textbf{S}tate model (FORBES), which incorporates normalizing flows into the variational inference to learn general continuous belief states for POMDPs. Furthermore, we show that the learned belief states can be plugged into downstream RL algorithms to improve performance. In experiments, we show that our methods successfully capture the complex belief states that enable multi-modal predictions as well as high quality reconstructions, and results on challenging visual-motor control tasks show that our method achieves superior performance and sample efficiency.

Structure-Aware Transformer for Graph Representation Learning

- Dexiong Chen, Leslie O'Bray, Karsten Borgwardt
- abstract: The Transformer architecture has gained growing attention in graph representation learning recently, as it naturally overcomes several limitations of graph neural networks (GNNs) by avoiding their strict structural inductive biases and instead only encoding the graph structure via positional encoding. Here, we show that the node representations generated by the Transformer with positional encoding do not necessarily capture structural similarity between them. To address this issue, we propose the Structure-Aware Transformer, a class of simple and flexible graph Transformers built upon a new self-attention mechanism. This new self-attention incorporates structural information into the original self-attention by extracting a subgraph representation rooted at each node before computing the attention. We propose several methods for automatically generating the subgraph representation and show theoretically that the resulting representations are at least as expressive as the subgraph representations. Empirically, our method achieves state-of-the-art performance on five graph prediction benchmarks. Our structure-aware framework can leverage any existing GNN to extract the subgraph representation, and we show that it systematically improves performance relative to the base GNN model, successfully combining the advantages of GNNs and Transformers. Our code is available at https://github.com/BorgwardtLab/SAT.

The Poisson Binomial Mechanism for Unbiased Federated Learning with Secure Aggregation

- Wei-Ning Chen, Ayfer Ozgur, Peter Kairouz
- abstract: We introduce the Poisson Binomial mechanism (PBM), a discrete differential privacy mechanism for distributed mean estimation (DME) with applications to federated learning and analytics. We provide a tight analysis of its privacy guarantees, showing that it achieves the same privacy-accuracy trade-offs as the continuous Gaussian mechanism. Our analysis is based on a novel bound on the Rényi divergence of two Poisson binomial distributions that may be of independent interest. Unlike previous discrete DP schemes based on additive noise, our mechanism encodes local information into a parameter of the binomial distribution, and hence the output distribution is discrete with bounded support. Moreover, the support does not increase as the privacy budget goes to zero as in the case of additive schemes which require the addition of more noise to achieve higher privacy; on the contrary, the support becomes smaller as eps goes to zero. The bounded support enables us to combine our mechanism with secure aggregation (SecAgg), a multi-party cryptographic protocol, without the need of performing modular clipping which results in an unbiased estimator of the sum of the local vectors. This in turn allows us to apply it in the private FL setting and provide an upper bound on the convergence rate of the SGD algorithm. Moreover, since the support of the output distribution becomes smaller as $\varepsilon \ra 0$, the communication cost of our scheme decreases with the privacy constraint ε, outperforming all previous distributed DP schemes based on additive noise in the high privacy or low communication regimes.

Learning Mixtures of Linear Dynamical Systems

- Yanxi Chen, H. Vincent Poor
- abstract: We study the problem of learning a mixture of multiple linear dynamical systems (LDSs) from unlabeled short sample trajectories, each generated by one of the LDS models. Despite the wide applicability of mixture models for time-series data, learning algorithms that come with end-to-end performance guarantees are largely absent from existing literature. There are multiple sources of technical challenges, including but not limited to (1) the presence of latent variables (i.e. the unknown labels of trajectories); (2) the possibility that the sample trajectories might have lengths much smaller than the dimension d of the LDS models; and (3) the complicated temporal dependence inherent to time-series data. To tackle these challenges, we develop a two-stage meta-algorithm, which is guaranteed to efficiently recover each ground-truth LDS model up to error $\tilde{O}(\sqrt{d/T})$, where T is the total sample size. We validate our theoretical studies with numerical experiments, confirming the efficacy of the proposed algorithm.

On Well-posedness and Minimax Optimal Rates of Nonparametric Q-function Estimation in Off-policy Evaluation

- Xiaohong Chen, Zhengling Qi
- abstract: We study the off-policy evaluation (OPE) problem in an infinite-horizon Markov decision process with continuous states and actions. We recast the Q-function estimation into a special form of the nonparametric instrumental variables (NPIV) estimation problem. We first show that under one mild condition the NPIV formulation of Q-function estimation is well-posed in the sense of L^2-measure of ill-posedness with respect to the data generating distribution, bypassing a strong assumption on the discount factor γ imposed in the recent literature for obtaining the L^2 convergence rates of various Q-function estimators. Thanks to this new well-posed property, we derive the first minimax lower bounds for the convergence rates of nonparametric estimation of Q-function and its derivatives in both sup-norm and L^2-norm, which are shown to be the same as those for the classical nonparametric regression (Stone, 1982). We then propose a sieve two-stage least squares estimator and establish its rate-optimality in both norms under some mild conditions. Our general results on the well-posedness and the minimax lower bounds are of independent interest to study not only other nonparametric estimators for Q-function but also efficient estimation on the value of any target policy in off-policy settings.

Faster Fundamental Graph Algorithms via Learned Predictions

- Justin Chen, Sandeep Silwal, Ali Vakilian, Fred Zhang

- abstract: We consider the question of speeding up classic graph algorithms with machine-learned predictions. In this model, algorithms are furnished with extra advice learned from past or similar instances. Given the additional information, we aim to improve upon the traditional worst-case run-time guarantees. Our contributions are the following: (i) We give a faster algorithm for minimum-weight bipartite matching via learned duals, improving the recent result by Dinitz, Im, Lavastida, Moseley and Vassilvitskii (NeurIPS, 2021); (ii) We extend the learned dual approach to the single-source shortest path problem (with negative edge lengths), achieving an almost linear runtime given sufficiently accurate predictions which improves upon the classic fastest algorithm due to Goldberg (SIAM J. Comput., 1995); (iii) We provide a general reduction-based framework for learning-based graph algorithms, leading to new algorithms for degree-constrained subgraph and minimum-cost 0-1 flow, based on reductions to bipartite matching and the shortest path problem. Finally, we give a set of general learnability theorems, showing that the predictions required by our algorithms can be efficiently learned in a PAC fashion.

Improve Single-Point Zeroth-Order Optimization Using High-Pass and Low-Pass Filters

- Xin Chen, Yujie Tang, Na Li
- abstract: Single-point zeroth-order optimization (SZO) is useful in solving online black-box optimization and control problems in time-varying environments, as it queries the function value only once at each time step. However, the vanilla SZO method is known to suffer from a large estimation variance and slow convergence, which seriously limits its practical application. In this work, we borrow the idea of high-pass and low-pass filters from extremum seeking control (continuous-time version of SZO) and develop a novel SZO method called HLF-SZO by integrating these filters. It turns out that the high-pass filter coincides with the residual feedback method, and the low-pass filter can be interpreted as the momentum method. As a result, the proposed HLF-SZO achieves a much smaller variance and much faster convergence than the vanilla SZO method, and empirically outperforms the residual-feedback SZO method, which are verified via extensive numerical experiments.

Deep Variational Graph Convolutional Recurrent Network for Multivariate Time Series Anomaly Detection

- Wenchao Chen, Long Tian, Bo Chen, Liang Dai, Zhibin Duan, Mingyuan Zhou
- abstract: Anomaly detection within multivariate time series (MTS) is an essential task in both data mining and service quality management. Many recent works on anomaly detection focus on designing unsupervised probabilistic models to extract robust normal patterns of MTS. In this paper, we model sensor dependency and stochasticity within MTS by developing an embedding-guided probabilistic generative network. We combine it with adaptive variational graph convolutional recurrent network %and get variational GCRN (VGCRN) to model both spatial and temporal fine-grained correlations in MTS. To explore hierarchical latent representations, we further extend VGCRN into a deep variational network, which captures multilevel information at different layers and is robust to noisy time series. Moreover, we develop an upward-downward variational inference scheme that considers both forecasting-based and reconstruction-based losses, achieving an accurate posterior approximation of latent variables with better MTS representations. The experiments verify the superiority of the proposed method over current state-of-the-art methods.

Auxiliary Learning with Joint Task and Data Scheduling

- Hong Chen, Xin Wang, Chaoyu Guan, Yue Liu, Wenwu Zhu
- abstract: Existing auxiliary learning approaches only consider the relationships between the target task and the auxiliary tasks, ignoring the fact that data samples within an auxiliary task could contribute differently to the target task, which results in inefficient auxiliary information usage and non-robustness to data noise. In this paper, we propose to learn a joint task and data schedule for auxiliary learning, which captures the importance of different data samples in each auxiliary task to the target task. However, learning such a joint schedule is challenging due to the large number of additional parameters required for the schedule. To tackle the challenge, we propose a joint task and data scheduling (JTDS) model for auxiliary learning. The JTDS model captures the joint task-data importance through a task-data scheduler, which creates a mapping from task, feature and label information to the schedule in a parameter-efficient way. Particularly, we formulate the scheduler and the task learning process as a bi-level optimization problem. In the lower optimization, the task learning model is updated with the scheduled gradient, while in the upper optimization, the task-data scheduler is updated with the implicit gradient. Experimental results show that our JTDS model significantly outperforms the state-of-the-art methods under supervised, semi-supervised and corrupted label settings.

Optimization-Induced Graph Implicit Nonlinear Diffusion

- Qi Chen, Yifei Wang, Yisen Wang, Jiansheng Yang, Zhouchen Lin
- abstract: Due to the over-smoothing issue, most existing graph neural networks can only capture limited dependencies with their inherently finite aggregation layers. To overcome this limitation, we propose a new kind of graph convolution, called Graph Implicit Nonlinear Diffusion (GIND), which implicitly has access to infinite hops of neighbors while adaptively aggregating features with nonlinear diffusion to prevent over-smoothing. Notably, we show that the learned representation can be formalized as the minimizer of an explicit convex optimization objective. With this property, we can theoretically characterize the equilibrium of our GIND from an optimization perspective. More interestingly, we can induce new structural variants by modifying the corresponding optimization objective. To be specific, we can embed prior properties to the equilibrium, as well as introducing skip connections to promote training stability. Extensive experiments show that GIND is good at capturing long-range dependencies, and performs well on both homophilic and heterophilic graphs with nonlinear diffusion. Moreover, we show that the optimization-induced variants of our models can boost the performance and improve training stability and efficiency as well. As a result, our GIND obtains significant improvements on both node-level and graph-level tasks.

Robust Meta-learning with Sampling Noise and Label Noise via Eigen-Reptile

- Dong Chen, Lingfei Wu, Siliang Tang, Xiao Yun, Bo Long, Yueting Zhuang
- abstract: Recent years have seen a surge of interest in meta-learning techniques for tackling the few-shot learning (FSL) problem. However, the meta-learner is prone to overfitting since there are only a few available samples, which can be identified as sampling noise on a clean dataset. Besides, when handling the data with noisy labels, the meta-learner could be extremely sensitive to label noise on a corrupted dataset. To address these two challenges, we present Eigen-Reptile (ER) that updates the meta-parameters with the main direction of historical task-specific parameters. Specifically, the main direction is computed in a fast way, where the scale of the calculated matrix is related to the number of gradient steps for the specific task instead of the number of parameters. Furthermore, to obtain a more accurate main direction for Eigen-Reptile in the presence of many noisy labels, we further propose Introspective Self-paced Learning (ISPL). We have theoretically and experimentally demonstrated the soundness and effectiveness of the proposed Eigen-Reptile and ISPL. Particularly, our experiments on different tasks show that the proposed method is able to outperform or achieve highly competitive performance compared with other gradient-based methods with or without noisy labels. The code and data for the proposed method are provided for research purposes https://github.com/Anfeather/Eigen-Reptile.

Adaptive Model Design for Markov Decision Process

- Siyu Chen, Donglin Yang, Jiayang Li, Senmiao Wang, Zhuoran Yang, Zhaoran Wang
- abstract: In a Markov decision process (MDP), an agent interacts with the environment via perceptions and actions. During this process, the agent aims to maximize its own gain. Hence, appropriate regulations are often required, if we hope to take the external costs/benefits of its actions into consideration. In this paper, we study how to regulate such an agent by redesigning model parameters that can affect the rewards and/or the transition kernels. We formulate this problem as a bilevel program, in which the lower-level MDP is regulated by the upper-level model designer. To solve the resulting problem, we

develop a scheme that allows the designer to iteratively predict the agent's reaction by solving the MDP and then adaptively update model parameters based on the predicted reaction. The algorithm is first theoretically analyzed and then empirically tested on several MDP models arising in economics and robotics.

State Transition of Dendritic Spines Improves Learning of Sparse Spiking Neural Networks

- Yanqi Chen, Zhaofei Yu, Wei Fang, Zhengyu Ma, Tiejun Huang, Yonghong Tian
- abstract: Spiking Neural Networks (SNNs) are considered a promising alternative to Artificial Neural Networks (ANNs) for their event-driven computing paradigm when deployed on energy-efficient neuromorphic hardware. Recently, deep SNNs have shown breathtaking performance improvement through cutting-edge training strategy and flexible structure, which also scales up the number of parameters and computational burdens in a single network. Inspired by the state transition of dendritic spines in the filopodial model of spinogenesis, we model different states of SNN weights, facilitating weight optimization for pruning. Furthermore, the pruning speed can be regulated by using different functions describing the growing threshold of state transition. We organize these techniques as a dynamic pruning algorithm based on nonlinear reparameterization mapping from spine size to SNN weights. Our approach yields sparse deep networks on the large-scale dataset (SEW ResNet18 on ImageNet) while maintaining state-of-the-art low performance loss (3% at 88.8% sparsity) compared to existing pruning methods on directly trained SNNs. Moreover, we find out pruning speed regulation while learning is crucial to avoiding disastrous performance degradation at the final stages of training, which may shed light on future work on SNN pruning.

Efficient Online ML API Selection for Multi-Label Classification Tasks

- Lingjiao Chen, Matei Zaharia, James Zou
- abstract: Multi-label classification tasks such as OCR and multi-object recognition are a major focus of the growing machine learning as a service industry. While many multi-label APIs are available, it is challenging for users to decide which API to use for their own data and budget, due to the heterogeneity in their prices and performance. Recent work has shown how to efficiently select and combine single label APIs to optimize performance and cost. However, its computation cost is exponential in the number of labels, and is not suitable for settings like OCR. In this work, we propose FrugalMCT, a principled framework that adaptively selects the APIs to use for different data in an online fashion while respecting the user's budget. It allows combining ML APIs' predictions for any single data point, and selects the best combination based on an accuracy estimator. We run systematic experiments using ML APIs from Google, Microsoft, Amazon, IBM, Tencent, and other providers for tasks including multi-label image classification, scene text recognition, and named entity recognition. Across these tasks, FrugalMCT can achieve over 90% cost reduction while matching the accuracy of the best single API, or up to 8% better accuracy while matching the best API's cost.

Data-Efficient Double-Win Lottery Tickets from Robust Pre-training

- Tianlong Chen, Zhenyu Zhang, Sijia Liu, Yang Zhang, Shiyu Chang, Zhangyang Wang
- abstract: Pre-training serves as a broadly adopted starting point for transfer learning on various downstream tasks. Recent investigations of lottery tickets hypothesis (LTH) demonstrate such enormous pre-trained models can be replaced by extremely sparse subnetworks (a.k.a. matching subnetworks) without sacrificing transferability. However, practical security-crucial applications usually pose more challenging requirements beyond standard transfer, which also demand these subnetworks to overcome adversarial vulnerability. In this paper, we formulate a more rigorous concept, Double-Win Lottery Tickets, in which a located subnetwork from a pre-trained model can be independently transferred on diverse downstream tasks, to reach BOTH the same standard and robust generalization, under BOTH standard and adversarial training regimes, as the full pre-trained model can do. We comprehensively examine various pre-training mechanisms and find that robust pre-training tends to craft sparser double-win lottery tickets with superior performance over the standard counterparts. For example, on downstream CIFAR-10/100 datasets, we identify double-win matching subnetworks with the standard, fast adversarial, and adversarial pre-training from ImageNet, at 89.26%/73.79%, 89.26%/79.03%, and 91.41%/83.22% sparsity, respectively. Furthermore, we observe the obtained double-win lottery tickets can be more data-efficient to transfer, under practical data-limited (e.g., 1% and 10%) downstream schemes. Our results show that the benefits from robust pre-training are amplified by the lottery ticket scheme, as well as the data-limited transfer setting. Codes are available at https://github.com/VITA-Group/Double-Win-LTH.

Linearity Grafting: Relaxed Neuron Pruning Helps Certifiable Robustness

- Tianlong Chen, Huan Zhang, Zhenyu Zhang, Shiyu Chang, Sijia Liu, Pin-Yu Chen, Zhangyang Wang
- abstract: Certifiable robustness is a highly desirable property for adopting deep neural networks (DNNs) in safety-critical scenarios, but often demands tedious computations to establish. The main hurdle lies in the massive amount of non-linearity in large DNNs. To trade off the DNN expressiveness (which calls for more non-linearity) and robustness certification scalability (which prefers more linearity), we propose a novel solution to strategically manipulate neurons, by "grafting" appropriate levels of linearity. The core of our proposal is to first linearize insignificant ReLU neurons, to eliminate the non-linear components that are both redundant for DNN performance and harmful to its certification. We then optimize the associated slopes and intercepts of the replaced linear activations for restoring model performance while maintaining certifiability. Hence, typical neuron pruning could be viewed as a special case of grafting a linear function of the fixed zero slopes and intercept, that might overly restrict the network flexibility and sacrifice its performance. Extensive experiments on multiple datasets and network backbones show that our linearity grafting can (1) effectively tighten certified bounds; (2) achieve competitive certifiable robustness without certified robust training (i.e., over 30% improvements on CIFAR-10 models); and (3) scale up complete verification to large adversarially trained models with 17M parameters. Codes are available at https://github.com/VITA-Group/Linearity-Grafting.

Human-in-the-loop: Provably Efficient Preference-based Reinforcement Learning with General Function Approximation

- Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, Liwei Wang
- abstract: We study human-in-the-loop reinforcement learning (RL) with trajectory preferences, where instead of receiving a numeric reward at each step, the RL agent only receives preferences over trajectory pairs from a human overseer. The goal of the RL agent is to learn the optimal policy which is most preferred by the human overseer. Despite the empirical success in various real-world applications, the theoretical understanding of preference-based RL (PbRL) is only limited to the tabular case. In this paper, we propose the first optimistic model-based algorithm for PbRL with general function approximation, which estimates the model using value-targeted regression and calculates the exploratory policies by solving an optimistic planning problem. We prove that our algorithm achieves the regret bound of $\tilde{O} (\operatorname{poly}(d H) \sqrt{K})$, where d is the complexity measure of the transition and preference model depending on the Eluder dimension and log-covering numbers, H is the planning horizon, K is the number of episodes, and $\tilde O(\cdot)$ omits logarithmic terms. Our lower bound indicates that our algorithm is near-optimal when specialized to the linear setting. Furthermore, we extend the PbRL problem by formulating a novel problem called RL with n-wise comparisons, and provide the first sample-efficient algorithm for this new setting. To the best of our knowledge, this is the first theoretical result for PbRL with (general) function approximation.

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

- Ziyi Chen, Yi Zhou, Rong-Rong Chen, Shaofeng Zou
- abstract: Actor-critic (AC) algorithms have been widely used in decentralized multi-agent systems to learn the optimal joint control policy. However, existing decentralized AC algorithms either need to share agents' sensitive information or lack communication-efficiency. In this work, we develop decentralized AC and natural AC (NAC) algorithms that avoid sharing agents' local information and are sample and communication-efficient. In both

algorithms, agents share only noisy rewards and use mini-batch local policy gradient updates to ensure high sample and communication efficiency. Particularly for decentralized NAC, we develop a decentralized Markovian SGD algorithm with an adaptive mini-batch size to efficiently compute the natural policy gradient. Under Markovian sampling and linear function approximation, we prove that the proposed decentralized AC and NAC algorithms achieve the state-of-the-art sample complexities $\mathcal{O}(\epsilon^{-2}\ln\epsilon^{-1})$ and $\mathcal{O}(\epsilon^{-3}\ln\epsilon^{-1})$, respectively, and achieve an improved communication complexity $\mathcal{O}(\epsilon^{-1}\ln\epsilon^{-1})$. Numerical experiments demonstrate that the proposed algorithms achieve lower sample and communication complexities than the existing decentralized AC algorithms.

Task-aware Privacy Preservation for Multi-dimensional Data

- Jiangnan Cheng, Ao Tang, Sandeep Chinchali
- abstract: Local differential privacy (LDP) can be adopted to anonymize richer user data attributes that will be input to sophisticated machine learning (ML) tasks. However, today's LDP approaches are largely task-agnostic and often lead to severe performance loss – they simply inject noise to all data attributes according to a given privacy budget, regardless of what features are most relevant for the ultimate task. In this paper, we address how to significantly improve the ultimate task performance with multi-dimensional user data by considering a task-aware privacy preservation problem. The key idea is to use an encoder-decoder framework to learn (and anonymize) a task-relevant latent representation of user data. We obtain an analytical near-optimal solution for the linear setting with mean-squared error (MSE) task loss. We also provide an approximate solution through a gradient-based learning algorithm for general nonlinear cases. Extensive experiments demonstrate that our task-aware approach significantly improves ultimate task accuracy compared to standard benchmark LDP approaches with the same level of privacy guarantee.

Adversarially Trained Actor Critic for Offline Reinforcement Learning

- Ching-An Cheng, Tengyang Xie, Nan Jiang, Alekh Agarwal
- abstract: We propose Adversarially Trained Actor Critic (ATAC), a new model-free algorithm for offline reinforcement learning (RL) under insufficient data coverage, based on the concept of relative pessimism. ATAC is designed as a two-player Stackelberg game framing of offline RL: A policy actor competes against an adversarially trained value critic, who finds data-consistent scenarios where the actor is inferior to the data-collection behavior policy. We prove that, when the actor attains no regret in the two-player game, running ATAC produces a policy that provably 1) outperforms the behavior policy over a wide range of hyperparameters that control the degree of pessimism, and 2) competes with the best policy covered by data with appropriately chosen hyperparameters. Compared with existing works, notably our framework offers both theoretical guarantees for general function approximation and a deep RL implementation scalable to complex environments and large datasets. In the D4RL benchmark, ATAC consistently outperforms state-of-the-art offline RL algorithms on a range of continuous control tasks.

Quantum-Inspired Algorithms from Randomized Numerical Linear Algebra

- Nadiia Chepurko, Kenneth Clarkson, Lior Horesh, Honghao Lin, David Woodruff
- abstract: We create classical (non-quantum) dynamic data structures supporting queries for recommender systems and least-squares regression that are comparable to their quantum analogues. De-quantizing such algorithms has received a flurry of attention in recent years; we obtain sharper bounds for these problems. More significantly, we achieve these improvements by arguing that the previous quantum-inspired algorithms for these problems are doing leverage or ridge-leverage score sampling in disguise; these are powerful and standard techniques in randomized numerical linear algebra. With this recognition, we are able to employ the large body of work in numerical linear algebra to obtain algorithms for these problems that are simpler or faster (or both) than existing approaches. Our experiments demonstrate that the proposed data structures also work well on real-world datasets.

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

- Victor Chernozhukov, Whitney Newey, Víctor M Quintas-Martínez, Vasilis Syrgkanis
- abstract: Many causal and policy effects of interest are defined by linear functionals of high-dimensional or non-parametric regression functions. $\sqrt{n}$-consistent and asymptotically normal estimation of the object of interest requires debiasing to reduce the effects of regularization and/or model selection on the object of interest. Debiasing is typically achieved by adding a correction term to the plug-in estimator of the functional, which leads to properties such as semi-parametric efficiency, double robustness, and Neyman orthogonality. We implement an automatic debiasing procedure based on automatically learning the Riesz representation of the linear functional using Neural Nets and Random Forests. Our method only relies on black-box evaluation oracle access to the linear functional and does not require knowledge of its analytic form. We propose a multitasking Neural Net debiasing method with stochastic gradient descent minimization of a combined Riesz representer and regression loss, while sharing representation layers for the two functions. We also propose a Random Forest method which learns a locally linear representation of the Riesz function. Even though our method applies to arbitrary functionals, we experimentally find that it performs well compared to the state of art neural net based algorithm of Shi et al. (2019) for the case of the average treatment effect functional. We also evaluate our method on the problem of estimating average marginal effects with continuous treatments, using semi-synthetic data of gasoline price changes on gasoline demand.

Self-supervised learning with random-projection quantizer for speech recognition

- Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, Yonghui Wu
- abstract: We present a simple and effective self-supervised learning approach for speech recognition. The approach learns a model to predict the masked speech signals, in the form of discrete labels generated with a random-projection quantizer. In particular the quantizer projects speech inputs with a randomly initialized matrix, and does a nearest-neighbor lookup in a randomly-initialized codebook. Neither the matrix nor the codebook are updated during self-supervised learning. Since the random-projection quantizer is not trained and is separated from the speech recognition model, the design makes the approach flexible and is compatible with universal speech recognition architecture. On LibriSpeech our approach achieves similar word-error-rates as previous work using self-supervised learning with non-streaming models, and provides lower word-error-rates than previous work with streaming models. On multilingual tasks the approach also provides significant improvement over wav2vec 2.0 and w2v-BERT.

Discrete Probabilistic Inverse Optimal Transport

- Wei-Ting Chiu, Pei Wang, Patrick Shafto
- abstract: Inverse Optimal Transport (IOT) studies the problem of inferring the underlying cost that gives rise to an observation on coupling two probability measures. Couplings appear as the outcome of matching sets (e.g. dating) and moving distributions (e.g. transportation). Compared to Optimal transport (OT), the mathematical theory of IOT is undeveloped. We formalize and systematically analyze the properties of IOT using tools from the study of entropy-regularized OT. Theoretical contributions include characterization of the manifold of cross-ratio equivalent costs, the implications of model priors, and derivation of an MCMC sampler. Empirical contributions include visualizations of cross-ratio equivalent effect on basic examples, simulations validating theoretical results and experiments on real world data.

Selective Network Linearization for Efficient Private Inference

- Minsu Cho, Ameya Joshi, Brandon Reagen, Siddharth Garg, Chinmay Hegde

- abstract: Private inference (PI) enables inferences directly on cryptographically secure data. While promising to address many privacy issues, it has seen limited use due to extreme runtimes. Unlike plaintext inference, where latency is dominated by FLOPs, in PI non-linear functions (namely ReLU) are the bottleneck. Thus, practical PI demands novel ReLU-aware optimizations. To reduce PI latency we propose a gradient-based algorithm that selectively linearizes ReLUs while maintaining prediction accuracy. We evaluate our algorithm on several standard PI benchmarks. The results demonstrate up to $4.25%$ more accuracy (iso-ReLU count at 50K) or $2.2\times$ less latency (iso-accuracy at 70%) than the current state of the art and advance the Pareto frontier across the latency-accuracy space. To complement empirical results, we present a “no free lunch" theorem that sheds light on how and when network linearization is possible while maintaining prediction accuracy.

From block-Toeplitz matrices to differential equations on graphs: towards a general theory for scalable masked Transformers

- Krzysztof Choromanski, Han Lin, Haoxian Chen, Tianyi Zhang, Arijit Sehanobish, Valerii Likhosherstov, Jack Parker-Holder, Tamas Sarlos, Adrian Weller, Thomas Weingarten
- abstract: In this paper we provide, to the best of our knowledge, the first comprehensive approach for incorporating various masking mechanisms into Transformers architectures in a scalable way. We show that recent results on linear causal attention (Choromanski et al., 2021) and log-linear RPE-attention (Luo et al., 2021) are special cases of this general mechanism. However by casting the problem as a topological (graph-based) modulation of unmasked attention, we obtain several results unknown before, including efficient d-dimensional RPE-masking and graph-kernel masking. We leverage many mathematical techniques ranging from spectral analysis through dynamic programming and random walks to new algorithms for solving Markov processes on graphs. We provide a corresponding empirical evaluation.

Shuffle Private Linear Contextual Bandits

- Sayak Ray Chowdhury, Xingyu Zhou
- abstract: Differential privacy (DP) has been recently introduced to linear contextual bandits to formally address the privacy concerns in its associated personalized services to participating users (e.g., recommendations). Prior work largely focus on two trust models of DP – the central model, where a central server is responsible for protecting users’ sensitive data, and the (stronger) local model, where information needs to be protected directly on users’ side. However, there remains a fundamental gap in the utility achieved by learning algorithms under these two privacy models, e.g., if all users are unique within a learning horizon T, $\widetilde{O}(\sqrt{T})$ regret in the central model as compared to $\widetilde{O}(T^{3/4})$ regret in the local model. In this work, we aim to achieve a stronger model of trust than the central model, while suffering a smaller regret than the local model by considering recently popular shuffle model of privacy. We propose a general algorithmic framework for linear contextual bandits under the shuffle trust model, where there exists a trusted shuffler – in between users and the central server– that randomly permutes a batch of users data before sending those to the server. We then instantiate this framework with two specific shuffle protocols – one relying on privacy amplification of local mechanisms, and another incorporating a protocol for summing vectors and matrices of bounded norms. We prove that both these instantiations lead to regret guarantees that significantly improve on that of the local model, and can potentially be of the order $\widetilde{O}(T^{3/5})$ if all users are unique. We also verify this regret behavior with simulations on synthetic data. Finally, under the practical scenario of non-unique users, we show that the regret of our shuffle private algorithm scale as $\widetilde{O}(T^{2/3})$, which matches what the central model could achieve in this case.

DNA: Domain Generalization with Diversified Neural Averaging

- Xu Chu, Yujie Jin, Wenwu Zhu, Yasha Wang, Xin Wang, Shanghang Zhang, Hong Mei
- abstract: The inaccessibility of the target domain data causes domain generalization (DG) methods prone to forget target discriminative features, and challenges the pervasive theme in existing literature in pursuing a single classifier with an ideal joint risk. In contrast, this paper investigates model misspecification and attempts to bridge DG with classifier ensemble theoretically and methodologically. By introducing a pruned Jensen-Shannon (PJS) loss, we show that the target square-root risk w.r.t. the PJS loss of the ρ-ensemble (the averaged classifier weighted by a quasi-posterior ρ) is bounded by the averaged source square-root risk of the Gibbs classifiers. We derive a tighter bound by enforcing a positive principled diversity measure of the classifiers. We give a PAC-Bayes upper bound on the target square-root risk of the ρ-ensemble. Methodologically, we propose a diversified neural averaging (DNA) method for DG, which optimizes the proposed PAC-Bayes bound approximately. The DNA method samples Gibbs classifiers transversely and longitudinally by simultaneously considering the dropout variational family and optimization trajectory. The ρ-ensemble is approximated by averaging the longitudinal weights in a single run with dropout shut down, ensuring a fast ensemble with low computational overhead. Empirically, the proposed DNA method achieves the state-of-the-art classification performance on standard DG benchmark datasets.

TPC: Transformation-Specific Smoothing for Point Cloud Models

- Wenda Chu, Linyi Li, Bo Li
- abstract: Point cloud models with neural network architectures have achieved great success and been widely used in safety-critical applications, such as Lidar-based recognition systems in autonomous vehicles. However, such models are shown vulnerable against adversarial attacks which aim to apply stealthy semantic transformations such as rotation and tapering to mislead model predictions. In this paper, we propose a transformation-specific smoothing framework TPC, which provides tight and scalable robustness guarantees for point cloud models against semantic transformation attacks. We first categorize common 3D transformations into two categories: composable (e.g., rotation) and indirectly composable (e.g., tapering), and we present generic robustness certification strategies for both categories. We then specify unique certification protocols for a range of specific semantic transformations and derive strong robustness guarantees. Extensive experiments on several common 3D transformations show that TPC significantly outperforms the state of the art. For example, our framework boosts the certified accuracy against twisting transformation along z-axis (within $\pm$20{\textdegree}) from 20.3% to 83.8%. Codes and models are available at https://github.com/Qianhewu/Point-Cloud-Smoothing.

Unified Scaling Laws for Routed Language Models

- Aidan Clark, Diego De Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George Bm Van Den Driessche, Eliza Rutherford, Tom Hennigan, Matthew J Johnson, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Marc’Aurelio Ranzato, Jack Rae, Erich Elsen, Koray Kavukcuoglu, Karen Simonyan
- abstract: The performance of a language model has been shown to be effectively modeled as a power-law in its parameter count. Here we study the scaling behaviors of Routing Networks: architectures that conditionally use only a subset of their parameters while processing an input. For these models, parameter count and computational requirement form two independent axes along which an increase leads to better performance. In this work we derive and justify scaling laws defined on these two variables which generalize those known for standard language models and describe the performance of a wide range of routing architectures trained via three different techniques. Afterwards we provide two applications of these laws: first deriving an Effective Parameter Count along which all models scale at the same rate, and then using the scaling coefficients to give a quantitative comparison of the three routing techniques considered. Our analysis derives from an extensive evaluation of Routing Networks across five orders of magnitude of size, including models with hundreds of experts and hundreds of billions of parameters.

Context-Aware Drift Detection

- Oliver Cobb, Arnaud Van Looveren

- abstract: When monitoring machine learning systems, two-sample tests of homogeneity form the foundation upon which existing approaches to drift detection build. They are used to test for evidence that the distribution underlying recent deployment data differs from that underlying the historical reference data. Often, however, various factors such as time-induced correlation mean that batches of recent deployment data are not expected to form an i.i.d. sample from the historical data distribution. Instead we may wish to test for differences in the distributions conditional on context that is permitted to change. To facilitate this we borrow machinery from the causal inference domain to develop a more general drift detection framework built upon a foundation of two-sample tests for conditional distributional treatment effects. We recommend a particular instantiation of the framework based on maximum conditional mean discrepancies. We then provide an empirical study demonstrating its effectiveness for various drift detection problems of practical interest, such as detecting drift in the distributions underlying subpopulations of data in a manner that is insensitive to their respective prevalences. The study additionally demonstrates applicability to ImageNet-scale vision problems.

On the Robustness of CountSketch to Adaptive Inputs

- Edith Cohen, Xin Lyu, Jelani Nelson, Tamas Sarlos, Moshe Shechner, Uri Stemmer
- abstract: The last decade saw impressive progress towards understanding the performance of algorithms in adaptive settings, where subsequent inputs may depend on the output from prior inputs. Adaptive settings arise in processes with feedback or with adversarial attacks. Existing designs of robust algorithms are generic wrappers of non-robust counterparts and leave open the possibility of better tailored designs. The lowers bounds (attacks) are similarly worst-case and their significance to practical setting is unclear. Aiming to understand these questions, we study the robustness of \texttt{CountSketch}, a popular dimensionality reduction technique that maps vectors to a lower dimension using randomized linear measurements. The sketch supports recovering ℓ_2-heavy hitters of a vector (entries with $v[i]^2 \geq \frac{1}{k}\\|\boldsymbol{v}\|^2_2$). We show that the classic estimator is not robust, and can be attacked with a number of queries of the order of the sketch size. We propose a robust estimator (for a slightly modified sketch) that allows for quadratic number of queries in the sketch size, which is an improvement factor of $\sqrt{k}$ (for k heavy hitters) over prior "blackbox" approaches.

Diffusion bridges vector quantized variational autoencoders

- Max Cohen, Guillaume Quispe, Sylvain Le Corff, Charles Ollion, Eric Moulines
- abstract: Vector Quantized-Variational AutoEncoders (VQ-VAE) are generative models based on discrete latent representations of the data, where inputs are mapped to a finite set of learned embeddings. To generate new samples, an autoregressive prior distribution over the discrete states must be trained separately. This prior is generally very complex and leads to slow generation. In this work, we propose a new model to train the prior and the encoder/decoder networks simultaneously. We build a diffusion bridge between a continuous coded vector and a non-informative prior distribution. The latent discrete states are then given as random functions of these continuous vectors. We show that our model is competitive with the autoregressive prior on the mini-Imagenet and CIFAR dataset and is efficient in both optimization and sampling. Our framework also extends the standard VQ-VAE and enables end-to-end training.

Online and Consistent Correlation Clustering

- Vincent Cohen-Addad, Silvio Lattanzi, Andreas Maggiori, Nikos Parotsidis
- abstract: In the correlation clustering problem the input is a signed graph where the sign indicates whether each pair of points should be placed in the same cluster or not. The goal of the problem is to compute a clustering which minimizes the number of disagreements with such recommendation. Thanks to its many practical applications, correlation clustering is a fundamental unsupervised learning problem and has been extensively studied in many different settings. In this paper we study the problem in the classic online setting with recourse; The vertices of the graphs arrive in an online manner and the goal is to maintain an approximate clustering while minimizing the number of times each vertex changes cluster. Our main contribution is an algorithm that achieves logarithmic recourse per vertex in the worst case. We also complement this result with a tight lower bound. Finally we show experimentally that our algorithm achieves better performances than state-of-the-art algorithms on real world data.

Massively Parallel k-Means Clustering for Perturbation Resilient Instances

- Vincent Cohen-Addad, Vahab Mirrokni, Peilin Zhong
- abstract: We consider k-means clustering of n data points in Euclidean space in the Massively Parallel Computation (MPC) model, a computational model which is an abstraction of modern massively parallel computing system such as MapReduce. Recent work provides evidence that getting $O(1)$-approximate k-means solution for general input points using $o(\log n)$ rounds in the MPC model may be impossible under certain conditions [Ghaffari, Kuhn \& Uitto'2019]. However, the real-world data points usually have better structures. One instance of interest is the set of data points which is perturbation resilient [Bilu \& Linial'2010]. In particular, a point set is α-perturbation resilient for k-means if perturbing pairwise distances by multiplicative factors in the range $[1,\alpha]$ does not change the optimum k-means clusters. We bypass the worst case lower bound by considering the perturbation resilient input points and showing $o(\log n)$ rounds k-means clustering algorithms for these instances in the MPC model. Specifically, we show a fully scalable $(1+\varepsilon)$-approximate k-means clustering algorithm for $O(\alpha)$-perturbation resilient instance in the MPC model using $O(1)$ rounds and ${O}_{\varepsilon,d}(n^{1+1/\alpha^2+o(1)})$ total space. If the space per machine is sufficiently larger than k, i.e., at least $k\cdot n^{\Omega(1)}$, we also develop an optimal k-means clustering algorithm for $O(\alpha)$-perturbation resilient instance in MPC using $O(1)$ rounds and ${O}_d(n^{1+o(1)}\cdot(n^{1/\alpha^2}+k))$ total space.

One-Pass Diversified Sampling with Application to Terabyte-Scale Genomic Sequence Streams

- Benjamin Coleman, Benito Geordie, Li Chou, R. A. Leo Elworth, Todd Treangen, Anshumali Shrivastava
- abstract: A popular approach to reduce the size of a massive dataset is to apply efficient online sampling to the stream of data as it is read or generated. Online sampling routines are currently restricted to variations of reservoir sampling, where each sample is selected uniformly and independently of other samples. This renders them unsuitable for large-scale applications in computational biology, such as metagenomic community profiling and protein function annotation, which suffer from severe class imbalance. To maintain a representative and diverse sample, we must identify and preferentially select data that are likely to belong to rare classes. We argue that existing schemes for diversity sampling have prohibitive overhead for large-scale problems and high-throughput streams. We propose an efficient sampling routine that uses an online representation of the data distribution as a prefilter to retain elements from rare groups. We apply this method to several genomic data analysis tasks and demonstrate significant speedup in downstream analysis without sacrificing the quality of the results. Because our algorithm is 2x faster and uses 1000x less memory than coreset, reservoir and sketch-based alternatives, we anticipate that it will become a useful preprocessing step for applications with large-scale streaming data.

Transfer and Marginalize: Explaining Away Label Noise with Privileged Information

- Mark Collier, Rodolphe Jenatton, Effrosyni Kokiopoulou, Jesse Berent
- abstract: Supervised learning datasets often have privileged information, in the form of features which are available at training time but are not available at test time e.g. the ID of the annotator that provided the label. We argue that privileged information is useful for explaining away label noise, thereby reducing the harmful impact of noisy labels. We develop a simple and efficient method for supervised learning with neural networks: it transfers via weight sharing the knowledge learned with privileged information and approximately marginalizes over privileged information at test time. Our method, TRAM (TRansfer and Marginalize), has minimal training time overhead and has the same test-time cost as not using privileged information. TRAM performs strongly on CIFAR-10H, ImageNet and Civil Comments benchmarks.

MAML and ANIL Provably Learn Representations

- Liam Collins, Aryan Mokhtari, Sewoong Oh, Sanjay Shakkottai
- abstract: Recent empirical evidence has driven conventional wisdom to believe that gradient-based meta-learning (GBML) methods perform well at few-shot learning because they learn an expressive data representation that is shared across tasks. However, the mechanics of GBML have remained largely mysterious from a theoretical perspective. In this paper, we prove that two well-known GBML methods, MAML and ANIL, as well as their first-order approximations, are capable of learning common representation among a set of given tasks. Specifically, in the well-known multi-task linear representation learning setting, they are able to recover the ground-truth representation at an exponentially fast rate. Moreover, our analysis illuminates that the driving force causing MAML and ANIL to recover the underlying representation is that they adapt the final layer of their model, which harnesses the underlying task diversity to improve the representation in all directions of interest. To the best of our knowledge, these are the first results to show that MAML and/or ANIL learn expressive representations and to rigorously explain why they do so.

Entropic Causal Inference: Graph Identifiability

- Spencer Compton, Kristjan Greenewald, Dmitriy A Katz, Murat Kocaoglu
- abstract: Entropic causal inference is a recent framework for learning the causal graph between two variables from observational data by finding the information-theoretically simplest structural explanation of the data, i.e., the model with smallest entropy. In our work, we first extend the causal graph identifiability result in the two-variable setting under relaxed assumptions. We then show the first identifiability result using the entropic approach for learning causal graphs with more than two nodes. Our approach utilizes the property that ancestrality between a source node and its descendants can be determined using the bivariate entropic tests. We provide a sound sequential peeling algorithm for general graphs that relies on this property. We also propose a heuristic algorithm for small graphs that shows strong empirical performance. We rigorously evaluate the performance of our algorithms on synthetic data generated from a variety of models, observing improvement over prior work. Finally we test our algorithms on real-world datasets.

Mitigating Gender Bias in Face Recognition using the von Mises-Fisher Mixture Model

- Jean-Rémy Conti, Nathan Noiry, Stephan Clemencon, Vincent Despiegel, Stéphane Gentric
- abstract: In spite of the high performance and reliability of deep learning algorithms in a wide range of everyday applications, many investigations tend to show that a lot of models exhibit biases, discriminating against specific subgroups of the population (e.g. gender, ethnicity). This urges the practitioner to develop fair systems with a uniform/comparable performance across sensitive groups. In this work, we investigate the gender bias of deep Face Recognition networks. In order to measure this bias, we introduce two new metrics, BFAR and BFRR, that better reflect the inherent deployment needs of Face Recognition systems. Motivated by geometric considerations, we mitigate gender bias through a new post-processing methodology which transforms the deep embeddings of a pre-trained model to give more representation power to discriminated subgroups. It consists in training a shallow neural network by minimizing a Fair von Mises-Fisher loss whose hyperparameters account for the intra-class variance of each gender. Interestingly, we empirically observe that these hyperparameters are correlated with our fairness metrics. In fact, extensive numerical experiments on a variety of datasets show that a careful selection significantly reduces gender bias.

Counterfactual Transportability: A Formal Approach

- Juan D Correa, Sanghack Lee, Elias Bareinboim
- abstract: Generalizing causal knowledge across environments is a common challenge shared across many of the data-driven disciplines, including AI and ML. Experiments are usually performed in one environment (e.g., in a lab, on Earth, in a training ground), almost invariably, with the intent of being used elsewhere (e.g., outside the lab, on Mars, in the real world), in an environment that is related but somewhat different than the original one, where certain conditions and mechanisms are likely to change. This generalization task has been studied in the causal inference literature under the rubric of transportability (Pearl and Bareinboim, 2011). While most transportability works focused on generalizing associational and interventional distributions, the generalization of counterfactual distributions has not been formally studied. In this paper, we investigate the transportability of counterfactuals from an arbitrary combination of observational and experimental distributions coming from disparate domains. Specifically, we introduce a sufficient and necessary graphical condition and develop an efficient, sound, and complete algorithm for transporting counterfactual quantities across domains in nonparametric settings. Failure of the algorithm implies the impossibility of generalizing the target counterfactual from the available data without further assumptions.

Label-Free Explainability for Unsupervised Models

- Jonathan Crabbé, Mihaela van der Schaar
- abstract: Unsupervised black-box models are challenging to interpret. Indeed, most existing explainability methods require labels to select which component(s) of the black-box's output to interpret. In the absence of labels, black-box outputs often are representation vectors whose components do not correspond to any meaningful quantity. Hence, choosing which component(s) to interpret in a label-free unsupervised/self-supervised setting is an important, yet unsolved problem. To bridge this gap in the literature, we introduce two crucial extensions of post-hoc explanation techniques: (1) label-free feature importance and (2) label-free example importance that respectively highlight influential features and training examples for a black-box to construct representations at inference time. We demonstrate that our extensions can be successfully implemented as simple wrappers around many existing feature and example importance methods. We illustrate the utility of our label-free explainability paradigm through a qualitative and quantitative comparison of representation spaces learned by various autoencoders trained on distinct unsupervised tasks.

Evaluating the Adversarial Robustness of Adaptive Test-time Defenses

- Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, Taylan Cemgil
- abstract: Adaptive defenses, which optimize at test time, promise to improve adversarial robustness. We categorize such adaptive test-time defenses, explain their potential benefits and drawbacks, and evaluate a representative variety of the latest adaptive defenses for image classification. Unfortunately, none significantly improve upon static defenses when subjected to our careful case study evaluation. Some even weaken the underlying static model while simultaneously increasing inference computation. While these results are disappointing, we still believe that adaptive test-time defenses are a promising avenue of research and, as such, we provide recommendations for their thorough evaluation. We extend the checklist of Carlini et al. (2019) by providing concrete steps specific to adaptive defenses.

Adversarial Robustness against Multiple and Single l_p-Threat Models via Quick Fine-Tuning of Robust Classifiers

- Francesco Croce, Matthias Hein
- abstract: A major drawback of adversarially robust models, in particular for large scale datasets like ImageNet, is the extremely long training time compared to standard models. Moreover, models should be robust not only to one l_p-threat model but ideally to all of them. In this paper we propose Extreme norm Adversarial Training (E-AT) for multiple-norm robustness which is based on geometric properties of l_p-balls. E-AT costs up to three times less than other adversarial training methods for multiple-norm robustness. Using E-AT we show that for ImageNet a single epoch and for CIFAR-10 three epochs are sufficient to turn any l_p-robust model into a multiple-norm robust model. In this way we get the first multiple-norm robust model for ImageNet and boost the state-of-the-art for multiple-norm robustness to more than $51%$ on CIFAR-10. Finally, we study the general transfer via fine-

tuning of adversarial robustness between different individual l_p-threat models and improve the previous SOTA l_1-robustness on both CIFAR-10 and ImageNet. Extensive experiments show that our scheme works across datasets and architectures including vision transformers.

Self-conditioning Pre-Trained Language Models

- Xavier Suau Cuadros, Luca Zappella, Nicholas Apostoloff
- abstract: In this paper we aim to investigate the mechanisms that guide text generation with pre-trained Transformer-based Language Models (TLMs). Grounded on the Product of Experts formulation by Hinton (1999), we describe a generative mechanism that exploits expert units which naturally exist in TLMs. Such units are responsible for detecting concepts in the input and conditioning text generation on such concepts. We describe how to identify expert units and how to activate them during inference in order to induce any desired concept in the generated output. We find that the activation of a surprisingly small amount of units is sufficient to steer text generation (as little as 3 units in a model with 345M parameters). While the objective of this work is to learn more about how TLMs work, we show that our method is effective for conditioning without fine-tuning or using extra parameters, even on fine-grained homograph concepts. Additionally, we show that our method can be used to correct gender bias present in the output of TLMs and achieves gender parity for all evaluated contexts. We compare our method with FUDGE and PPLM-BoW, and show that our approach is able to achieve gender parity at a lower perplexity and better Self-BLEU score. The proposed method is accessible to a wide audience thanks to its simplicity and minimal compute needs. The findings in this paper are a step forward in understanding the generative mechanisms of TLMs.

Only tails matter: Average-Case Universality and Robustness in the Convex Regime

- Leonardo Cunha, Gauthier Gidel, Fabian Pedregosa, Damien Scieur, Courtney Paquette
- abstract: The recently developed average-case analysis of optimization methods allows a more fine-grained and representative convergence analysis than usual worst-case results. In exchange, this analysis requires a more precise hypothesis over the data generating process, namely assuming knowledge of the expected spectral distribution (ESD) of the random matrix associated with the problem. This work shows that the concentration of eigenvalues near the edges of the ESD determines a problem's asymptotic average complexity. This a priori information on this concentration is a more grounded assumption than complete knowledge of the ESD. This approximate concentration is effectively a middle ground between the coarseness of the worst-case scenario convergence and the restrictive previous average-case analysis. We also introduce the Generalized Chebyshev method, asymptotically optimal under a hypothesis on this concentration and globally optimal when the ESD follows a Beta distribution. We compare its performance to classical optimization algorithms, such as gradient descent or Nesterov's scheme, and we show that, in the average-case context, Nesterov's method is universally nearly optimal asymptotically.

Principal Component Flows

- Edmond Cunningham, Adam D Cobb, Susmit Jha
- abstract: Normalizing flows map an independent set of latent variables to their samples using a bijective transformation. Despite the exact correspondence between samples and latent variables, their high level relationship is not well understood. In this paper we characterize the geometric structure of flows using principal manifolds and understand the relationship between latent variables and samples using contours. We introduce a novel class of normalizing flows, called principal component flows (PCF), whose contours are its principal manifolds, and a variant for injective flows (iPCF) that is more efficient to train than regular injective flows. PCFs can be constructed using any flow architecture, are trained with a regularized maximum likelihood objective and can perform density estimation on all of their principal manifolds. In our experiments we show that PCFs and iPCFs are able to learn the principal manifolds over a variety of datasets. Additionally, we show that PCFs can perform density estimation on data that lie on a manifold with variable dimensionality, which is not possible with existing normalizing flows.

Deep symbolic regression for recurrence prediction

- Stéphane D'Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample, Francois Charton
- abstract: Symbolic regression, i.e. predicting a function from the observation of its values, is well-known to be a challenging task. In this paper, we train Transformers to infer the function or recurrence relation underlying sequences of integers or floats, a typical task in human IQ tests which has hardly been tackled in the machine learning literature. We evaluate our integer model on a subset of OEIS sequences, and show that it outperforms built-in Mathematica functions for recurrence prediction. We also demonstrate that our float model is able to yield informative approximations of out-of-vocabulary functions and constants, e.g. $\operatorname{bessel0}(x)\approx \frac{\sin(x)+\cos(x)}{\sqrt{\pi x}}$ and $1.644934\approx \pi^2/6$.

Continuous Control with Action Quantization from Demonstrations

- Robert Dadashi, Léonard Hussenot, Damien Vincent, Sertan Girgin, Anton Raichuk, Matthieu Geist, Olivier Pietquin
- abstract: In this paper, we propose a novel Reinforcement Learning (RL) framework for problems with continuous action spaces: Action Quantization from Demonstrations (AQuaDem). The proposed approach consists in learning a discretization of continuous action spaces from human demonstrations. This discretization returns a set of plausible actions (in light of the demonstrations) for each input state, thus capturing the priors of the demonstrator and their multimodal behavior. By discretizing the action space, any discrete action deep RL technique can be readily applied to the continuous control problem. Experiments show that the proposed approach outperforms state-of-the-art methods such as SAC in the RL setup, and GAIL in the Imitation Learning setup. We provide a website with interactive videos: https://google-research.github.io/aquadem/ and make the code available: https://github.com/google-research/google-research/tree/master/aquadem.

Dialog Inpainting: Turning Documents into Dialogs

- Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, Kelvin Guu
- abstract: Many important questions (e.g. "How to eat healthier?") require conversation to establish context and explore in depth. However, conversational question answering (ConvQA) systems have long been stymied by scarce training data that is expensive to collect. To address this problem, we propose a new technique for synthetically generating diverse and high-quality dialog data: dialog inpainting. Our approach takes the text of any document and transforms it into a two-person dialog between the writer and an imagined reader: we treat sentences from the article as utterances spoken by the writer, and then use a dialog inpainter to predict what the imagined reader asked or said in between each of the writer's utterances. By applying this approach to passages from Wikipedia and the web, we produce WikiDialog and WebDialog, two datasets totalling 19 million diverse information-seeking dialogs – 1,000x larger than the largest existing ConvQA dataset. Furthermore, human raters judge the answer adequacy and conversationality of WikiDialog to be as good or better than existing manually-collected datasets. Remarkably, our approach shows strong zero-shot capability, generating high quality synthetic data without using any in-domain ConvQA data. Using our inpainted data to pre-train ConvQA retrieval systems, we significantly advance state-of-the-art across three benchmarks (QReCC, OR-QuAC, TREC CAsT) yielding up to 40% relative gains on standard evaluation metrics.

DisPFL: Towards Communication-Efficient Personalized Federated Learning via Decentralized Sparse Training

- Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, Dacheng Tao
- abstract: Personalized federated learning is proposed to handle the data heterogeneity problem amongst clients by learning dedicated tailored local models for each user. However, existing works are often built in a centralized way, leading to high communication pressure and high vulnerability when a failure

or an attack on the central server occurs. In this work, we propose a novel personalized federated learning framework in a decentralized (peer-to-peer) communication protocol named DisPFL, which employs personalized sparse masks to customize sparse local models on the edge. To further save the communication and computation cost, we propose a decentralized sparse training technique, which means that each local model in DisPFL only maintains a fixed number of active parameters throughout the whole local training and peer-to-peer communication process. Comprehensive experiments demonstrate that DisPFL significantly saves the communication bottleneck for the busiest node among all clients and, at the same time, achieves higher model accuracy with less computation cost and communication rounds. Furthermore, we demonstrate that our method can easily adapt to heterogeneous local clients with varying computation complexities and achieves better personalized performances.

Marginal Distribution Adaptation for Discrete Sets via Module-Oriented Divergence Minimization

- Hanjun Dai, Mengjiao Yang, Yuan Xue, Dale Schuurmans, Bo Dai
- abstract: Distributions over discrete sets capture the essential statistics including the high-order correlation among elements. Such information provides powerful insight for decision making across various application domains, e.g., product assortment based on product distribution in shopping carts. While deep generative models trained on pre-collected data can capture existing distributions, such pre-trained models are usually not capable of aligning with a target domain in the presence of distribution shift due to reasons such as temporal shift or the change in the population mix. We develop a general framework to adapt a generative model subject to a (possibly counterfactual) target data distribution with both sampling and computation efficiency. Concretely, instead of re-training a full model from scratch, we reuse the learned modules to preserve the correlations between set elements, while only adjusting corresponding components to align with target marginal constraints. We instantiate the approach for three commonly used forms of discrete set distribution—latent variable, autoregressive, and energy based models—and provide efficient solutions for marginal-constrained optimization in either primal or dual forms. Experiments on both synthetic and real-world e-commerce and EHR datasets show that the proposed framework is able to practically align a generative model to match marginal constraints under distribution shift.

Balancing Sample Efficiency and Suboptimality in Inverse Reinforcement Learning

- Angelo Damiani, Giorgio Manganini, Alberto Maria Metelli, Marcello Restelli
- abstract: We propose a novel formulation for the Inverse Reinforcement Learning (IRL) problem, which jointly accounts for the compatibility with the expert behavior of the identified reward and its effectiveness for the subsequent forward learning phase. Albeit quite natural, especially when the final goal is apprenticeship learning (learning policies from an expert), this aspect has been completely overlooked by IRL approaches so far. We propose a new model-free IRL method that is remarkably able to autonomously find a trade-off between the error induced on the learned policy when potentially choosing a sub-optimal reward, and the estimation error caused by using finite samples in the forward learning phase, which can be controlled by explicitly optimizing also the discount factor of the related learning problem. The approach is based on a min-max formulation for the robust selection of the reward parameters and the discount factor so that the distance between the expert's policy and the learned policy is minimized in the successive forward learning task when a finite and possibly small number of samples is available. Differently from the majority of other IRL techniques, our approach does not involve any planning or forward Reinforcement Learning problems to be solved. After presenting the formulation, we provide a numerical scheme for the optimization, and we show its effectiveness on an illustrative numerical case.

Understanding Robust Generalization in Learning Regular Languages

- Soham Dan, Osbert Bastani, Dan Roth
- abstract: A key feature of human intelligence is the ability to generalize beyond the training distribution, for instance, parsing longer sentences than seen in the past. Currently, deep neural networks struggle to generalize robustly to such shifts in the data distribution. We study robust generalization in the context of using recurrent neural networks (RNNs) to learn regular languages. We hypothesize that standard end-to-end modeling strategies cannot generalize well to systematic distribution shifts and propose a compositional strategy to address this. We compare an end-to-end strategy that maps strings to labels with a compositional strategy that predicts the structure of the deterministic finite state automaton (DFA) that accepts the regular language. We theoretically prove that the compositional strategy generalizes significantly better than the end-to-end strategy. In our experiments, we implement the compositional strategy via an auxiliary task where the goal is to predict the intermediate states visited by the DFA when parsing a string. Our empirical results support our hypothesis, showing that auxiliary tasks can enable robust generalization. Interestingly, the end-to-end RNN generalizes significantly better than the theoretical lower bound, suggesting that it is able to achieve atleast some degree of robust generalization.

Unsupervised Image Representation Learning with Deep Latent Particles

- Tal Daniel, Aviv Tamar
- abstract: We propose a new representation of visual data that disentangles object position from appearance. Our method, termed Deep Latent Particles (DLP), decomposes the visual input into low-dimensional latent "particles", where each particle is described by its spatial location and features of its surrounding region. To drive learning of such representations, we follow a VAE-based based approach and introduce a prior for particle positions based on a spatial-Softmax architecture, and a modification of the evidence lower bound loss inspired by the Chamfer distance between particles. We demonstrate that our DLP representations are useful for downstream tasks such as unsupervised keypoint (KP) detection, image manipulation, and video prediction for scenes composed of multiple dynamic objects. In addition, we show that our probabilistic interpretation of the problem naturally provides uncertainty estimates for particle locations, which can be used for model selection, among other tasks.

Guarantees for Epsilon-Greedy Reinforcement Learning with Function Approximation

- Chris Dann, Yishay Mansour, Mehryar Mohri, Ayush Sekhari, Karthik Sridharan
- abstract: Myopic exploration policies such as epsilon-greedy, softmax, or Gaussian noise fail to explore efficiently in some reinforcement learning tasks and yet, they perform well in many others. In fact, in practice, they are often selected as the top choices, due to their simplicity. But, for what tasks do such policies succeed? Can we give theoretical guarantees for their favorable performance? These crucial questions have been scarcely investigated, despite the prominent practical importance of these policies. This paper presents a theoretical analysis of such policies and provides the first regret and sample-complexity bounds for reinforcement learning with myopic exploration. Our results apply to value-function-based algorithms in episodic MDPs with bounded Bellman Eluder dimension. We propose a new complexity measure called myopic exploration gap, denoted by alpha, that captures a structural property of the MDP, the exploration policy and the given value function class. We show that the sample-complexity of myopic exploration scales quadratically with the inverse of this quantity, 1 / alpha^2. We further demonstrate through concrete examples that myopic exploration gap is indeed favorable in several tasks where myopic exploration succeeds, due to the corresponding dynamics and reward structure.

Monarch: Expressive Structured Matrices for Efficient and Accurate Training

- Tri Dao, Beidi Chen, Nimit S Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, Christopher Re
- abstract: Large neural networks excel in many domains, but they are expensive to train and fine-tune. A popular approach to reduce their compute or memory requirements is to replace dense weight matrices with structured ones (e.g., sparse, low-rank, Fourier transform). These methods have not seen widespread adoption (1) in end-to-end training due to unfavorable efficiency–quality tradeoffs, and (2) in dense-to-sparse fine-tuning due to lack of tractable algorithms to approximate a given dense weight matrix. To address these issues, we propose a class of matrices (Monarch) that is hardware-efficient (they are parameterized as products of two block-diagonal matrices for better hardware utilization) and expressive (they can represent many commonly used transforms). Surprisingly, the problem of approximating a dense weight matrix with a Monarch matrix, though nonconvex, has an

analytical optimal solution. These properties of Monarch matrices unlock new ways to train and fine-tune sparse and dense models. We empirically validate that Monarch can achieve favorable accuracy-efficiency tradeoffs in several end-to-end sparse training applications: speeding up ViT and GPT-2 training on ImageNet classification and Wikitext-103 language modeling by 2x with comparable model quality, and reducing the error on PDE solving and MRI reconstruction tasks by 40%. In sparse-to-dense training, with a simple technique called "reverse sparsification," Monarch matrices serve as a useful intermediate representation to speed up GPT-2 pretraining on OpenWebText by 2x without quality drop. The same technique brings 23% faster BERT pretraining than even the very optimized implementation from Nvidia that set the MLPerf 1.1 record. In dense-to-sparse fine-tuning, as a proof-of-concept, our Monarch approximation algorithm speeds up BERT fine-tuning on GLUE by 1.7x with comparable accuracy.

Score-Guided Intermediate Level Optimization: Fast Langevin Mixing for Inverse Problems

- Giannis Daras, Yuval Dagan, Alex Dimakis, Constantinos Daskalakis
- abstract: We prove fast mixing and characterize the stationary distribution of the Langevin Algorithm for inverting random weighted DNN generators. This result extends the work of Hand and Voroninski from efficient inversion to efficient posterior sampling. In practice, to allow for increased expressivity, we propose to do posterior sampling in the latent space of a pre-trained generative model. To achieve that, we train a score-based model in the latent space of a StyleGAN-2 and we use it to solve inverse problems. Our framework, Score-Guided Intermediate Layer Optimization (SGILO), extends prior work by replacing the sparsity regularization with a generative prior in the intermediate layer. Experimentally, we obtain significant improvements over the previous state-of-the-art, especially in the low measurement regime.

Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing

- Mohammad Zalbagi Darestani, Jiayu Liu, Reinhard Heckel
- abstract: Deep learning based image reconstruction methods outperform traditional methods. However, neural networks suffer from a performance drop when applied to images from a different distribution than the training images. For example, a model trained for reconstructing knees in accelerated magnetic resonance imaging (MRI) does not reconstruct brains well, even though the same network trained on brains reconstructs brains perfectly well. Thus there is a distribution shift performance gap for a given neural network, defined as the difference in performance when training on a distribution P and training on another distribution Q, and evaluating both models on Q. In this work, we propose a domain adaptation method for deep learning based compressive sensing that relies on self-supervision during training paired with test-time training at inference. We show that for four natural distribution shifts, this method essentially closes the distribution shift performance gap for state-of-the-art architectures for accelerated MRI.

Knowledge Base Question Answering by Case-based Reasoning over Subgraphs

- Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, Andrew Mccallum
- abstract: Question answering (QA) over knowledge bases (KBs) is challenging because of the diverse, essentially unbounded, types of reasoning patterns needed. However, we hypothesize in a large KB, reasoning patterns required to answer a query type reoccur for various entities in their respective subgraph neighborhoods. Leveraging this structural similarity between local neighborhoods of different subgraphs, we introduce a semiparametric model (CBR-SUBG) with (i) a nonparametric component that for each query, dynamically retrieves other similar k-nearest neighbor (KNN) training queries along with query-specific subgraphs and (ii) a parametric component that is trained to identify the (latent) reasoning patterns from the subgraphs of KNN queries and then apply them to the subgraph of the target query. We also propose an adaptive subgraph collection strategy to select a query-specific compact subgraph, allowing us to scale to full Freebase KB containing billions of facts. We show that CBR-SUBG can answer queries requiring subgraph reasoning patterns and performs competitively with the best models on several KBQA benchmarks. Our subgraph collection strategy also produces more compact subgraphs (e.g. 55% reduction in size for WebQSP while increasing answer recall by 4.85%)\footnote{Code, model, and subgraphs are available at \url{https://github.com/rajarshd/CBR-SUBG}}.

Framework for Evaluating Faithfulness of Local Explanations

- Sanjoy Dasgupta, Nave Frost, Michal Moshkovitz
- abstract: We study the faithfulness of an explanation system to the underlying prediction model. We show that this can be captured by two properties, consistency and sufficiency, and introduce quantitative measures of the extent to which these hold. Interestingly, these measures depend on the test-time data distribution. For a variety of existing explanation systems, such as anchors, we analytically study these quantities. We also provide estimators and sample complexity bounds for empirically determining the faithfulness of black-box explanation systems. Finally, we experimentally validate the new properties and estimators.

Distinguishing rule and exemplar-based generalization in learning systems

- Ishita Dasgupta, Erin Grant, Tom Griffiths
- abstract: Machine learning systems often do not share the same inductive biases as humans and, as a result, extrapolate or generalize in ways that are inconsistent with our expectations. The trade-off between exemplar- and rule-based generalization has been studied extensively in cognitive psychology; in this work, we present a protocol inspired by these experimental approaches to probe the inductive biases that control this trade-off in category-learning systems such as artificial neural networks. We isolate two such inductive biases: feature-level bias (differences in which features are more readily learned) and exemplar-vs-rule bias (differences in how these learned features are used for generalization of category labels). We find that standard neural network models are feature-biased and have a propensity towards exemplar-based extrapolation; we discuss the implications of these findings for machine-learning research on data augmentation, fairness, and systematic generalization.

Robust Multi-Objective Bayesian Optimization Under Input Noise

- Samuel Daulton, Sait Cakmak, Maximilian Balandat, Michael A. Osborne, Enlu Zhou, Eytan Bakshy
- abstract: Bayesian optimization (BO) is a sample-efficient approach for tuning design parameters to optimize expensive-to-evaluate, black-box performance metrics. In many manufacturing processes, the design parameters are subject to random input noise, resulting in a product that is often less performant than expected. Although BO methods have been proposed for optimizing a single objective under input noise, no existing method addresses the practical scenario where there are multiple objectives that are sensitive to input perturbations. In this work, we propose the first multi-objective BO method that is robust to input noise. We formalize our goal as optimizing the multivariate value-at-risk (MVaR), a risk measure of the uncertain objectives. Since directly optimizing MVaR is computationally infeasible in many settings, we propose a scalable, theoretically-grounded approach for optimizing MVaR using random scalarizations. Empirically, we find that our approach significantly outperforms alternative methods and efficiently identifies optimal robust designs that will satisfy specifications across multiple metrics with high probability.

Attentional Meta-learners for Few-shot Polythetic Classification

- Ben J Day, Ramon Viñas Torné, Nikola Simidjievski, Pietro Lió
- abstract: Polythetic classifications, based on shared patterns of features that need neither be universal nor constant among members of a class, are common in the natural world and greatly outnumber monothetic classifications over a set of features. We show that threshold meta-learners, such as Prototypical Networks, require an embedding dimension that is exponential in the number of task-relevant features to emulate these functions. In contrast, attentional

classifiers, such as Matching Networks, are polythetic by default and able to solve these problems with a linear embedding dimension. However, we find that in the presence of task-irrelevant features, inherent to meta-learning problems, attentional models are susceptible to misclassification. To address this challenge, we propose a self-attention feature-selection mechanism that adaptively dilutes non-discriminative features. We demonstrate the effectiveness of our approach in meta-learning Boolean functions, and synthetic and real-world few-shot learning tasks.

Adversarial Vulnerability of Randomized Ensembles

- Hassan Dbouk, Naresh Shanbhag
- abstract: Despite the tremendous success of deep neural networks across various tasks, their vulnerability to imperceptible adversarial perturbations has hindered their deployment in the real world. Recently, works on randomized ensembles have empirically demonstrated significant improvements in adversarial robustness over standard adversarially trained (AT) models with minimal computational overhead, making them a promising solution for safety-critical resource-constrained applications. However, this impressive performance raises the question: Are these robustness gains provided by randomized ensembles real? In this work we address this question both theoretically and empirically. We first establish theoretically that commonly employed robustness evaluation methods such as adaptive PGD provide a false sense of security in this setting. Subsequently, we propose a theoretically-sound and efficient adversarial attack algorithm (ARC) capable of compromising random ensembles even in cases where adaptive PGD fails to do so. We conduct comprehensive experiments across a variety of network architectures, training schemes, datasets, and norms to support our claims, and empirically establish that randomized ensembles are in fact more vulnerable to ℓ_p-bounded adversarial perturbations than even standard AT models. Our code can be found at https://github.com/hsndbk4/ARC.

Born-Infeld (BI) for AI: Energy-Conserving Descent (ECD) for Optimization

- Giuseppe Bruno De Luca, Eva Silverstein
- abstract: We introduce a novel framework for optimization based on energy-conserving Hamiltonian dynamics in a strongly mixing (chaotic) regime and establish its key properties analytically and numerically. The prototype is a discretization of Born-Infeld dynamics, with a squared relativistic speed limit depending on the objective function. This class of frictionless, energy-conserving optimizers proceeds unobstructed until slowing naturally near the minimal loss, which dominates the phase space volume of the system. Building from studies of chaotic systems such as dynamical billiards, we formulate a specific algorithm with good performance on machine learning and PDE-solving tasks, including generalization. It cannot stop at a high local minimum, an advantage in non-convex loss functions, and proceeds faster than GD+momentum in shallow valleys.

Error-driven Input Modulation: Solving the Credit Assignment Problem without a Backward Pass

- Giorgia Dellaferrera, Gabriel Kreiman
- abstract: Supervised learning in artificial neural networks typically relies on backpropagation, where the weights are updated based on the error-function gradients and sequentially propagated from the output layer to the input layer. Although this approach has proven effective in a wide domain of applications, it lacks biological plausibility in many regards, including the weight symmetry problem, the dependence of learning on non-local signals, the freezing of neural activity during error propagation, and the update locking problem. Alternative training schemes have been introduced, including sign symmetry, feedback alignment, and direct feedback alignment, but they invariably rely on a backward pass that hinders the possibility of solving all the issues simultaneously. Here, we propose to replace the backward pass with a second forward pass in which the input signal is modulated based on the error of the network. We show that this novel learning rule comprehensively addresses all the above-mentioned issues and can be applied to both fully connected and convolutional models. We test this learning rule on MNIST, CIFAR-10, and CIFAR-100. These results help incorporate biological principles into machine learning.

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

- Fei Deng, Ingook Jang, Sungjin Ahn
- abstract: Reconstruction-based Model-Based Reinforcement Learning (MBRL) agents, such as Dreamer, often fail to discard task-irrelevant visual distractions that are prevalent in natural scenes. In this paper, we propose a reconstruction-free MBRL agent, called DreamerPro, that can enhance robustness to distractions. Motivated by the recent success of prototypical representations, a non-contrastive self-supervised learning approach in computer vision, DreamerPro combines Dreamer with prototypes. In order for the prototypes to benefit temporal dynamics learning in MBRL, we propose to additionally learn the prototypes from the recurrent states of the world model, thereby distilling temporal structures from past observations and actions into the prototypes. Experiments on the DeepMind Control suite show that DreamerPro achieves better overall performance than state-of-the-art contrastive MBRL agents when there are complex background distractions, and maintains similar performance as Dreamer in standard tasks where contrastive MBRL agents can perform much worse.

NeuralEF: Deconstructing Kernels by Deep Neural Networks

- Zhijie Deng, Jiaxin Shi, Jun Zhu
- abstract: Learning the principal eigenfunctions of an integral operator defined by a kernel and a data distribution is at the core of many machine learning problems. Traditional nonparametric solutions based on the Nystrom formula suffer from scalability issues. Recent work has resorted to a parametric approach, i.e., training neural networks to approximate the eigenfunctions. However, the existing method relies on an expensive orthogonalization step and is difficult to implement. We show that these problems can be fixed by using a new series of objective functions that generalizes the EigenGame to function space. We test our method on a variety of supervised and unsupervised learning problems and show it provides accurate approximations to the eigenfunctions of polynomial, radial basis, neural network Gaussian process, and neural tangent kernels. Finally, we demonstrate our method can scale up linearised Laplace approximation of deep neural networks to modern image classification datasets through approximating the Gauss-Newton matrix. Code is available at https://github.com/thudzj/neuraleigenfunction.

Deep Causal Metric Learning

- Xiang Deng, Zhongfei Zhang
- abstract: Deep metric learning aims to learn distance metrics that measure similarities and dissimilarities between samples. The existing approaches typically focus on designing different hard sample mining or distance margin strategies and then minimize a pair/triplet-based or proxy-based loss over the training data. However, this can lead the model to recklessly learn all the correlated distances found in training data including the spurious distance (e.g., background differences) that is not the distance of interest and can harm the generalization of the learned metric. To address this issue, we study metric learning from a causality perspective and accordingly propose deep causal metric learning (DCML) that pursues the true causality of the distance between samples. DCML is achieved through explicitly learning environment-invariant attention and task-invariant embedding based on causal inference. Extensive experiments on several benchmark datasets demonstrate the superiority of DCML over the existing methods.

On the Convergence of Inexact Predictor-Corrector Methods for Linear Programming

- Gregory Dexter, Agniva Chowdhury, Haim Avron, Petros Drineas

- abstract: Interior point methods (IPMs) are a common approach for solving linear programs (LPs) with strong theoretical guarantees and solid empirical performance. The time complexity of these methods is dominated by the cost of solving a linear system of equations at each iteration. In common applications of linear programming, particularly in machine learning and scientific computing, the size of this linear system can become prohibitively large, requiring the use of iterative solvers, which provide an approximate solution to the linear system. However, approximately solving the linear system at each iteration of an IPM invalidates the theoretical guarantees of common IPM analyses. To remedy this, we theoretically and empirically analyze (slightly modified) predictor-corrector IPMs when using approximate linear solvers: our approach guarantees that, when certain conditions are satisfied, the number of IPM iterations does not increase and that the final solution remains feasible. We also provide practical instantiations of approximate linear solvers that satisfy these conditions for special classes of constraint matrices using randomized linear algebra.

Analysis of Stochastic Processes through Replay Buffers

- Shirli Di-Castro, Shie Mannor, Dotan Di Castro
- abstract: Replay buffers are a key component in many reinforcement learning schemes. Yet, their theoretical properties are not fully understood. In this paper we analyze a system where a stochastic process X is pushed into a replay buffer and then randomly sampled to generate a stochastic process Y from the replay buffer. We provide an analysis of the properties of the sampled process such as stationarity, Markovity and autocorrelation in terms of the properties of the original process. Our theoretical analysis sheds light on why replay buffer may be a good de-correlator. Our analysis provides theoretical tools for proving the convergence of replay buffer based algorithms which are prevalent in reinforcement learning schemes.

Streaming Algorithms for High-Dimensional Robust Statistics

- Ilias Diakonikolas, Daniel M. Kane, Ankit Pensia, Thanasis Pittas
- abstract: We study high-dimensional robust statistics tasks in the streaming model. A recent line of work obtained computationally efficient algorithms for a range of high-dimensional robust statistics tasks. Unfortunately, all previous algorithms require storing the entire dataset, incurring memory at least quadratic in the dimension. In this work, we develop the first efficient streaming algorithms for high-dimensional robust statistics with near-optimal memory requirements (up to logarithmic factors). Our main result is for the task of high-dimensional robust mean estimation in (a strengthening of) Huber's contamination model. We give an efficient single-pass streaming algorithm for this task with near-optimal error guarantees and space complexity nearly-linear in the dimension. As a corollary, we obtain streaming algorithms with near-optimal space complexity for several more complex tasks, including robust covariance estimation, robust regression, and more generally robust stochastic optimization.

Learning General Halfspaces with Adversarial Label Noise via Online Gradient Descent

- Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, Nikos Zarifis
- abstract: We study the problem of learning general {—} i.e., not necessarily homogeneous {—} halfspaces with adversarial label noise under the Gaussian distribution. Prior work has provided a sophisticated polynomial-time algorithm for this problem. In this work, we show that the problem can be solved directly via online gradient descent applied to a sequence of natural non-convex surrogates. This approach yields a simple iterative learning algorithm for general halfspaces with near-optimal sample complexity, runtime, and error guarantee. At the conceptual level, our work establishes an intriguing connection between learning halfspaces with adversarial noise and online optimization that may find other applications.

Variational Feature Pyramid Networks

- Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou
- abstract: Recent architectures for object detection adopt a Feature Pyramid Network as a backbone for deep feature extraction. Many works focus on the design of pyramid networks which produce richer feature representations. In this work, we opt to learn a dataset-specific architecture for Feature Pyramid Networks. With the proposed method, the network fuses features at multiple scales, it is efficient in terms of parameters and operations, and yields better results across a variety of tasks and datasets. Starting by a complex network, we adopt Variational Inference to prune redundant connections. Our model, integrated with standard detectors, outperforms the state-of-the-art feature fusion networks.

Understanding Doubly Stochastic Clustering

- Tianjiao Ding, Derek Lim, Rene Vidal, Benjamin D Haeffele
- abstract: The problem of projecting a matrix onto the space of doubly stochastic matrices finds several applications in machine learning. For example, in spectral clustering, it has been shown that forming the normalized Laplacian matrix from a data affinity matrix has close connections to projecting it onto the set of doubly stochastic matrices. However, the analysis of why this projection improves clustering has been limited. In this paper we present theoretical conditions on the given affinity matrix under which its doubly stochastic projection is an ideal affinity matrix (i.e., it has no false connections between clusters, and is well-connected within each cluster). In particular, we show that a necessary and sufficient condition for a projected affinity matrix to be ideal reduces to a set of conditions on the input affinity that decompose along each cluster. Further, in the subspace clustering problem, where each cluster is defined by a linear subspace, we provide geometric conditions on the underlying subspaces which guarantee correct clustering via a continuous version of the problem. This allows us to explain theoretically the remarkable performance of a recently proposed doubly stochastic subspace clustering method.

Independent Policy Gradient for Large-Scale Markov Potential Games: Sharper Rates, Function Approximation, and Game-Agnostic Convergence

- Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, Mihailo Jovanovic
- abstract: We examine global non-asymptotic convergence properties of policy gradient methods for multi-agent reinforcement learning (RL) problems in Markov potential games (MPGs). To learn a Nash equilibrium of an MPG in which the size of state space and/or the number of players can be very large, we propose new independent policy gradient algorithms that are run by all players in tandem. When there is no uncertainty in the gradient evaluation, we show that our algorithm finds an ϵ-Nash equilibrium with $O(1/\epsilon^2)$ iteration complexity which does not explicitly depend on the state space size. When the exact gradient is not available, we establish $O(1/\epsilon^5)$ sample complexity bound in a potentially infinitely large state space for a sample-based algorithm that utilizes function approximation. Moreover, we identify a class of independent policy gradient algorithms that enjoy convergence for both zero-sum Markov games and Markov cooperative games with the players that are oblivious to the types of games being played. Finally, we provide computational experiments to corroborate the merits and the effectiveness of our theoretical developments.

Generalization and Robustness Implications in Object-Centric Learning

- Andrea Dittadi, Samuele S Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, Francesco Locatello
- abstract: The idea behind object-centric representation learning is that natural scenes can better be modeled as compositions of objects and their relations as opposed to distributed representations. This inductive bias can be injected into neural networks to potentially improve systematic generalization and performance of downstream tasks in scenes with multiple objects. In this paper, we train state-of-the-art unsupervised models on five common multi-object datasets and evaluate segmentation metrics and downstream object property prediction. In addition, we study generalization and robustness by investigating the settings where either a single object is out of distribution – e.g., having an unseen color, texture, or shape – or global properties of the

scene are altered – e.g., by occlusions, cropping, or increasing the number of objects. From our experimental study, we find object-centric representations to be useful for downstream tasks and generally robust to most distribution shifts affecting objects. However, when the distribution shift affects the input in a less structured manner, robustness in terms of segmentation and downstream task performance may vary significantly across models and distribution shifts.

Fair Generalized Linear Models with a Convex Penalty

- Hyungrok Do, Preston Putzel, Axel S Martin, Padhraic Smyth, Judy Zhong
- abstract: Despite recent advances in algorithmic fairness, methodologies for achieving fairness with generalized linear models (GLMs) have yet to be explored in general, despite GLMs being widely used in practice. In this paper we introduce two fairness criteria for GLMs based on equalizing expected outcomes or log-likelihoods. We prove that for GLMs both criteria can be achieved via a convex penalty term based solely on the linear components of the GLM, thus permitting efficient optimization. We also derive theoretical properties for the resulting fair GLM estimator. To empirically demonstrate the efficacy of the proposed fair GLM, we compare it with other well-known fair prediction methods on an extensive set of benchmark datasets for binary classification and regression. In addition, we demonstrate that the fair GLM can generate fair predictions for a range of response variables, other than binary and continuous outcomes.

Bayesian Learning with Information Gain Provably Bounds Risk for a Robust Adversarial Defense

- Bao Gia Doan, Ehsan M Abbasnejad, Javen Qinfeng Shi, Damith Ranashinghe
- abstract: We present a new algorithm to learn a deep neural network model robust against adversarial attacks. Previous algorithms demonstrate an adversarially trained Bayesian Neural Network (BNN) provides improved robustness. We recognize the learning approach for approximating the multi-modal posterior distribution of an adversarially trained Bayesian model can lead to mode collapse; consequently, the model's achievements in robustness and performance are sub-optimal. Instead, we first propose preventing mode collapse to better approximate the multi-modal posterior distribution. Second, based on the intuition that a robust model should ignore perturbations and only consider the informative content of the input, we conceptualize and formulate an information gain objective to measure and force the information learned from both benign and adversarial training instances to be similar. Importantly. we prove and demonstrate that minimizing the information gain objective allows the adversarial risk to approach the conventional empirical risk. We believe our efforts provide a step towards a basis for a principled method of adversarially training BNNs. Our extensive experimental results demonstrate significantly improved robustness up to 20% compared with adversarial training and Adv-BNN under PGD attacks with 0.035 distortion on both CIFAR-10 and STL-10 dataset.

On the Adversarial Robustness of Causal Algorithmic Recourse

- Ricardo Dominguez-Olmedo, Amir H Karimi, Bernhard Schölkopf
- abstract: Algorithmic recourse seeks to provide actionable recommendations for individuals to overcome unfavorable classification outcomes from automated decision-making systems. Recourse recommendations should ideally be robust to reasonably small uncertainty in the features of the individual seeking recourse. In this work, we formulate the adversarially robust recourse problem and show that recourse methods that offer minimally costly recourse fail to be robust. We then present methods for generating adversarially robust recourse for linear and for differentiable classifiers. Finally, we show that regularizing the decision-making classifier to behave locally linearly and to rely more strongly on actionable features facilitates the existence of adversarially robust recourse.

Finding the Task-Optimal Low-Bit Sub-Distribution in Deep Neural Networks

- Runpei Dong, Zhanhong Tan, Mengdi Wu, Linfeng Zhang, Kaisheng Ma
- abstract: Quantized neural networks typically require smaller memory footprints and lower computation complexity, which is crucial for efficient deployment. However, quantization inevitably leads to a distribution divergence from the original network, which generally degrades the performance. To tackle this issue, massive efforts have been made, but most existing approaches lack statistical considerations and depend on several manual configurations. In this paper, we present an adaptive-mapping quantization method to learn an optimal latent sub-distribution that is inherent within models and smoothly approximated with a concrete Gaussian Mixture (GM). In particular, the network weights are projected in compliance with the GM-approximated sub-distribution. This sub-distribution evolves along with the weight update in a co-tuning schema guided by the direct task-objective optimization. Sufficient experiments on image classification and object detection over various modern architectures demonstrate the effectiveness, generalization property, and transferability of the proposed method. Besides, an efficient deployment flow for the mobile CPU is developed, achieving up to 7.46$\times$ inference acceleration on an octa-core ARM CPU. Our codes have been publicly released at https://github.com/RunpeiDong/DGMS.

PACE: A Parallelizable Computation Encoder for Directed Acyclic Graphs

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

Privacy for Free: How does Dataset Condensation Help Privacy?

- Tian Dong, Bo Zhao, Lingjuan Lyu
- abstract: To prevent unintentional data leakage, research community has resorted to data generators that can produce differentially private data for model training. However, for the sake of the data privacy, existing solutions suffer from either expensive training cost or poor generalization performance. Therefore, we raise the question whether training efficiency and privacy can be achieved simultaneously. In this work, we for the first time identify that dataset condensation (DC) which is originally designed for improving training efficiency is also a better solution to replace the traditional data generators for private data generation, thus providing privacy for free. To demonstrate the privacy benefit of DC, we build a connection between DC and differential privacy, and theoretically prove on linear feature extractors (and then extended to non-linear feature extractors) that the existence of one sample has limited impact ($O(m/n)$) on the parameter distribution of networks trained on m samples synthesized from $n (n \gg m)$ raw samples by DC. We also empirically validate the visual privacy and membership privacy of DC-synthesized data by launching both the loss-based and the state-of-the-art likelihood-based membership inference attacks. We envision this work as a milestone for data-efficient and privacy-preserving machine learning.

Fast rates for noisy interpolation require rethinking the effect of inductive bias

- Konstantin Donhauser, Nicolò Ruggeri, Stefan Stojanovic, Fanny Yang
- abstract: Good generalization performance on high-dimensional data crucially hinges on a simple structure of the ground truth and a corresponding strong inductive bias of the estimator. Even though this intuition is valid for regularized models, in this paper we caution against a strong inductive bias for interpolation in the presence of noise: While a stronger inductive bias encourages a simpler structure that is more aligned with the ground truth, it also increases the detrimental effect of noise. Specifically, for both linear regression and classification with a sparse ground truth, we prove that minimum ℓ_p-norm and maximum ℓ_p-margin interpolators achieve fast polynomial rates close to order $1/n$ for $p > 1$ compared to a logarithmic rate for $p = 1$. Finally, we provide preliminary experimental evidence that this trade-off may also play a crucial role in understanding non-linear interpolating models used in practice.

Adapting to Mixing Time in Stochastic Optimization with Markovian Data

- Ron Dorfman, Kfir Yehuda Levy
- abstract: We consider stochastic optimization problems where data is drawn from a Markov chain. Existing methods for this setting crucially rely on knowing the mixing time of the chain, which in real-world applications is usually unknown. We propose the first optimization method that does not require the knowledge of the mixing time, yet obtains the optimal asymptotic convergence rate when applied to convex problems. We further show that our approach can be extended to: (i) finding stationary points in non-convex optimization with Markovian data, and (ii) obtaining better dependence on the mixing time in temporal difference (TD) learning; in both cases, our method is completely oblivious to the mixing time. Our method relies on a novel combination of multi-level Monte Carlo (MLMC) gradient estimation together with an adaptive learning method.

TACTiS: Transformer-Attentional Copulas for Time Series

- Alexandre Drouin, Étienne Marcotte, Nicolas Chapados
- abstract: The estimation of time-varying quantities is a fundamental component of decision making in fields such as healthcare and finance. However, the practical utility of such estimates is limited by how accurately they quantify predictive uncertainty. In this work, we address the problem of estimating the joint predictive distribution of high-dimensional multivariate time series. We propose a versatile method, based on the transformer architecture, that estimates joint distributions using an attention-based decoder that provably learns to mimic the properties of non-parametric copulas. The resulting model has several desirable properties: it can scale to hundreds of time series, supports both forecasting and interpolation, can handle unaligned and non-uniformly sampled data, and can seamlessly adapt to missing data during training. We demonstrate these properties empirically and show that our model produces state-of-the-art predictions on multiple real-world datasets.

Branching Reinforcement Learning

- Yihan Du, Wei Chen
- abstract: In this paper, we propose a novel Branching Reinforcement Learning (Branching RL) model, and investigate both Regret Minimization (RM) and Reward-Free Exploration (RFE) metrics for this model. Unlike standard RL where the trajectory of each episode is a single H-step path, branching RL allows an agent to take multiple base actions in a state such that transitions branch out to multiple successor states correspondingly, and thus it generates a tree-structured trajectory. This model finds important applications in hierarchical recommendation systems and online advertising. For branching RL, we establish new Bellman equations and key lemmas, i.e., branching value difference lemma and branching law of total variance, and also bound the total variance by only $O(H^2)$ under an exponentially-large trajectory. For RM and RFE metrics, we propose computationally efficient algorithms BranchVI and BranchRFE, respectively, and derive nearly matching upper and lower bounds. Our regret and sample complexity results are polynomial in all problem parameters despite exponentially-large trajectories.

Bayesian Imitation Learning for End-to-End Mobile Manipulation

- Yuqing Du, Daniel Ho, Alex Alemi, Eric Jang, Mohi Khansari
- abstract: In this work we investigate and demonstrate benefits of a Bayesian approach to imitation learning from multiple sensor inputs, as applied to the task of opening office doors with a mobile manipulator. Augmenting policies with additional sensor inputs{—}such as RGB + depth cameras{—}is a straightforward approach to improving robot perception capabilities, especially for tasks that may favor different sensors in different situations. As we scale multi-sensor robotic learning to unstructured real-world settings (e.g. offices, homes) and more complex robot behaviors, we also increase reliance on simulators for cost, efficiency, and safety. Consequently, the sim-to-real gap across multiple sensor modalities also increases, making simulated validation more difficult. We show that using the Variational Information Bottleneck (Alemi et al., 2016) to regularize convolutional neural networks improves generalization to heldout domains and reduces the sim-to-real gap in a sensor-agnostic manner. As a side effect, the learned embeddings also provide useful estimates of model uncertainty for each sensor. We demonstrate that our method is able to help close the sim-to-real gap and successfully fuse RGB and depth modalities based on understanding of the situational uncertainty of each sensor. In a real-world office environment, we achieve 96% task success, improving upon the baseline by +16%.

GLaM: Efficient Scaling of Language Models with Mixture-of-Experts

- Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, Claire Cui
- abstract: Scaling language models with more data, compute and parameters has driven significant progress in natural language processing. For example, thanks to scaling, GPT-3 was able to achieve strong results on in-context learning tasks. However, training these large dense models requires significant amounts of computing resources. In this paper, we propose and develop a family of language models named \glam (\textbf{G}eneralist \textbf{La}nguage \textbf{M}odel), which uses a sparsely activated mixture-of-experts architecture to scale the model capacity while also incurring substantially less training cost compared to dense variants. The largest \glam has 1.2 trillion parameters, which is approximately 7x larger than GPT-3. It consumes only 1/3 of the energy used to train GPT-3 and requires half of the computation flops for inference, while still achieving better overall fewshot performance across 29 NLP tasks.

Learning Iterative Reasoning through Energy Minimization

- Yilun Du, Shuang Li, Joshua Tenenbaum, Igor Mordatch
- abstract: Deep learning has excelled on complex pattern recognition tasks such as image classification and object recognition. However, it struggles with tasks requiring nontrivial reasoning, such as algorithmic computation. Humans are able to solve such tasks through iterative reasoning – spending more time to think about harder tasks. Most existing neural networks, however, exhibit a fixed computational budget controlled by the neural network architecture, preventing additional computational processing on harder tasks. In this work, we present a new framework for iterative reasoning with neural networks. We train a neural network to parameterize an energy landscape over all outputs, and implement each step of the iterative reasoning as an energy minimization step to find a minimal energy solution. By formulating reasoning as an energy minimization problem, for harder problems that lead to more complex energy landscapes, we may then adjust our underlying computational budget by running a more complex optimization procedure. We empirically illustrate that our iterative reasoning approach can solve more accurate and generalizable algorithmic reasoning tasks in both graph and continuous domains. Finally, we illustrate that our approach can recursively solve algorithmic problems requiring nested reasoning.

SE(3) Equivariant Graph Neural Networks with Complete Local Frames

- Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, Tie-Yan Liu
- abstract: Group equivariance (e.g. SE(3) equivariance) is a critical physical symmetry in science, from classical and quantum physics to computational biology. It enables robust and accurate prediction under arbitrary reference transformations. In light of this, great efforts have been put on encoding this symmetry into deep neural networks, which has been shown to improve the generalization performance and data efficiency for downstream tasks. Constructing an equivariant neural network generally brings high computational costs to ensure expressiveness. Therefore, how to better trade-off the expressiveness and computational efficiency plays a core role in the design of the equivariant deep learning models. In this paper, we propose a framework to construct SE(3) equivariant graph neural networks that can approximate the geometric quantities efficiently. Inspired by differential geometry and physics, we introduce equivariant local complete frames to graph neural networks, such that tensor information at given orders can be projected onto the frames. The local frame is constructed to form an orthonormal basis that avoids direction degeneration and ensure completeness. Since the frames are built only by cross product operations, our method is computationally efficient. We evaluate our method on two tasks: Newton mechanics modeling and equilibrium molecule conformation generation. Extensive experimental results demonstrate that our model achieves the best or competitive performance in two types of datasets.

A Context-Integrated Transformer-Based Neural Network for Auction Design

- Zhijian Duan, Jingwu Tang, Yutong Yin, Zhe Feng, Xiang Yan, Manzil Zaheer, Xiaotie Deng
- abstract: One of the central problems in auction design is developing an incentive-compatible mechanism that maximizes the auctioneer's expected revenue. While theoretical approaches have encountered bottlenecks in multi-item auctions, recently, there has been much progress on finding the optimal mechanism through deep learning. However, these works either focus on a fixed set of bidders and items, or restrict the auction to be symmetric. In this work, we overcome such limitations by factoring public contextual information of bidders and items into the auction learning framework. We propose $\mathtt{CITransNet}$, a context-integrated transformer-based neural network for optimal auction design, which maintains permutation-equivariance over bids and contexts while being able to find asymmetric solutions. We show by extensive experiments that $\mathtt{CITransNet}$ can recover the known optimal solutions in single-item settings, outperform strong baselines in multi-item auctions, and generalize well to cases other than those in training.

Augment with Care: Contrastive Learning for Combinatorial Problems

- Haonan Duan, Pashootan Vaezipoor, Max B Paulus, Yangjun Ruan, Chris Maddison
- abstract: Supervised learning can improve the design of state-of-the-art solvers for combinatorial problems, but labelling large numbers of combinatorial instances is often impractical due to exponential worst-case complexity. Inspired by the recent success of contrastive pre-training for images, we conduct a scientific study of the effect of augmentation design on contrastive pre-training for the Boolean satisfiability problem. While typical graph contrastive pre-training uses label-agnostic augmentations, our key insight is that many combinatorial problems have well-studied invariances, which allow for the design of label-preserving augmentations. We find that label-preserving augmentations are critical for the success of contrastive pre-training. We show that our representations are able to achieve comparable test accuracy to fully-supervised learning while using only 1% of the labels. We also demonstrate that our representations are more transferable to larger problems from unseen domains. Our code is available at https://github.com/h4duan/contrastive-sat.

Parametric Visual Program Induction with Function Modularization

- Xuguang Duan, Xin Wang, Ziwei Zhang, Wenwu Zhu
- abstract: Generating programs to describe visual observations has gained much research attention recently. However, most of the existing approaches are based on non-parametric primitive functions, making them unable to handle complex visual scenes involving many attributes and details. In this paper, we propose the concept of parametric visual program induction. Learning to generate parametric programs for visual scenes is challenging due to the huge number of function variants and the complex function correlations. To solve these challenges, we propose the method of function modularization, capable of dealing with numerous function variants and complex correlations. Specifically, we model each parametric function as a multi-head self-contained neural module to cover different function variants. Moreover, to eliminate the complex correlations between functions, we propose the hierarchical heterogeneous Monto-Carlo tree search (H2MCTS) algorithm which can provide high-quality uncorrelated supervision during training, and serve as an efficient searching technique during testing. We demonstrate the superiority of the proposed method on three visual program induction datasets involving parametric primitive functions. Experimental results show that our proposed model is able to significantly outperform the state-of-the-art baseline methods in terms of generating accurate programs.

Bayesian Deep Embedding Topic Meta-Learner

- Zhibin Duan, Yishi Xu, Jianqiao Sun, Bo Chen, Wenchao Chen, Chaojie Wang, Mingyuan Zhou
- abstract: Existing deep topic models are effective in capturing the latent semantic structures in textual data but usually rely on a plethora of documents. This is less than satisfactory in practical applications when only a limited amount of data is available. In this paper, we propose a novel framework that efficiently solves the problem of topic modeling under the small data regime. Specifically, the framework involves two innovations: a bi-level generative model that aims to exploit the task information to guide the document generation, and a topic meta-learner that strives to learn a group of global topic embeddings so that fast adaptation to the task-specific topic embeddings can be achieved with a few examples. We apply the proposed framework to a hierarchical embedded topic model and achieve better performance than various baseline models on diverse experiments, including few-shot topic discovery and few-shot document classification.

Deletion Robust Submodular Maximization over Matroids

- Paul Duetting, Federico Fusco, Silvio Lattanzi, Ashkan Norouzi-Fard, Morteza Zadimoghaddam
- abstract: Maximizing a monotone submodular function is a fundamental task in machine learning. In this paper we study the deletion robust version of the problem under the classic matroids constraint. Here the goal is to extract a small size summary of the dataset that contains a high value independent set even after an adversary deleted some elements. We present constant-factor approximation algorithms, whose space complexity depends on the rank k of the matroid and the number d of deleted elements. In the centralized setting we present a $(3.582+O(\varepsilon))$-approximation algorithm with summary size $O(k + \frac{d}{\eps^2}\log \frac{k}{\eps})$. In the streaming setting we provide a $(5.582+O(\varepsilon))$-approximation algorithm with summary size and memory $O(k + \frac{d}{\eps^2}\log \frac{k}{\eps})$. We complement our theoretical results with an in-depth experimental analysis showing the effectiveness of our algorithms on real-world datasets.

From data to functa: Your data point is a function and you can treat it like one

- Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo Jimenez Rezende, Dan Rosenbaum
- abstract: It is common practice in deep learning to represent a measurement of the world on a discrete grid, e.g. a 2D grid of pixels. However, the underlying signal represented by these measurements is often continuous, e.g. the scene depicted in an image. A powerful continuous alternative is then to represent these measurements using an implicit neural representation, a neural function trained to output the appropriate measurement value for any input spatial location. In this paper, we take this idea to its next level: what would it take to perform deep learning on these functions instead, treating them as

data? In this context we refer to the data as functa, and propose a framework for deep learning on functa. This view presents a number of challenges around efficient conversion from data to functa, compact representation of functa, and effectively solving downstream tasks on functa. We outline a recipe to overcome these challenges and apply it to a wide range of data modalities including images, 3D shapes, neural radiance fields (NeRF) and data on manifolds. We demonstrate that this approach has various compelling properties across data modalities, in particular on the canonical tasks of generative modeling, data imputation, novel view synthesis and classification.

Efficient Low Rank Convex Bounds for Pairwise Discrete Graphical Models

- Valentin Durante, George Katsirelos, Thomas Schiex
- abstract: In this paper, we extend a Burer-Monteiro style method to compute low rank Semi-Definite Programming (SDP) bounds for the MAP problem on discrete graphical models with an arbitrary number of states and arbitrary pairwise potentials. We consider both a penalized constraint approach and a dedicated Block Coordinate Descent (BCD) approach which avoids large penalty coefficients in the cost matrix. We show our algorithm is decreasing. Experiments show that the BCD approach compares favorably to the penalized approach and to usual linear bounds relying on convergent message passing approaches.

Robust Counterfactual Explanations for Tree-Based Ensembles

- Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, Daniele Magazzeni
- abstract: Counterfactual explanations inform ways to achieve a desired outcome from a machine learning model. However, such explanations are not robust to certain real-world changes in the underlying model (e.g., retraining the model, changing hyperparameters, etc.), questioning their reliability in several applications, e.g., credit lending. In this work, we propose a novel strategy - that we call RobX - to generate robust counterfactuals for tree-based ensembles, e.g., XGBoost. Tree-based ensembles pose additional challenges in robust counterfactual generation, e.g., they have a non-smooth and non-differentiable objective function, and they can change a lot in the parameter space under retraining on very similar data. We first introduce a novel metric - that we call Counterfactual Stability - that attempts to quantify how robust a counterfactual is going to be to model changes under retraining, and comes with desirable theoretical properties. Our proposed strategy RobX works with any counterfactual generation method (base method) and searches for robust counterfactuals by iteratively refining the counterfactual generated by the base method using our metric Counterfactual Stability. We compare the performance of RobX with popular counterfactual generation methods (for tree-based ensembles) across benchmark datasets. The results demonstrate that our strategy generates counterfactuals that are significantly more robust (nearly 100% validity after actual model changes) and also realistic (in terms of local outlier factor) over existing state-of-the-art methods.

On the Difficulty of Defending Self-Supervised Learning against Model Extraction

- Adam Dziedzic, Nikita Dhawan, Muhammad Ahmad Kaleem, Jonas Guan, Nicolas Papernot
- abstract: Self-Supervised Learning (SSL) is an increasingly popular ML paradigm that trains models to transform complex inputs into representations without relying on explicit labels. These representations encode similarity structures that enable efficient learning of multiple downstream tasks. Recently, ML-as-a-Service providers have commenced offering trained SSL models over inference APIs, which transform user inputs into useful representations for a fee. However, the high cost involved to train these models and their exposure over APIs both make black-box extraction a realistic security threat. We thus explore model stealing attacks against SSL. Unlike traditional model extraction on classifiers that output labels, the victim models here output representations; these representations are of significantly higher dimensionality compared to the low-dimensional prediction scores output by classifiers. We construct several novel attacks and find that approaches that train directly on a victim's stolen representations are query efficient and enable high accuracy for downstream models. We then show that existing defenses against model extraction are inadequate and not easily retrofitted to the specificities of SSL.

LIMO: Latent Inceptionism for Targeted Molecule Generation

- Peter Eckmann, Kunyang Sun, Bo Zhao, Mudong Feng, Michael Gilson, Rose Yu
- abstract: Generation of drug-like molecules with high binding affinity to target proteins remains a difficult and resource-intensive task in drug discovery. Existing approaches primarily employ reinforcement learning, Markov sampling, or deep generative models guided by Gaussian processes, which can be prohibitively slow when generating molecules with high binding affinity calculated by computationally-expensive physics-based methods. We present Latent Inceptionism on Molecules (LIMO), which significantly accelerates molecule generation with an inceptionism-like technique. LIMO employs a variational autoencoder-generated latent space and property prediction by two neural networks in sequence to enable faster gradient-based reverse-optimization of molecular properties. Comprehensive experiments show that LIMO performs competitively on benchmark tasks and markedly outperforms state-of-the-art techniques on the novel task of generating drug-like compounds with high binding affinity, reaching nanomolar range against two protein targets. We corroborate these docking-based results with more accurate molecular dynamics-based calculations of absolute binding free energy and show that one of our generated drug-like compounds has a predicted K_D (a measure of binding affinity) of $6 \cdot 10^{-14}$ M against the human estrogen receptor, well beyond the affinities of typical early-stage drug candidates and most FDA-approved drugs to their respective targets. Code is available at https://github.com/Rose-STL-Lab/LIMO.

Inductive Biases and Variable Creation in Self-Attention Mechanisms

- Benjamin L Edelman, Surbhi Goel, Sham Kakade, Cyril Zhang
- abstract: Self-attention, an architectural motif designed to model long-range interactions in sequential data, has driven numerous recent breakthroughs in natural language processing and beyond. This work provides a theoretical analysis of the inductive biases of self-attention modules. Our focus is to rigorously establish which functions and long-range dependencies self-attention blocks prefer to represent. Our main result shows that bounded-norm Transformer networks "create sparse variables": a single self-attention head can represent a sparse function of the input sequence, with sample complexity scaling only logarithmically with the context length. To support our analysis, we present synthetic experiments to probe the sample complexity of learning sparse Boolean functions with Transformers.

Provable Reinforcement Learning with a Short-Term Memory

- Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, Sobhan Miryoosefi
- abstract: Real-world sequential decision making problems commonly involve partial observability, which requires the agent to maintain a memory of history in order to infer the latent states, plan and make good decisions. Coping with partial observability in general is extremely challenging, as a number of worst-case statistical and computational barriers are known in learning Partially Observable Markov Decision Processes (POMDPs). Motivated by the problem structure in several physical applications, as well as a commonly used technique known as "frame stacking", this paper proposes to study a new subclass of POMDPs, whose latent states can be decoded by the most recent history of a short length m. We establish a set of upper and lower bounds on the sample complexity for learning near-optimal policies for this class of problems in both tabular and rich-observation settings (where the number of observations is enormous). In particular, in the rich-observation setting, we develop new algorithms using a novel "moment matching" approach with a sample complexity that scales exponentially with the short length m rather than the problem horizon, and is independent of the number of observations. Our results show that a short-term memory suffices for reinforcement learning in these environments.

Sparsity in Partially Controllable Linear Systems

- Yonathan Efroni, Sham Kakade, Akshay Krishnamurthy, Cyril Zhang
- abstract: A fundamental concept in control theory is that of controllability, where any system state can be reached through an appropriate choice of control inputs. Indeed, a large body of classical and modern approaches are designed for controllable linear dynamical systems. However, in practice, we often encounter systems in which a large set of state variables evolve exogenously and independently of the control inputs; such systems are only partially controllable. The focus of this work is on a large class of partially controllable linear dynamical systems, specified by an underlying sparsity pattern. Our main results establish structural conditions and finite-sample guarantees for learning to control such systems. In particular, our structural results characterize those state variables which are irrelevant for optimal control, an analysis which departs from classical control techniques. Our algorithmic results adapt techniques from high-dimensional statistics{—}specifically soft-thresholding and semiparametric least-squares{—}to exploit the underlying sparsity pattern in order to obtain finite-sample guarantees that significantly improve over those based on certainty-equivalence. We also corroborate these theoretical improvements over certainty-equivalent control through a simulation study.

FedNew: A Communication-Efficient and Privacy-Preserving Newton-Type Method for Federated Learning

- Anis Elgabli, Chaouki Ben Issaid, Amrit Singh Bedi, Ketan Rajawat, Mehdi Bennis, Vaneet Aggarwal
- abstract: Newton-type methods are popular in federated learning due to their fast convergence. Still, they suffer from two main issues, namely: low communication efficiency and low privacy due to the requirement of sending Hessian information from clients to parameter server (PS). In this work, we introduced a novel framework called FedNew in which there is no need to transmit Hessian information from clients to PS, hence resolving the bottleneck to improve communication efficiency. In addition, FedNew hides the gradient information and results in a privacy-preserving approach compared to the existing state-of-the-art. The core novel idea in FedNew is to introduce a two level framework, and alternate between updating the inverse Hessian-gradient product using only one alternating direction method of multipliers (ADMM) step and then performing the global model update using Newton's method. Though only one ADMM pass is used to approximate the inverse Hessian-gradient product at each iteration, we develop a novel theoretical approach to show the converging behavior of FedNew for convex problems. Additionally, a significant reduction in communication overhead is achieved by utilizing stochastic quantization. Numerical results using real datasets show the superiority of FedNew compared to existing methods in terms of communication costs.

pathGCN: Learning General Graph Spatial Operators from Paths

- Moshe Eliasof, Eldad Haber, Eran Treister
- abstract: Graph Convolutional Networks (GCNs), similarly to Convolutional Neural Networks (CNNs), are typically based on two main operations - spatial and point-wise convolutions. In the context of GCNs, differently from CNNs, a pre-determined spatial operator based on the graph Laplacian is often chosen, allowing only the point-wise operations to be learnt. However, learning a meaningful spatial operator is critical for developing more expressive GCNs for improved performance. In this paper we propose pathGCN, a novel approach to learn the spatial operator from random paths on the graph. We analyze the convergence of our method and its difference from existing GCNs. Furthermore, we discuss several options of combining our learnt spatial operator with point-wise convolutions. Our extensive experiments on numerous datasets suggest that by properly learning both the spatial and point-wise convolutions, phenomena like over-smoothing can be inherently avoided, and new state-of-the-art performance is achieved.

Discrete Tree Flows via Tree-Structured Permutations

- Mai Elkady, Hyung Zin Lim, David I Inouye
- abstract: While normalizing flows for continuous data have been extensively researched, flows for discrete data have only recently been explored. These prior models, however, suffer from limitations that are distinct from those of continuous flows. Most notably, discrete flow-based models cannot be straightforwardly optimized with conventional deep learning methods because gradients of discrete functions are undefined or zero. Previous works approximate pseudo-gradients of the discrete functions but do not solve the problem on a fundamental level. In addition to that, backpropagation can be computationally burdensome compared to alternative discrete algorithms such as decision tree algorithms. Our approach seeks to reduce computational burden and remove the need for pseudo-gradients by developing a discrete flow based on decision trees—building upon the success of efficient tree-based methods for classification and regression for discrete data. We first define a tree-structured permutation (TSP) that compactly encodes a permutation of discrete data where the inverse is easy to compute; thus, we can efficiently compute the density value and sample new data. We then propose a decision tree algorithm to build TSPs that learns the tree structure and permutations at each node via novel criteria. We empirically demonstrate the feasibility of our method on multiple datasets.

For Learning in Symmetric Teams, Local Optima are Global Nash Equilibria

- Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, Stuart Russell
- abstract: Although it has been known since the 1970s that a globally optimal strategy profile in a common-payoff game is a Nash equilibrium, global optimality is a strict requirement that limits the result's applicability. In this work, we show that any locally optimal symmetric strategy profile is also a (global) Nash equilibrium. Furthermore, we show that this result is robust to perturbations to the common payoff and to the local optimum. Applied to machine learning, our result provides a global guarantee for any gradient method that finds a local optimum in symmetric strategy space. While this result indicates stability to unilateral deviation, we nevertheless identify broad classes of games where mixed local optima are unstable under joint, asymmetric deviations. We analyze the prevalence of instability by running learning algorithms in a suite of symmetric games, and we conclude by discussing the applicability of our results to multi-agent RL, cooperative inverse RL, and decentralized POMDPs.

Streaming Algorithm for Monotone k-Submodular Maximization with Cardinality Constraints

- Alina Ene, Huy Nguyen
- abstract: Maximizing a monotone k-submodular function subject to cardinality constraints is a general model for several applications ranging from influence maximization with multiple products to sensor placement with multiple sensor types and online ad allocation. Due to the large problem scale in many applications and the online nature of ad allocation, a need arises for algorithms that process elements in a streaming fashion and possibly make online decisions. In this work, we develop a new streaming algorithm for maximizing a monotone k-submodular function subject to a per-coordinate cardinality constraint attaining an approximation guarantee close to the state of the art guarantee in the offline setting. Though not typical for streaming algorithms, our streaming algorithm also readily applies to the online setting with free disposal. Our algorithm is combinatorial and enjoys fast running time and small number of function evaluations. Furthermore, its guarantee improves as the cardinality constraints get larger, which is especially suited for the large scale applications. For the special case of maximizing a submodular function with large budgets, our combinatorial algorithm matches the guarantee of the state-of-the-art continuous algorithm, which requires significantly more time and function evaluations.

Towards Scaling Difference Target Propagation by Learning Backprop Targets

- Maxence M Ernoult, Fabrice Normandin, Abhinav Moudgil, Sean Spinney, Eugene Belilovsky, Irina Rish, Blake Richards, Yoshua Bengio
- abstract: The development of biologically-plausible learning algorithms is important for understanding learning in the brain, but most of them fail to scale-up to real-world tasks, limiting their potential as explanations for learning by real brains. As such, it is important to explore learning algorithms that come

with strong theoretical guarantees and can match the performance of backpropagation (BP) on complex tasks. One such algorithm is Difference Target Propagation (DTP), a biologically-plausible learning algorithm whose close relation with Gauss-Newton (GN) optimization has been recently established. However, the conditions under which this connection rigorously holds preclude layer-wise training of the feedback pathway synaptic weights (which is more biologically plausible). Moreover, good alignment between DTP weight updates and loss gradients is only loosely guaranteed and under very specific conditions for the architecture being trained. In this paper, we propose a novel feedback weight training scheme that ensures both that DTP approximates BP and that layer-wise feedback weight training can be restored without sacrificing any theoretical guarantees. Our theory is corroborated by experimental results and we report the best performance ever achieved by DTP on CIFAR-10 and ImageNet 32x32.

Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information

- Kawin Ethayarajh, Yejin Choi, Swabha Swayamdipta
- abstract: Estimating the difficulty of a dataset typically involves comparing state-of-the-art models to humans; the bigger the performance gap, the harder the dataset is said to be. However, this comparison provides little understanding of how difficult each instance in a given distribution is, or what attributes make the dataset difficult for a given model. To address these questions, we frame dataset difficulty—w.r.t. a model $\mathcal{V}$—as the lack of $\mathcal{V}$-usable information (Xu et al., 2019), where a lower value indicates a more difficult dataset for $\mathcal{V}$. We further introduce pointwise $\mathcal{V}$-information (PVI) for measuring the difficulty of individual instances w.r.t. a given distribution. While standard evaluation metrics typically only compare different models for the same dataset, $\mathcal{V}$-usable information and PVI also permit the converse: for a given model $\mathcal{V}$, we can compare different datasets, as well as different instances/slices of the same dataset. Furthermore, our framework allows for the interpretability of different input attributes via transformations of the input, which we use to discover annotation artefacts in widely-used NLP benchmarks.

Head2Toe: Utilizing Intermediate Representations for Better Transfer Learning

- Utku Evci, Vincent Dumoulin, Hugo Larochelle, Michael C Mozer
- abstract: Transfer-learning methods aim to improve performance in a data-scarce target domain using a model pretrained on a data-rich source domain. A cost-efficient strategy, linear probing, involves freezing the source model and training a new classification head for the target domain. This strategy is outperformed by a more costly but state-of-the-art method – fine-tuning all parameters of the source model to the target domain – possibly because fine-tuning allows the model to leverage useful information from intermediate layers which is otherwise discarded by the later previously trained layers. We explore the hypothesis that these intermediate layers might be directly exploited. We propose a method, Head-to-Toe probing (Head2Toe), that selects features from all layers of the source model to train a classification head for the target-domain. In evaluations on the Visual Task Adaptation Benchmark-1k, Head2Toe matches performance obtained with fine-tuning on average while reducing training and storage cost hundred folds or more, but critically, for out-of-distribution transfer, Head2Toe outperforms fine-tuning. Code used in our experiments can be found in supplementary materials.

Variational Sparse Coding with Learned Thresholding

- Kion Fallah, Christopher J Rozell
- abstract: Sparse coding strategies have been lauded for their parsimonious representations of data that leverage low dimensional structure. However, inference of these codes typically relies on an optimization procedure with poor computational scaling in high-dimensional problems. For example, sparse inference in the representations learned in the high-dimensional intermediary layers of deep neural networks (DNNs) requires an iterative minimization to be performed at each training step. As such, recent, quick methods in variational inference have been proposed to infer sparse codes by learning a distribution over the codes with a DNN. In this work, we propose a new approach to variational sparse coding that allows us to learn sparse distributions by thresholding samples, avoiding the use of problematic relaxations. We first evaluate and analyze our method by training a linear generator, showing that it has superior performance, statistical efficiency, and gradient estimation compared to other sparse distributions. We then compare to a standard variational autoencoder using a DNN generator on the CelebA dataset.

Training Discrete Deep Generative Models via Gapped Straight-Through Estimator

- Ting-Han Fan, Ta-Chung Chi, Alexander I. Rudnicky, Peter J Ramadge
- abstract: While deep generative models have succeeded in image processing, natural language processing, and reinforcement learning, training that involves discrete random variables remains challenging due to the high variance of its gradient estimation process. Monte Carlo is a common solution used in most variance reduction approaches. However, this involves time-consuming resampling and multiple function evaluations. We propose a Gapped Straight-Through (GST) estimator to reduce the variance without incurring resampling overhead. This estimator is inspired by the essential properties of Straight-Through Gumbel-Softmax. We determine these properties and show via an ablation study that they are essential. Experiments demonstrate that the proposed GST estimator enjoys better performance compared to strong baselines on two discrete deep generative modeling tasks, MNIST-VAE and ListOps.

DRIBO: Robust Deep Reinforcement Learning via Multi-View Information Bottleneck

- Jiameng Fan, Wenchao Li
- abstract: Deep reinforcement learning (DRL) agents are often sensitive to visual changes that were unseen in their training environments. To address this problem, we leverage the sequential nature of RL to learn robust representations that encode only task-relevant information from observations based on the unsupervised multi-view setting. Specif- ically, we introduce a novel contrastive version of the Multi-View Information Bottleneck (MIB) objective for temporal data. We train RL agents from pixels with this auxiliary objective to learn robust representations that can compress away task-irrelevant information and are predictive of task-relevant dynamics. This approach enables us to train high-performance policies that are robust to visual distractions and can generalize well to unseen environments. We demonstrate that our approach can achieve SOTA performance on a di- verse set of visual control tasks in the DeepMind Control Suite when the background is replaced with natural videos. In addition, we show that our approach outperforms well-established base- lines for generalization to unseen environments on the Procgen benchmark. Our code is open- sourced and available at https://github.com/BU-DEPEND-Lab/DRIBO.

Generalized Data Distribution Iteration

- Jiajun Fan, Changnan Xiao
- abstract: To obtain higher sample efficiency and superior final performance simultaneously has been one of the major challenges for deep reinforcement learning (DRL). Previous work could handle one of these challenges but typically failed to address them concurrently. In this paper, we try to tackle these two challenges simultaneously. To achieve this, we firstly decouple these challenges into two classic RL problems: data richness and exploration-exploitation trade-off. Then, we cast these two problems into the training data distribution optimization problem, namely to obtain desired training data within limited interactions, and address them concurrently via i) explicit modeling and control of the capacity and diversity of behavior policy and ii) more fine-grained and adaptive control of selective/sampling distribution of the behavior policy using a monotonic data distribution optimization. Finally, we integrate this process into Generalized Policy Iteration (GPI) and obtain a more general framework called Generalized Data Distribution Iteration (GDI). We use the GDI framework to introduce operator-based versions of well-known RL methods from DQN to Agent57. Theoretical guarantee of the superiority of GDI compared with GPI is concluded. We also demonstrate our state-of-the-art (SOTA) performance on Arcade Learning Environment (ALE), wherein our algorithm has achieved 9620.98% mean human normalized score (HNS), 1146.39% median HNS, and surpassed 22 human world

records using only 200M training frames. Our performance is comparable to Agent57's while we consume 500 times less data. We argue that there is still a long way to go before obtaining real superhuman agents in ALE.

Variational Wasserstein gradient flow

- Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, Yongxin Chen
- abstract: Wasserstein gradient flow has emerged as a promising approach to solve optimization problems over the space of probability distributions. A recent trend is to use the well-known JKO scheme in combination with input convex neural networks to numerically implement the proximal step. The most challenging step, in this setup, is to evaluate functions involving density explicitly, such as entropy, in terms of samples. This paper builds on the recent works with a slight but crucial difference: we propose to utilize a variational formulation of the objective function formulated as maximization over a parametric class of functions. Theoretically, the proposed variational formulation allows the construction of gradient flows directly for empirical distributions with a well-defined and meaningful objective function. Computationally, this approach replaces the computationally expensive step in existing methods, to handle objective functions involving density, with inner loop updates that only require a small batch of samples and scale well with the dimension. The performance and scalability of the proposed method are illustrated with the aid of several numerical experiments involving high-dimensional synthetic and real datasets.

Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP)

- Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, Ludwig Schmidt
- abstract: Contrastively trained language-image models such as CLIP, ALIGN, and BASIC have demonstrated unprecedented robustness to multiple challenging natural distribution shifts. Since these language-image models differ from previous training approaches in several ways, an important question is what causes the large robustness gains. We answer this question via a systematic experimental investigation. Concretely, we study five different possible causes for the robustness gains: (i) the training set size, (ii) the training distribution, (iii) language supervision at training time, (iv) language supervision at test time, and (v) the contrastive loss function. Our experiments show that the more diverse training distribution is the main cause for the robustness gains, with the other factors contributing little to no robustness. Beyond our experimental results, we also introduce ImageNet-Captions, a version of ImageNet with original text annotations from Flickr, to enable further controlled experiments of language-image training.

Bayesian Continuous-Time Tucker Decomposition

- Shikai Fang, Akil Narayan, Robert Kirby, Shandian Zhe
- abstract: Tensor decomposition is a dominant framework for multiway data analysis and prediction. Although practical data often contains timestamps for the observed entries, existing tensor decomposition approaches overlook or under-use this valuable time information. They either drop the timestamps or bin them into crude steps and hence ignore the temporal dynamics within each step or use simple parametric time coefficients. To overcome these limitations, we propose Bayesian Continuous-Time Tucker Decomposition. We model the tensor-core of the classical Tucker decomposition as a time-varying function, and place a Gaussian process prior to flexibly estimate all kinds of temporal dynamics. In this way, our model maintains the interpretability while is flexible enough to capture various complex temporal relationships between the tensor nodes. For efficient and high-quality posterior inference, we use the stochastic differential equation (SDE) representation of temporal GPs to build an equivalent state-space prior, which avoids huge kernel matrix computation and sparse/low-rank approximations. We then use Kalman filtering, RTS smoothing, and conditional moment matching to develop a scalable message passing inference algorithm. We show the advantage of our method in simulation and several real-world applications.

Byzantine Machine Learning Made Easy By Resilient Averaging of Momentums

- Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, John Stephan
- abstract: Byzantine resilience emerged as a prominent topic within the distributed machine learning community. Essentially, the goal is to enhance distributed optimization algorithms, such as distributed SGD, in a way that guarantees convergence despite the presence of some misbehaving (a.k.a., Byzantine) workers. Although a myriad of techniques addressing the problem have been proposed, the field arguably rests on fragile foundations. These techniques are hard to prove correct and rely on assumptions that are (a) quite unrealistic, i.e., often violated in practice, and (b) heterogeneous, i.e., making it difficult to compare approaches. We present RESAM (RESilient Averaging of Momentums), a unified framework that makes it simple to establish optimal Byzantine resilience, relying only on standard machine learning assumptions. Our framework is mainly composed of two operators: resilient averaging at the server and distributed momentum at the workers. We prove a general theorem stating the convergence of distributed SGD under RESAM. Interestingly, demonstrating and comparing the convergence of many existing techniques become direct corollaries of our theorem, without resorting to stringent assumptions. We also present an empirical evaluation of the practical relevance of RESAM.

An Equivalence Between Data Poisoning and Byzantine Gradient Attacks

- Sadegh Farhadkhani, Rachid Guerraoui, Lê Nguyên Hoang, Oscar Villemaud
- abstract: To study the resilience of distributed learning, the "Byzantine" literature considers a strong threat model where workers can report arbitrary gradients to the parameter server. Whereas this model helped obtain several fundamental results, it has sometimes been considered unrealistic, when the workers are mostly trustworthy machines. In this paper, we show a surprising equivalence between this model and data poisoning, a threat considered much more realistic. More specifically, we prove that every gradient attack can be reduced to data poisoning, in any personalized federated learning system with PAC guarantees (which we show are both desirable and realistic). This equivalence makes it possible to obtain new impossibility results on the resilience of any "robust" learning algorithm to data poisoning in highly heterogeneous applications, as corollaries of existing impossibility theorems on Byzantine machine learning. Moreover, using our equivalence, we derive a practical attack that we show (theoretically and empirically) can be very effective against classical personalized federated learning models.

Investigating Generalization by Controlling Normalized Margin

- Alexander R Farhang, Jeremy D Bernstein, Kushal Tirumala, Yang Liu, Yisong Yue
- abstract: Weight norm $\|w\|$ and margin γ participate in learning theory via the normalized margin $\gamma/\|w\|$. Since standard neural net optimizers do not control normalized margin, it is hard to test whether this quantity causally relates to generalization. This paper designs a series of experimental studies that explicitly control normalized margin and thereby tackle two central questions. First: does normalized margin always have a causal effect on generalization? The paper finds that no—networks can be produced where normalized margin has seemingly no relationship with generalization, counter to the theory of Bartlett et al. (2017). Second: does normalized margin ever have a causal effect on generalization? The paper finds that yes—in a standard training setup, test performance closely tracks normalized margin. The paper suggests a Gaussian process model as a promising explanation for this behavior.

Kernelized Multiplicative Weights for 0/1-Polyhedral Games: Bridging the Gap Between Learning in Extensive-Form and Normal-Form Games

- Gabriele Farina, Chung-Wei Lee, Haipeng Luo, Christian Kroer

- abstract: While extensive-form games (EFGs) can be converted into normal-form games (NFGs), doing so comes at the cost of an exponential blowup of the strategy space. So, progress on NFGs and EFGs has historically followed separate tracks, with the EFG community often having to catch up with advances (\eg last-iterate convergence and predictive regret bounds) from the larger NFG community. In this paper we show that the Optimistic Multiplicative Weights Update (OMWU) algorithm—the premier learning algorithm for NFGs—can be simulated on the normal-form equivalent of an EFG in linear time per iteration in the game tree size using a kernel trick. The resulting algorithm, Kernelized OMWU (KOMWU), applies more broadly to all convex games whose strategy space is a polytope with 0/1 integral vertices, as long as the kernel can be evaluated efficiently. In the particular case of EFGs, KOMWU closes several standing gaps between NFG and EFG learning, by enabling direct, black-box transfer to EFGs of desirable properties of learning dynamics that were so far known to be achievable only in NFGs. Specifically, KOMWU gives the first algorithm that guarantees at the same time last-iterate convergence, lower dependence on the size of the game tree than all prior algorithms, and $\tilde{\bigOh}(1)$ regret when followed by all players.

Local Linear Convergence of Douglas-Rachford for Linear Programming: a Probabilistic Analysis

- Oisin Faust, Hamza Fawzi
- abstract: Douglas-Rachford splitting/ADMM (henceforth DRS) is a very popular algorithm for solving convex optimisation problems to low or moderate accuracy, and in particular for solving large-scale linear programs. Despite recent progress, obtaining highly accurate solutions to linear programs with DRS remains elusive. In this paper we analyze the local linear convergence rate r of the DRS method for random linear programs, and give explicit and tight bounds on r. We show that $1-r^2$ is typically of the order of $m^{-1}(n-m)^{-1}$, where n is the number of variables and m is the number of constraints. This provides a quantitative explanation for the very slow convergence of DRS/ADMM on random LPs. The proof of our result relies on an established characterisation of the linear rate of convergence as the cosine of the Friedrichs angle between two subspaces associated to the problem. We also show that the cosecant of this angle can be interpreted as a condition number for the LP. The proof of our result relies on a characterization of the linear rate of convergence as the cosine of the Friedrichs angle between two subspaces associated to the problem. We also show that the cosecant of this angle can be interpreted as a condition number for the LP.

Matching Structure for Dual Learning

- Hao Fei, Shengqiong Wu, Yafeng Ren, Meishan Zhang
- abstract: Many natural language processing (NLP) tasks appear in dual forms, which are generally solved by dual learning technique that models the dualities between the coupled tasks. In this work, we propose to further enhance dual learning with structure matching that explicitly builds structural connections in between. Starting with the dual text$\leftrightarrow$text generation, we perform dually-syntactic structure co-echoing of the region of interest (RoI) between the task pair, together with a syntax cross-reconstruction at the decoding side. We next extend the idea to a text$\leftrightarrow$non-text setup, making alignment between the syntactic-semantic structure. Over 2*14 tasks covering 5 dual learning scenarios, the proposed structure matching method shows its significant effectiveness in enhancing existing dual learning. Our method can retrieve the key RoIs that are highly crucial to the task performance. Besides NLP tasks, it is also revealed that our approach has great potential in facilitating more non-text$\leftrightarrow$non-text scenarios.

Cascaded Gaps: Towards Logarithmic Regret for Risk-Sensitive Reinforcement Learning

- Yingjie Fei, Ruitu Xu
- abstract: In this paper, we study gap-dependent regret guarantees for risk-sensitive reinforcement learning based on the entropic risk measure. We propose a novel definition of sub-optimality gaps, which we call cascaded gaps, and we discuss their key components that adapt to underlying structures of the problem. Based on the cascaded gaps, we derive non-asymptotic and logarithmic regret bounds for two model-free algorithms under episodic Markov decision processes. We show that, in appropriate settings, these bounds feature exponential improvement over existing ones that are independent of gaps. We also prove gap-dependent lower bounds, which certify the near optimality of the upper bounds.

Private frequency estimation via projective geometry

- Vitaly Feldman, Jelani Nelson, Huy Nguyen, Kunal Talwar
- abstract: In this work, we propose a new algorithm ProjectiveGeometryResponse (PGR) for locally differentially private (LDP) frequency estimation. For universe size of k and with n users, our eps-LDP algorithm has communication cost ceil(log_2 k) and computation cost O(n + k\exp(eps) log k) for the server to approximately reconstruct the frequency histogram, while achieve optimal privacy-utility tradeoff. In many practical settings this is a significant improvement over the O (n+k^2) computation cost that is achieved by the recent PI-RAPPOR algorithm (Feldman and Talwar; 2021). Our empirical evaluation shows a speedup of over 50x over PI-RAPPOR while using approximately 75x less memory. In addition, the running time of our algorithm is comparable to that of HadamardResponse (Acharya, Sun, and Zhang; 2019) and RecursiveHadamardResponse (Chen, Kairouz, and Ozgur; 2020) which have significantly worse reconstruction error. The error of our algorithm essentially matches that of the communication- and time-inefficient but utility-optimal SubsetSelection (SS) algorithm (Ye and Barg; 2017). Our new algorithm is based on using Projective Planes over a finite field to define a small collection of sets that are close to being pairwise independent and a dynamic programming algorithm for approximate histogram reconstruction for the server.

An Intriguing Property of Geophysics Inversion

- Yinan Feng, Yinpeng Chen, Shihang Feng, Peng Jin, Zicheng Liu, Youzuo Lin
- abstract: Inversion techniques are widely used to reconstruct subsurface physical properties (e.g., velocity, conductivity) from surface-based geophysical measurements (e.g., seismic, electric/magnetic (EM) data). The problems are governed by partial differential equations (PDEs) like the wave or Maxwell's equations. Solving geophysical inversion problems is challenging due to the ill-posedness and high computational cost. To alleviate those issues, recent studies leverage deep neural networks to learn the inversion mappings from measurements to the property directly. In this paper, we show that such a mapping can be well modeled by a very shallow (but not wide) network with only five layers. This is achieved based on our new finding of an intriguing property: a near-linear relationship between the input and output, after applying integral transform in high dimensional space. In particular, when dealing with the inversion from seismic data to subsurface velocity governed by a wave equation, the integral results of velocity with Gaussian kernels are linearly correlated to the integral of seismic data with sine kernels. Furthermore, this property can be easily turned into a light-weight encoder-decoder network for inversion. The encoder contains the integration of seismic data and the linear transformation without need for fine-tuning. The decoder only consists of a single transformer block to reverse the integral of velocity. Experiments show that this interesting property holds for two geophysics inversion problems over four different datasets. Compared to much deeper InversionNet, our method achieves comparable accuracy, but consumes significantly fewer parameters

Principled Knowledge Extrapolation with GANs

- Ruili Feng, Jie Xiao, Kecheng Zheng, Deli Zhao, Jingren Zhou, Qibin Sun, Zheng-Jun Zha
- abstract: Human can extrapolate well, generalize daily knowledge into unseen scenarios, raise and answer counterfactual questions. To imitate this ability via generative models, previous works have extensively studied explicitly encoding Structural Causal Models (SCMs) into architectures of generator networks. This methodology, however, limits the flexibility of the generator as they must be carefully crafted to follow the causal graph, and demands a ground truth SCM with strong ignorability assumption as prior, which is a nontrivial assumption in many real scenarios. Thus, many current causal GAN

methods fail to generate high fidelity counterfactual results as they cannot easily leverage state-of-the-art generative models. In this paper, we propose to study counterfactual synthesis from a new perspective of knowledge extrapolation, where a given knowledge dimension of the data distribution is extrapolated, but the remaining knowledge is kept indistinguishable from the original distribution. We show that an adversarial game with a closed-form discriminator can be used to address the knowledge extrapolation problem, and a novel principal knowledge descent method can efficiently estimate the extrapolated distribution through the adversarial game. Our method enjoys both elegant theoretical guarantees and superior performance in many scenarios.

A Resilient Distributed Boosting Algorithm

- Yuval Filmus, Idan Mehalel, Shay Moran
- abstract: Given a learning task where the data is distributed among several parties, communication is one of the fundamental resources which the parties would like to minimize. We present a distributed boosting algorithm which is resilient to a limited amount of noise. Our algorithm is similar to classical boosting algorithms, although it is equipped with a new component, inspired by Impagliazzo's hard-core lemma (Impagliazzo, 1995), adding a robustness quality to the algorithm. We also complement this result by showing that resilience to any asymptotically larger noise is not achievable by a communication-efficient algorithm.

Model-Value Inconsistency as a Signal for Epistemic Uncertainty

- Angelos Filos, Eszter Vértes, Zita Marinho, Gregory Farquhar, Diana Borsa, Abram Friesen, Feryal Behbahani, Tom Schaul, Andre Barreto, Simon Osindero
- abstract: Using a model of the environment and a value function, an agent can construct many estimates of a state's value, by unrolling the model for different lengths and bootstrapping with its value function. Our key insight is that one can treat this set of value estimates as a type of ensemble, which we call an implicit value ensemble (IVE). Consequently, the discrepancy between these estimates can be used as a proxy for the agent's epistemic uncertainty; we term this signal model-value inconsistency or self-inconsistency for short. Unlike prior work which estimates uncertainty by training an ensemble of many models and/or value functions, this approach requires only the single model and value function which are already being learned in most model-based reinforcement learning algorithms. We provide empirical evidence in both tabular and function approximation settings from pixels that self-inconsistency is useful (i) as a signal for exploration, (ii) for acting safely under distribution shifts, and (iii) for robustifying value-based planning with a learned model.

Coordinated Double Machine Learning

- Nitai Fingerhut, Matteo Sesia, Yaniv Romano
- abstract: Double machine learning is a statistical method for leveraging complex black-box models to construct approximately unbiased treatment effect estimates given observational data with high-dimensional covariates, under the assumption of a partially linear model. The idea is to first fit on a subset of the samples two non-linear predictive models, one for the continuous outcome of interest and one for the observed treatment, and then to estimate a linear coefficient for the treatment using the remaining samples through a simple orthogonalized regression. While this methodology is flexible and can accommodate arbitrary predictive models, typically trained independently of one another, this paper argues that a carefully coordinated learning algorithm for deep neural networks may reduce the estimation bias. The improved empirical performance of the proposed method is demonstrated through numerical experiments on both simulated and real data.

Conformal Prediction Sets with Limited False Positives

- Adam Fisch, Tal Schuster, Tommi Jaakkola, Dr.Regina Barzilay
- abstract: We develop a new approach to multi-label conformal prediction in which we aim to output a precise set of promising prediction candidates with a bounded number of incorrect answers. Standard conformal prediction provides the ability to adapt to model uncertainty by constructing a calibrated candidate set in place of a single prediction, with guarantees that the set contains the correct answer with high probability. In order to obey this coverage property, however, conformal sets can become inundated with noisy candidates—which can render them unhelpful in practice. This is particularly relevant to practical applications where there is a limited budget, and the cost (monetary or otherwise) associated with false positives is non-negligible. We propose to trade coverage for a notion of precision by enforcing that the presence of incorrect candidates in the predicted conformal sets (i.e., the total number of false positives) is bounded according to a user-specified tolerance. Subject to this constraint, our algorithm then optimizes for a generalized notion of set coverage (i.e., the true positive rate) that allows for any number of true answers for a given query (including zero). We demonstrate the effectiveness of this approach across a number of classification tasks in natural language processing, computer vision, and computational chemistry.

Fast Population-Based Reinforcement Learning on a Single Machine

- Arthur Flajolet, Claire Bizon Monroc, Karim Beguir, Thomas Pierrot
- abstract: Training populations of agents has demonstrated great promise in Reinforcement Learning for stabilizing training, improving exploration and asymptotic performance, and generating a diverse set of solutions. However, population-based training is often not considered by practitioners as it is perceived to be either prohibitively slow (when implemented sequentially), or computationally expensive (if agents are trained in parallel on independent accelerators). In this work, we compare implementations and revisit previous studies to show that the judicious use of compilation and vectorization allows population-based training to be performed on a single machine with one accelerator with minimal overhead compared to training a single agent. We also show that, when provided with a few accelerators, our protocols extend to large population sizes for applications such as hyperparameter tuning. We hope that this work and the public release of our code will encourage practitioners to use population-based learning techniques more frequently for their research and applications.

Fast Relative Entropy Coding with A* coding

- Gergely Flamich, Stratis Markou, Jose Miguel Hernandez-Lobato
- abstract: Relative entropy coding (REC) algorithms encode a sample from a target distribution Q using a proposal distribution P, such that the expected codelength is O(KL[Q || P]). REC can be seamlessly integrated with existing learned compression models since, unlike entropy coding, it does not assume discrete Q or P, and does not require quantisation. However, general REC algorithms require an intractable Ω(exp(KL[Q || P])) runtime. We introduce AS *and AD* coding, two REC algorithms based on A *sampling*. *We prove that, for continuous distributions over the reals, if the density ratio is unimodal, AS* has O(D∞[Q || P]) expected runtime, where D∞[Q || P] is the Renyi ∞-divergence. We provide experimental evidence that AD *also has O(D∞[Q || P]) expected runtime*. *We prove that AS* and AD *achieve an expected codelength of O(KL[Q || P])*. *Further, we introduce DAD*, an approximate algorithm based on AD *which retains its favourable runtime and has bias similar to that of alternative methods*. *Focusing on VAEs, we propose the IsoKL VAE (IKVAE), which can be used with DAD* to further improve compression efficiency. We evaluate A* coding with (IK)VAEs on MNIST, showing that it can losslessly compress images near the theoretically optimal limit.

Contrastive Mixture of Posteriors for Counterfactual Inference, Data Integration and Fairness

- Adam Foster, Arpi Vezer, Craig A. Glastonbury, Paidi Creed, Samer Abujudeh, Aaron Sim

- abstract: Learning meaningful representations of data that can address challenges such as batch effect correction and counterfactual inference is a central problem in many domains including computational biology. Adopting a Conditional VAE framework, we show that marginal independence between the representation and a condition variable plays a key role in both of these challenges. We propose the Contrastive Mixture of Posteriors (CoMP) method that uses a novel misalignment penalty defined in terms of mixtures of the variational posteriors to enforce this independence in latent space. We show that CoMP has attractive theoretical properties compared to previous approaches, and we prove counterfactual identifiability of CoMP under additional assumptions. We demonstrate state-of-the-art performance on a set of challenging tasks including aligning human tumour samples with cancer cell-lines, predicting transcriptome-level perturbation responses, and batch correction on single-cell RNA sequencing data. We also find parallels to fair representation learning and demonstrate that CoMP is competitive on a common task in the field.

Label Ranking through Nonparametric Regression

- Dimitris Fotakis, Alkis Kalavasis, Eleni Psaroudaki
- abstract: Label Ranking (LR) corresponds to the problem of learning a hypothesis that maps features to rankings over a finite set of labels. We adopt a nonparametric regression approach to LR and obtain theoretical performance guarantees for this fundamental practical problem. We introduce a generative model for Label Ranking, in noiseless and noisy nonparametric regression settings, and provide sample complexity bounds for learning algorithms in both cases. In the noiseless setting, we study the LR problem with full rankings and provide computationally efficient algorithms using decision trees and random forests in the high-dimensional regime. In the noisy setting, we consider the more general cases of LR with incomplete and partial rankings from a statistical viewpoint and obtain sample complexity bounds using the One-Versus-One approach of multiclass classification. Finally, we complement our theoretical contributions with experiments, aiming to understand how the input regression noise affects the observed output.

A Neural Tangent Kernel Perspective of GANs

- Jean-Yves Franceschi, Emmanuel De Bézenac, Ibrahim Ayed, Mickael Chen, Sylvain Lamprier, Patrick Gallinari
- abstract: We propose a novel theoretical framework of analysis for Generative Adversarial Networks (GANs). We reveal a fundamental flaw of previous analyses which, by incorrectly modeling GANs' training scheme, are subject to ill-defined discriminator gradients. We overcome this issue which impedes a principled study of GAN training, solving it within our framework by taking into account the discriminator's architecture. To this end, we leverage the theory of infinite-width neural networks for the discriminator via its Neural Tangent Kernel. We characterize the trained discriminator for a wide range of losses and establish general differentiability properties of the network. From this, we derive new insights about the convergence of the generated distribution, advancing our understanding of GANs' training dynamics. We empirically corroborate these results via an analysis toolkit based on our framework, unveiling intuitions that are consistent with GAN practice.

Extracting Latent State Representations with Linear Dynamics from Rich Observations

- Abraham Frandsen, Rong Ge, Holden Lee
- abstract: Recently, many reinforcement learning techniques have been shown to have provable guarantees in the simple case of linear dynamics, especially in problems like linear quadratic regulators. However, in practice many tasks require learning a policy from rich, high-dimensional features such as images, which are unlikely to be linear. We consider a setting where there is a hidden linear subspace of the high-dimensional feature space in which the dynamics are linear. We design natural objectives based on forward and inverse dynamics models. We prove that these objectives can be efficiently optimized and their local optimizers extract the hidden linear subspace. We empirically verify our theoretical results with synthetic data and explore the effectiveness of our approach (generalized to nonlinear settings) in simple control tasks with rich observations.

SPDY: Accurate Pruning with Speedup Guarantees

- Elias Frantar, Dan Alistarh
- abstract: The recent focus on the efficiency of deep neural networks (DNNs) has led to significant work on model compression approaches, of which weight pruning is one of the most popular. At the same time, there is rapidly-growing computational support for efficiently executing the unstructured-sparse models obtained via pruning. Yet, most existing pruning methods minimize just the number of remaining weights, i.e. the size of the model, rather than optimizing for inference time. We address this gap by introducing SPDY, a new compression method which automatically determines layer-wise sparsity targets achieving a desired inference speedup on a given system, while minimizing accuracy loss. SPDY is the composition of two new techniques. The first is an efficient and general dynamic programming algorithm for solving constrained layer-wise compression problems, given a set of layer-wise error scores. The second technique is a local search procedure for automatically determining such scores in an accurate and robust manner. Experiments across popular vision and language models show that SPDY guarantees speedups while recovering higher accuracy relative to existing strategies, both for one-shot and gradual pruning scenarios, and is compatible with most existing pruning approaches. We also extend our approach to the recently-proposed task of pruning with very little data, where we achieve the best known accuracy recovery when pruning to the GPU-supported 2:4 sparsity pattern.

Revisiting the Effects of Stochasticity for Hamiltonian Samplers

- Giulio Franzese, Dimitrios Milios, Maurizio Filippone, Pietro Michiardi
- abstract: We revisit the theoretical properties of Hamiltonian stochastic differential equations (SDES) for Bayesian posterior sampling, and we study the two types of errors that arise from numerical SDE simulation: the discretization error and the error due to noisy gradient estimates in the context of data subsampling. Our main result is a novel analysis for the effect of mini-batches through the lens of differential operator splitting, revising previous literature results. The stochastic component of a Hamiltonian SDE is decoupled from the gradient noise, for which we make no normality assumptions. This leads to the identification of a convergence bottleneck: when considering mini-batches, the best achievable error rate is $\mathcal{O}(\eta^2)$, with η being the integrator step size. Our theoretical results are supported by an empirical study on a variety of regression and classification tasks for Bayesian neural networks.

Bregman Neural Networks

- Jordan Frecon, Gilles Gasso, Massimiliano Pontil, Saverio Salzo
- abstract: We present a framework based on bilevel optimization for learning multilayer, deep data representations. On the one hand, the lower-level problem finds a representation by successively minimizing layer-wise objectives made of the sum of a prescribed regularizer as well as a fidelity term and some linear function both depending on the representation found at the previous layer. On the other hand, the upper-level problem optimizes over the linear functions to yield a linearly separable final representation. We show that, by choosing the fidelity term as the quadratic distance between two successive layer-wise representations, the bilevel problem reduces to the training of a feed-forward neural network. Instead, by elaborating on Bregman distances, we devise a novel neural network architecture additionally involving the inverse of the activation function reminiscent of the skip connection used in ResNets. Numerical experiments suggest that the proposed Bregman variant benefits from better learning properties and more robust prediction performance.

(Non-)Convergence Results for Predictive Coding Networks

- Simon Frieder, Thomas Lukasiewicz
- abstract: Predictive coding networks (PCNs) are (un)supervised learning models, coming from neuroscience, that approximate how the brain works. One major open problem around PCNs is their convergence behavior. In this paper, we use dynamical systems theory to formally investigate the convergence of PCNs as they are used in machine learning. Doing so, we put their theory on a firm, rigorous basis, by developing a precise mathematical framework for PCN and show that for sufficiently small weights and initializations, PCNs converge for any input. Thereby, we provide the theoretical assurance that previous implementations, whose convergence was assessed solely by numerical experiments, can indeed capture the correct behavior of PCNs. Outside of the identified regime of small weights and small initializations, we show via a counterexample that PCNs can diverge, countering common beliefs held in the community. This is achieved by identifying a Neimark-Sacker bifurcation in a PCN of small size, which gives rise to an unstable fixed point and an invariant curve around it.

Scaling Structured Inference with Randomization

- Yao Fu, John Cunningham, Mirella Lapata
- abstract: Deep discrete structured models have seen considerable progress recently, but traditional inference using dynamic programming (DP) typically works with a small number of states (less than hundreds), which severely limits model capacity. At the same time, across machine learning, there is a recent trend of using randomized truncation techniques to accelerate computations involving large sums. Here, we propose a family of randomized dynamic programming (RDP) algorithms for scaling structured models to tens of thousands of latent states. Our method is widely applicable to classical DP-based inference (partition, marginal, reparameterization, entropy) and different graph structures (chains, trees, and more general hypergraphs). It is also compatible with automatic differentiation: it can be integrated with neural networks seamlessly and learned with gradient-based optimizers. Our core technique approximates the sum-product by restricting and reweighting DP on a small subset of nodes, which reduces computation by orders of magnitude. We further achieve low bias and variance via Rao-Blackwellization and importance sampling. Experiments over different graphs demonstrate the accuracy and efficiency of our approach. Furthermore, when using RDP for training a structured variational autoencoder with a scaled inference network, we achieve better test likelihood than baselines and successfully prevent posterior collapse.

Greedy when Sure and Conservative when Uncertain about the Opponents

- Haobo Fu, Ye Tian, Hongxiang Yu, Weiming Liu, Shuang Wu, Jiechao Xiong, Ying Wen, Kai Li, Junliang Xing, Qiang Fu, Wei Yang
- abstract: We develop a new approach, named Greedy when Sure and Conservative when Uncertain (GSCU), to competing online against unknown and nonstationary opponents. GSCU improves in four aspects: 1) introduces a novel way of learning opponent policy embeddings offline; 2) trains offline a single best response (conditional additionally on our opponent policy embedding) instead of a finite set of separate best responses against any opponent; 3) computes online a posterior of the current opponent policy embedding, without making the discrete and ineffective decision which type the current opponent belongs to; and 4) selects online between a real-time greedy policy and a fixed conservative policy via an adversarial bandit algorithm, gaining a theoretically better regret than adhering to either. Experimental studies on popular benchmarks demonstrate GSCU's superiority over the state-of-the-art methods. The code is available online at \url{https://github.com/YeTianJHU/GSCU}.

DepthShrinker: A New Compression Paradigm Towards Boosting Real-Hardware Efficiency of Compact Neural Networks

- Yonggan Fu, Haichuan Yang, Jiayi Yuan, Meng Li, Cheng Wan, Raghuraman Krishnamoorthi, Vikas Chandra, Yingyan Lin
- abstract: Efficient deep neural network (DNN) models equipped with compact operators (e.g., depthwise convolutions) have shown great potential in reducing DNNs' theoretical complexity (e.g., the total number of weights/operations) while maintaining a decent model accuracy. However, existing efficient DNNs are still limited in fulfilling their promise in boosting real-hardware efficiency, due to their commonly adopted compact operators' low hardware utilization. In this work, we open up a new compression paradigm for developing real-hardware efficient DNNs, leading to boosted hardware efficiency while maintaining model accuracy. Interestingly, we observe that while some DNN layers' activation functions help DNNs' training optimization and achievable accuracy, they can be properly removed after training without compromising the model accuracy. Inspired by this observation, we propose a framework dubbed DepthShrinker, which develops hardware-friendly compact networks via shrinking the basic building blocks of existing efficient DNNs that feature irregular computation patterns into dense ones with much improved hardware utilization and thus real-hardware efficiency. Excitingly, our DepthShrinker framework delivers hardware-friendly compact networks that outperform both state-of-the-art efficient DNNs and compression techniques, e.g., a 3.06% higher accuracy and 1.53x throughput on Tesla V100 over SOTA channel-wise pruning method MetaPruning. Our codes are available at: https://github.com/facebookresearch/DepthShrinker.

Revisiting Some Common Practices in Cooperative Multi-Agent Reinforcement Learning

- Wei Fu, Chao Yu, Zelai Xu, Jiaqi Yang, Yi Wu
- abstract: Many advances in cooperative multi-agent reinforcement learning (MARL) are based on two common design principles: value decomposition and parameter sharing. A typical MARL algorithm of this fashion decomposes a centralized Q-function into local Q-networks with parameters shared across agents. Such an algorithmic paradigm enables centralized training and decentralized execution (CTDE) and leads to efficient learning in practice. Despite all the advantages, we revisit these two principles and show that in certain scenarios, e.g., environments with a highly multi-modal reward landscape, value decomposition, and parameter sharing can be problematic and lead to undesired outcomes. In contrast, policy gradient (PG) methods with individual policies provably converge to an optimal solution in these cases, which partially supports some recent empirical observations that PG can be effective in many MARL testbeds. Inspired by our theoretical analysis, we present practical suggestions on implementing multi-agent PG algorithms for either high rewards or diverse emergent behaviors and empirically validate our findings on a variety of domains, ranging from the simplified matrix and grid-world games to complex benchmarks such as StarCraft Multi-Agent Challenge and Google Research Football. We hope our insights could benefit the community towards developing more general and more powerful MARL algorithms.

p-Laplacian Based Graph Neural Networks

- Guoji Fu, Peilin Zhao, Yatao Bian
- abstract: Graph neural networks (GNNs) have demonstrated superior performance for semi-supervised node classification on graphs, as a result of their ability to exploit node features and topological information simultaneously. However, most GNNs implicitly assume that the labels of nodes and their neighbors in a graph are the same or consistent, which does not hold in heterophilic graphs, where the labels of linked nodes are likely to differ. Moreover, when the topology is non-informative for label prediction, ordinary GNNs may work significantly worse than simply applying multi-layer perceptrons (MLPs) on each node. To tackle the above problem, we propose a new p-Laplacian based GNN model, termed as ^{p}GNN, whose message passing mechanism is derived from a discrete regularization framework and could be theoretically explained as an approximation of a polynomial graph filter defined on the spectral domain of p-Laplacians. The spectral analysis shows that the new message passing mechanism works as low-high-pass filters, thus making ^{p}GNNs are effective on both homophilic and heterophilic graphs. Empirical studies on real-world and synthetic datasets validate our findings and demonstrate that ^{p}GNNs significantly outperform several state-of-the-art GNN architectures on heterophilic benchmarks while achieving competitive performance on homophilic benchmarks. Moreover, ^{p}GNNs can adaptively learn aggregation weights and are robust to noisy edges.

Why Should I Trust You, Bellman? The Bellman Error is a Poor Replacement for Value Error

- Scott Fujimoto, David Meger, Doina Precup, Ofir Nachum, Shixiang Shane Gu

- abstract: In this work, we study the use of the Bellman equation as a surrogate objective for value prediction accuracy. While the Bellman equation is uniquely solved by the true value function over all state-action pairs, we find that the Bellman error (the difference between both sides of the equation) is a poor proxy for the accuracy of the value function. In particular, we show that (1) due to cancellations from both sides of the Bellman equation, the magnitude of the Bellman error is only weakly related to the distance to the true value function, even when considering all state-action pairs, and (2) in the finite data regime, the Bellman equation can be satisfied exactly by infinitely many suboptimal solutions. This means that the Bellman error can be minimized without improving the accuracy of the value function. We demonstrate these phenomena through a series of propositions, illustrative toy examples, and empirical analysis in standard benchmark domains.

Robin Hood and Matthew Effects: Differential Privacy Has Disparate Impact on Synthetic Data

- Georgi Ganev, Bristena Oprisanu, Emiliano De Cristofaro
- abstract: Generative models trained with Differential Privacy (DP) can be used to generate synthetic data while minimizing privacy risks. We analyze the impact of DP on these models vis-a-vis underrepresented classes/subgroups of data, specifically, studying: 1) the size of classes/subgroups in the synthetic data, and 2) the accuracy of classification tasks run on them. We also evaluate the effect of various levels of imbalance and privacy budgets. Our analysis uses three state-of-the-art DP models (PrivBayes, DP-WGAN, and PATE-GAN) and shows that DP yields opposite size distributions in the generated synthetic data. It affects the gap between the majority and minority classes/subgroups; in some cases by reducing it (a "Robin Hood" effect) and, in others, by increasing it (a "Matthew" effect). Either way, this leads to (similar) disparate impacts on the accuracy of classification tasks on the synthetic data, affecting disproportionately more the underrepresented subparts of the data. Consequently, when training models on synthetic data, one might incur the risk of treating different subpopulations unevenly, leading to unreliable or unfair conclusions.

The Complexity of k-Means Clustering when Little is Known

- Robert Ganian, Thekla Hamm, Viktoriia Korchemna, Karolina Okrasa, Kirill Simonov
- abstract: In the area of data analysis and arguably even in machine learning as a whole, few approaches have been as impactful as the classical k-means clustering. Here, we study the complexity of k-means clustering in settings where most of the data is not known or simply irrelevant. To obtain a more fine-grained understanding of the tractability of this clustering problem, we apply the parameterized complexity paradigm and obtain three new algorithms for k-means clustering of incomplete data: one for the clustering of bounded-domain (i.e., integer) data, and two incomparable algorithms that target real-valued data. Our approach is based on exploiting structural properties of a graphical encoding of the missing entries, and we show that tractability can be achieved using significantly less restrictive parameterizations than in the complementary case of few missing entries.

IDYNO: Learning Nonparametric DAGs from Interventional Dynamic Data

- Tian Gao, Debarun Bhattacharjya, Elliot Nelson, Miao Liu, Yue Yu
- abstract: Causal discovery in the form of a directed acyclic graph (DAG) for time series data has been widely studied in various domains. The resulting DAG typically represents a dynamic Bayesian network (DBN), capturing both the instantaneous and time-delayed relationships among variables of interest. We propose a new algorithm, IDYNO, to learn the DAG structure from potentially nonlinear times series data by using a continuous optimization framework that includes a recent formulation for continuous acyclicity constraint. The proposed algorithm is designed to handle both observational and interventional time series data. We demonstrate the promising performance of our method on synthetic benchmark datasets against state-of-the-art baselines. In addition, we show that the proposed method can more accurately learn the underlying structure of a sequential decision model, such as a Markov decision process, with a fixed policy in typical continuous control tasks.

Loss Function Learning for Domain Generalization by Implicit Gradient

- Boyan Gao, Henry Gouk, Yongxin Yang, Timothy Hospedales
- abstract: Generalising robustly to distribution shift is a major challenge that is pervasive across most real-world applications of machine learning. A recent study highlighted that many advanced algorithms proposed to tackle such domain generalisation (DG) fail to outperform a properly tuned empirical risk minimisation (ERM) baseline. We take a different approach, and explore the impact of the ERM loss function on out-of-domain generalisation. In particular, we introduce a novel meta-learning approach to loss function search based on implicit gradient. This enables us to discover a general purpose parametric loss function that provides a drop-in replacement for cross-entropy. Our loss can be used in standard training pipelines to efficiently train robust models using any neural architecture on new datasets. The results show that it clearly surpasses cross-entropy, enables simple ERM to outperform some more complicated prior DG methods, and provides state-of-the-art performance across a variety of DG benchmarks. Furthermore, unlike most existing DG approaches, our setup applies to the most practical setting of single-source domain generalisation, on which we show significant improvement.

On the Convergence of Local Stochastic Compositional Gradient Descent with Momentum

- Hongchang Gao, Junyi Li, Heng Huang
- abstract: Federated Learning has been actively studied due to its efficiency in numerous real-world applications in the past few years. However, the federated stochastic compositional optimization problem is still underexplored, even though it has widespread applications in machine learning. In this paper, we developed a novel local stochastic compositional gradient descent with momentum method, which facilitates Federated Learning for the stochastic compositional problem. Importantly, we investigated the convergence rate of our proposed method and proved that it can achieve the $O(1/\epsilon^4)$ sample complexity, which is better than existing methods. Meanwhile, our communication complexity $O(1/\epsilon^3)$ can match existing methods. To the best of our knowledge, this is the first work achieving such favorable sample and communication complexities. Additionally, extensive experimental results demonstrate the superior empirical performance over existing methods, confirming the efficacy of our method.

Deep Reference Priors: What is the best way to pretrain a model?

- Yansong Gao, Rahul Ramesh, Pratik Chaudhari
- abstract: What is the best way to exploit extra data – be it unlabeled data from the same task, or labeled data from a related task – to learn a given task? This paper formalizes the question using the theory of reference priors. Reference priors are objective, uninformative Bayesian priors that maximize the mutual information between the task and the weights of the model. Such priors enable the task to maximally affect the Bayesian posterior, e.g., reference priors depend upon the number of samples available for learning the task and for very small sample sizes, the prior puts more probability mass on low-complexity models in the hypothesis space. This paper presents the first demonstration of reference priors for medium-scale deep networks and image-based data. We develop generalizations of reference priors and demonstrate applications to two problems. First, by using unlabeled data to compute the reference prior, we develop new Bayesian semi-supervised learning methods that remain effective even with very few samples per class. Second, by using labeled data from the source task to compute the reference prior, we develop a new pretraining method for transfer learning that allows data from the target task to maximally affect the Bayesian posterior. Empirical validation of these methods is conducted on image classification datasets. Code is available at https://github.com/grasp-lyrl/deep_reference_priors

On the Equivalence Between Temporal and Static Equivariant Graph Representations

- Jianfei Gao, Bruno Ribeiro

- abstract: This work formalizes the associational task of predicting node attribute evolution in temporal graphs from the perspective of learning equivariant representations. We show that node representations in temporal graphs can be cast into two distinct frameworks: (a) The most popular approach, which we denote as time-and-graph, where equivariant graph (e.g., GNN) and sequence (e.g., RNN) representations are intertwined to represent the temporal evolution of node attributes in the graph; and (b) an approach that we denote as time-then-graph, where the sequences describing the node and edge dynamics are represented first, then fed as node and edge attributes into a static equivariant graph representation that comes after. Interestingly, we show that time-then-graph representations have an expressivity advantage over time-and-graph representations when both use component GNNs that are not most-expressive (e.g., 1-Weisfeiler-Lehman GNNs). Moreover, while our goal is not necessarily to obtain state-of-the-art results, our experiments show that time-then-graph methods are capable of achieving better performance and efficiency than state-of-the-art time-and-graph methods in some real-world tasks, thereby showcasing that the time-then-graph framework is a worthy addition to the graph ML toolbox.

Generalizing Gaussian Smoothing for Random Search

- Katelyn Gao, Ozan Sener
- abstract: Gaussian smoothing (GS) is a derivative-free optimization (DFO) algorithm that estimates the gradient of an objective using perturbations of the current parameters sampled from a standard normal distribution. We generalize it to sampling perturbations from a larger family of distributions. Based on an analysis of DFO for non-convex functions, we propose to choose a distribution for perturbations that minimizes the mean squared error (MSE) of the gradient estimate. We derive three such distributions with provably smaller MSE than Gaussian smoothing. We conduct evaluations of the three sampling distributions on linear regression, reinforcement learning, and DFO benchmarks in order to validate our claims. Our proposal improves on GS with the same computational complexity, and are competitive with and usually outperform Guided ES and Orthogonal ES, two computationally more expensive algorithms that adapt the covariance matrix of normally distributed perturbations.

Rethinking Image-Scaling Attacks: The Interplay Between Vulnerabilities in Machine Learning Systems

- Yue Gao, Ilia Shumailov, Kassem Fawaz
- abstract: As real-world images come in varying sizes, the machine learning model is part of a larger system that includes an upstream image scaling algorithm. In this paper, we investigate the interplay between vulnerabilities of the image scaling procedure and machine learning models in the decision-based black-box setting. We propose a novel sampling strategy to make a black-box attack exploit vulnerabilities in scaling algorithms, scaling defenses, and the final machine learning model in an end-to-end manner. Based on this scaling-aware attack, we reveal that most existing scaling defenses are ineffective under threat from downstream models. Moreover, we empirically observe that standard black-box attacks can significantly improve their performance by exploiting the vulnerable scaling procedure. We further demonstrate this problem on a commercial Image Analysis API with decision-based black-box attacks.

Lazy Estimation of Variable Importance for Large Neural Networks

- Yue Gao, Abby Stevens, Garvesh Raskutti, Rebecca Willett
- abstract: As opaque predictive models increasingly impact many areas of modern life, interest in quantifying the importance of a given input variable for making a specific prediction has grown. Recently, there has been a proliferation of model-agnostic methods to measure variable importance (VI) that analyze the difference in predictive power between a full model trained on all variables and a reduced model that excludes the variable(s) of interest. A bottleneck common to these methods is the estimation of the reduced model for each variable (or subset of variables), which is an expensive process that often does not come with theoretical guarantees. In this work, we propose a fast and flexible method for approximating the reduced model with important inferential guarantees. We replace the need for fully retraining a wide neural network by a linearization initialized at the full model parameters. By adding a ridge-like penalty to make the problem convex, we prove that when the ridge penalty parameter is sufficiently large, our method estimates the variable importance measure with an error rate of O(1/n) where n is the number of training samples. We also show that our estimator is asymptotically normal, enabling us to provide confidence bounds for the VI estimates. We demonstrate through simulations that our method is fast and accurate under several data-generating regimes, and we demonstrate its real-world applicability on a seasonal climate forecasting example.

Fast and Reliable Evaluation of Adversarial Robustness with Minimum-Margin Attack

- Ruize Gao, Jiongxiao Wang, Kaiwen Zhou, Feng Liu, Binghui Xie, Gang Niu, Bo Han, James Cheng
- abstract: The AutoAttack (AA) has been the most reliable method to evaluate adversarial robustness when considerable computational resources are available. However, the high computational cost (e.g., 100 times more than that of the project gradient descent attack) makes AA infeasible for practitioners with limited computational resources, and also hinders applications of AA in the adversarial training (AT). In this paper, we propose a novel method, minimum-margin (MM) attack, to fast and reliably evaluate adversarial robustness. Compared with AA, our method achieves comparable performance but only costs 3% of the computational time in extensive experiments. The reliability of our method lies in that we evaluate the quality of adversarial examples using the margin between two targets that can precisely identify the most adversarial example. The computational efficiency of our method lies in an effective Sequential TArget Ranking Selection (STARS) method, ensuring that the cost of the MM attack is independent of the number of classes. The MM attack opens a new way for evaluating adversarial robustness and provides a feasible and reliable way to generate high-quality adversarial examples in AT.

Value Function based Difference-of-Convex Algorithm for Bilevel Hyperparameter Selection Problems

- Lucy L Gao, Jane Ye, Haian Yin, Shangzhi Zeng, Jin Zhang
- abstract: Existing gradient-based optimization methods for hyperparameter tuning can only guarantee theoretical convergence to stationary solutions when the bilevel program satisfies the condition that for fixed upper-level variables, the lower-level is strongly convex (LLSC) and smooth (LLS). This condition is not satisfied for bilevel programs arising from tuning hyperparameters in many machine learning algorithms. In this work, we develop a sequentially convergent Value Function based Difference-of-Convex Algorithm with inexactness (VF-iDCA). We then ask: can this algorithm achieve stationary solutions without LLSC and LLS assumptions? We provide a positive answer to this question for bilevel programs from a broad class of hyperparameter tuning applications. Extensive experiments justify our theoretical results and demonstrate the superiority of the proposed VF-iDCA when applied to tune hyperparameters.

Learning to Incorporate Texture Saliency Adaptive Attention to Image Cartoonization

- Xiang Gao, Yuqi Zhang, Yingjie Tian
- abstract: Image cartoonization is recently dominated by generative adversarial networks (GANs) from the perspective of unsupervised image-to-image translation, in which an inherent challenge is to precisely capture and sufficiently transfer characteristic cartoon styles (e.g., clear edges, smooth color shading, vivid colors, etc.). Existing advanced models try to enhance cartoonization effect by learning to promote edges adversarially, introducing style transfer loss, or learning to align style from multiple representation space. This paper demonstrates that more distinct and vivid cartoonization effect could be easily achieved with only basic adversarial loss. Observing that cartoon style is more evident in cartoon-texture-salient local image regions, we build a region-level adversarial learning branch in parallel with the normal image-level one, which constrains adversarial learning on cartoon-texture-salient local patches for better perceiving and transferring cartoon texture features. To this end, a novel cartoon-texture-saliency-sampler (CTSS) module is proposed to adaptively sample cartoon-texture-salient patches from training data. We present that such texture saliency adaptive attention is of significant importance

in facilitating and enhancing cartoon stylization, which is a key missing ingredient of related methods. The superiority of our model in promoting cartoonization effect, especially for high-resolution input images, are fully demonstrated with extensive experiments.

Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification

- Camille Garcin, Maximilien Servajean, Alexis Joly, Joseph Salmon
- abstract: In modern classification tasks, the number of labels is getting larger and larger, as is the size of the datasets encountered in practice. As the number of classes increases, class ambiguity and class imbalance become more and more problematic to achieve high top-1 accuracy. Meanwhile, Top-K metrics (metrics allowing K guesses) have become popular, especially for performance reporting. Yet, proposing top-K losses tailored for deep learning remains a challenge, both theoretically and practically. In this paper we introduce a stochastic top-K hinge loss inspired by recent developments on top-K calibrated losses. Our proposal is based on the smoothing of the top-K operator building on the flexible "perturbed optimizer" framework. We show that our loss function performs very well in the case of balanced datasets, while benefiting from a significantly lower computational time than the state-of-the-art top-K loss function. In addition, we propose a simple variant of our loss for the imbalanced case. Experiments on a heavy-tailed dataset show that our loss function significantly outperforms other baseline loss functions.

PAGE-PG: A Simple and Loopless Variance-Reduced Policy Gradient Method with Probabilistic Gradient Estimation

- Matilde Gargiani, Andrea Zanelli, Andrea Martinelli, Tyler Summers, John Lygeros
- abstract: Despite their success, policy gradient methods suffer from high variance of the gradient estimator, which can result in unsatisfactory sample complexity. Recently, numerous variance-reduced extensions of policy gradient methods with provably better sample complexity and competitive numerical performance have been proposed. After a compact survey on some of the main variance-reduced REINFORCE-type methods, we propose ProbAbilistic Gradient Estimation for Policy Gradient (PAGE-PG), a novel loopless variance-reduced policy gradient method based on a probabilistic switch between two types of update. Our method is inspired by the PAGE estimator for supervised learning and leverages importance sampling to obtain an unbiased gradient estimator. We show that PAGE-PG enjoys a $\mathcal{O}\left(\epsilon^{-3} \right)$ average sample complexity to reach an ϵ-stationary solution, which matches the sample complexity of its most competitive counterparts under the same setting. A numerical evaluation confirms the competitive performance of our method on classical control tasks.

The power of first-order smooth optimization for black-box non-smooth problems

- Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takac, Pavel Dvurechensky, Bin Gu
- abstract: Gradient-free/zeroth-order methods for black-box convex optimization have been extensively studied in the last decade with the main focus on oracle calls complexity. In this paper, besides the oracle complexity, we focus also on iteration complexity, and propose a generic approach that, based on optimal first-order methods, allows to obtain in a black-box fashion new zeroth-order algorithms for non-smooth convex optimization problems. Our approach not only leads to optimal oracle complexity, but also allows to obtain iteration complexity similar to first-order methods, which, in turn, allows to exploit parallel computations to accelerate the convergence of our algorithms. We also elaborate on extensions for stochastic optimization problems, saddle-point problems, and distributed optimization.

A Functional Information Perspective on Model Interpretation

- Itai Gat, Nitay Calderon, Roi Reichart, Tamir Hazan
- abstract: Contemporary predictive models are hard to interpret as their deep nets exploit numerous complex relations between input elements. This work suggests a theoretical framework for model interpretability by measuring the contribution of relevant features to the functional entropy of the network with respect to the input. We rely on the log-Sobolev inequality that bounds the functional entropy by the functional Fisher information with respect to the covariance of the data. This provides a principled way to measure the amount of information contribution of a subset of features to the decision function. Through extensive experiments, we show that our method surpasses existing interpretability sampling-based methods on various data signals such as image, text, and audio.

UniRank: Unimodal Bandit Algorithms for Online Ranking

- Camille-Sovanneary Gauthier, Romaric Gaudel, Elisa Fromont
- abstract: We tackle, in the multiple-play bandit setting, the online ranking problem of assigning L items to K predefined positions on a web page in order to maximize the number of user clicks. We propose a generic algorithm, UniRank, that tackles state-of-the-art click models. The regret bound of this algorithm is a direct consequence of the pseudo-unimodality property of the bandit setting with respect to a graph where nodes are ordered sets of indistinguishable items. The main contribution of UniRank is its O(L/Δ logT) regret for T consecutive assignments, where Δ relates to the reward-gap between two items. This regret bound is based on the usually implicit condition that two items may not have the same attractiveness. Experiments against state-of-the-art learning algorithms specialized or not for different click models, show that our method has better regret performance than other generic algorithms on real life and synthetic datasets.

Variational Inference with Locally Enhanced Bounds for Hierarchical Models

- Tomas Geffner, Justin Domke
- abstract: Hierarchical models represent a challenging setting for inference algorithms. MCMC methods struggle to scale to large models with many local variables and observations, and variational inference (VI) may fail to provide accurate approximations due to the use of simple variational families. Some variational methods (e.g. importance weighted VI) integrate Monte Carlo methods to give better accuracy, but these tend to be unsuitable for hierarchical models, as they do not allow for subsampling and their performance tends to degrade for high dimensional models. We propose a new family of variational bounds for hierarchical models, based on the application of tightening methods (e.g. importance weighting) separately for each group of local random variables. We show that our approach naturally allows the use of subsampling to get unbiased gradients, and that it fully leverages the power of methods that build tighter lower bounds by applying them independently in lower dimensional spaces, leading to better results and more accurate posterior approximations than relevant baselines.

Inducing Causal Structure for Interpretable Neural Networks

- Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, Christopher Potts
- abstract: In many areas, we have well-founded insights about causal structure that would be useful to bring into our trained models while still allowing them to learn in a data-driven fashion. To achieve this, we present the new method of interchange intervention training (IIT). In IIT, we (1) align variables in a causal model (e.g., a deterministic program or Bayesian network) with representations in a neural model and (2) train the neural model to match the counterfactual behavior of the causal model on a base input when aligned representations in both models are set to be the value they would be for a source input. IIT is fully differentiable, flexibly combines with other objectives, and guarantees that the target causal model is a causal abstraction of the neural model when its loss is zero. We evaluate IIT on a structural vision task (MNIST-PVR), a navigational language task (ReaSCAN), and a natural language

inference task (MQNLI). We compare IIT against multi-task training objectives and data augmentation. In all our experiments, IIT achieves the best results and produces neural models that are more interpretable in the sense that they more successfully realize the target causal model.

Achieving Minimax Rates in Pool-Based Batch Active Learning

- Claudio Gentile, Zhilei Wang, Tong Zhang
- abstract: We consider a batch active learning scenario where the learner adaptively issues batches of points to a labeling oracle. Sampling labels in batches is highly desirable in practice due to the smaller number of interactive rounds with the labeling oracle (often human beings). However, batch active learning typically pays the price of a reduced adaptivity, leading to suboptimal results. In this paper we propose a solution which requires a careful trade off between the informativeness of the queried points and their diversity. We theoretically investigate batch active learning in the practically relevant scenario where the unlabeled pool of data is available beforehand (pool-based active learning). We analyze a novel stage-wise greedy algorithm and show that, as a function of the label complexity, the excess risk of this algorithm %operating in the realizable setting for which we prove matches the known minimax rates in standard statistical learning settings. Our results also exhibit a mild dependence on the batch size. These are the first theoretical results that employ careful trade offs between informativeness and diversity to rigorously quantify the statistical performance of batch active learning in the pool-based scenario.

Near-Exact Recovery for Tomographic Inverse Problems via Deep Learning

- Martin Genzel, Ingo Gühring, Jan Macdonald, Maximilian März
- abstract: This work is concerned with the following fundamental question in scientific machine learning: Can deep-learning-based methods solve noise-free inverse problems to near-perfect accuracy? Positive evidence is provided for the first time, focusing on a prototypical computed tomography (CT) setup. We demonstrate that an iterative end-to-end network scheme enables reconstructions close to numerical precision, comparable to classical compressed sensing strategies. Our results build on our winning submission to the recent AAPM DL-Sparse-View CT Challenge. Its goal was to identify the state-of-the-art in solving the sparse-view CT inverse problem with data-driven techniques. A specific difficulty of the challenge setup was that the precise forward model remained unknown to the participants. Therefore, a key feature of our approach was to initially estimate the unknown fanbeam geometry in a data-driven calibration step. Apart from an in-depth analysis of our methodology, we also demonstrate its state-of-the-art performance on the open-access real-world dataset LoDoPaB CT.

Online Learning for Min Sum Set Cover and Pandora's Box

- Evangelia Gergatsouli, Christos Tzamos
- abstract: Two central problems in Stochastic Optimization are Min-Sum Set Cover and Pandora's Box. In Pandora's Box, we are presented with n boxes, each containing an unknown value and the goal is to open the boxes in some order to minimize the sum of the search cost and the smallest value found. Given a distribution of value vectors, we are asked to identify a near-optimal search order. Min-Sum Set Cover corresponds to the case where values are either 0 or infinity. In this work, we study the case where the value vectors are not drawn from a distribution but are presented to a learner in an online fashion. We present a computationally efficient algorithm that is constant-competitive against the cost of the optimal search order. We extend our results to a bandit setting where only the values of the boxes opened are revealed to the learner after every round. We also generalize our results to other commonly studied variants of Pandora's Box and Min-Sum Set Cover that involve selecting more than a single value subject to a matroid constraint.

Equivariance versus Augmentation for Spherical Images

- Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, Daniel Persson
- abstract: We analyze the role of rotational equivariance in convolutional neural networks (CNNs) applied to spherical images. We compare the performance of the group equivariant networks known as S2CNNs and standard non-equivariant CNNs trained with an increasing amount of data augmentation. The chosen architectures can be considered baseline references for the respective design paradigms. Our models are trained and evaluated on single or multiple items from the MNIST- or FashionMNIST dataset projected onto the sphere. For the task of image classification, which is inherently rotationally invariant, we find that by considerably increasing the amount of data augmentation and the size of the networks, it is possible for the standard CNNs to reach at least the same performance as the equivariant network. In contrast, for the inherently equivariant task of semantic segmentation, the non-equivariant networks are consistently outperformed by the equivariant networks with significantly fewer parameters. We also analyze and compare the inference latency and training times of the different networks, enabling detailed tradeoff considerations between equivariant architectures and data augmentation for practical problems.

A Regret Minimization Approach to Multi-Agent Control

- Udaya Ghai, Udari Madhushani, Naomi Leonard, Elad Hazan
- abstract: We study the problem of multi-agent control of a dynamical system with known dynamics and adversarial disturbances. Our study focuses on optimal control without centralized precomputed policies, but rather with adaptive control policies for the different agents that are only equipped with a stabilizing controller. We give a reduction from any (standard) regret minimizing control method to a distributed algorithm. The reduction guarantees that the resulting distributed algorithm has low regret relative to the optimal precomputed joint policy. Our methodology involves generalizing online convex optimization to a multi-agent setting and applying recent tools from nonstochastic control derived for a single agent. We empirically evaluate our method on a model of an overactuated aircraft. We show that the distributed method is robust to failure and to adversarial perturbations in the dynamics.

Blocks Assemble! Learning to Assemble with Large-Scale Structured Reinforcement Learning

- Seyed Kamyar Seyed Ghasemipour, Satoshi Kataoka, Byron David, Daniel Freeman, Shixiang Shane Gu, Igor Mordatch
- abstract: Assembly of multi-part physical structures is both a valuable end product for autonomous robotics, as well as a valuable diagnostic task for open-ended training of embodied intelligent agents. We introduce a naturalistic physics-based environment with a set of connectable magnet blocks inspired by children's toy kits. The objective is to assemble blocks into a succession of target blueprints. Despite the simplicity of this objective, the compositional nature of building diverse blueprints from a set of blocks leads to an explosion of complexity in structures that agents encounter. Furthermore, assembly stresses agents' multi-step planning, physical reasoning, and bimanual coordination. We find that the combination of large-scale reinforcement learning and graph-based policies – surprisingly without any additional complexity – is an effective recipe for training agents that not only generalize to complex unseen blueprints in a zero-shot manner, but even operate in a reset-free setting without being trained to do so. Through extensive experiments, we highlight the importance of large-scale training, structured representations, contributions of multi-task vs. single-task learning, as well as the effects of curriculums, and discuss qualitative behaviors of trained agents. Our accompanying project webpage can be found at: https://sites.google.com/view/learning-direct-assembly/home

Faster Privacy Accounting via Evolving Discretization

- Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi
- abstract: We introduce a new algorithm for numerical composition of privacy random variables, useful for computing the accurate differential privacy parameters for compositions of mechanisms. Our algorithm achieves a running time and memory usage of $polylog(k)$ for the task of self-composing a

mechanism, from a broad class of mechanisms, k times; this class, e.g., includes the sub-sampled Gaussian mechanism, that appears in the analysis of differentially private stochastic gradient descent (DP-SGD). By comparison, recent work by Gopi et al. (NeurIPS 2021) has obtained a running time of $\widetilde{O}(\sqrt{k})$ for the same task. Our approach extends to the case of composing k different mechanisms in the same class, improving upon the running time and memory usage in their work from $\widetilde{O}(k^{1.5})$ to $\wtilde{O}(k)$.

Plug-In Inversion: Model-Agnostic Inversion for Vision with Data Augmentations

- Amin Ghiasi, Hamid Kazemi, Steven Reich, Chen Zhu, Micah Goldblum, Tom Goldstein
- abstract: Existing techniques for model inversion typically rely on hard-to-tune regularizers, such as total variation or feature regularization, which must be individually calibrated for each network in order to produce adequate images. In this work, we introduce Plug-In Inversion, which relies on a simple set of augmentations and does not require excessive hyper-parameter tuning. Under our proposed augmentation-based scheme, the same set of augmentation hyper-parameters can be used for inverting a wide range of image classification models, regardless of input dimensions or the architecture. We illustrate the practicality of our approach by inverting Vision Transformers (ViTs) and Multi-Layer Perceptrons (MLPs) trained on the ImageNet dataset, tasks which to the best of our knowledge have not been successfully accomplished by any previous works.

Offline RL Policies Should Be Trained to be Adaptive

- Dibya Ghosh, Anurag Ajay, Pulkit Agrawal, Sergey Levine
- abstract: Offline RL algorithms must account for the fact that the dataset they are provided may leave many facets of the environment unknown. The most common way to approach this challenge is to employ pessimistic or conservative methods, which avoid behaviors that are too dissimilar from those in the training dataset. However, relying exclusively on conservatism has drawbacks: performance is sensitive to the exact degree of conservatism, and conservative objectives can recover highly suboptimal policies. In this work, we propose that offline RL methods should instead be adaptive in the presence of uncertainty. We show that acting optimally in offline RL in a Bayesian sense involves solving an implicit POMDP. As a result, optimal policies for offline RL must be adaptive, depending not just on the current state but rather all the transitions seen so far during evaluation. We present a model-free algorithm for approximating this optimal adaptive policy, and demonstrate the efficacy of learning such adaptive policies in offline RL benchmarks.

Breaking the $\sqrtT$ Barrier: Instance-Independent Logarithmic Regret in Stochastic Contextual Linear Bandits

- Avishek Ghosh, Abishek Sankararaman
- abstract: We prove an instance independent (poly) logarithmic regret for stochastic contextual bandits with linear payoff. Previously, in \cite{chu2011contextual}, a lower bound of $\mathcal{O}(\sqrt{T})$ is shown for the contextual linear bandit problem with arbitrary (adversarily chosen) contexts. In this paper, we show that stochastic contexts indeed help to reduce the regret from $\sqrt{T}$ to $\polylog(T)$. We propose Low Regret Stochastic Contextual Bandits (\texttt{LR-SCB}), which takes advantage of the stochastic contexts and performs parameter estimation (in ℓ_2 norm) and regret minimization simultaneously. \texttt{LR-SCB} works in epochs, where the parameter estimation of the previous epoch is used to reduce the regret of the current epoch. The (poly) logarithmic regret of \texttt{LR-SCB} stems from two crucial facts: (a) the application of a norm adaptive algorithm to exploit the parameter estimation and (b) an analysis of the shifted linear contextual bandit algorithm, showing that shifting results in increasing regret. We have also shown experimentally that stochastic contexts indeed incurs a regret that scales with $\polylog(T)$.

SCHA-VAE: Hierarchical Context Aggregation for Few-Shot Generation

- Giorgio Giannone, Ole Winther
- abstract: A few-shot generative model should be able to generate data from a novel distribution by only observing a limited set of examples. In few-shot learning the model is trained on data from many sets from distributions sharing some underlying properties such as sets of characters from different alphabets or objects from different categories. We extend current latent variable models for sets to a fully hierarchical approach with an attention-based point to set-level aggregation and call our method SCHA-VAE for Set-Context-Hierarchical-Aggregation Variational Autoencoder. We explore likelihood-based model comparison, iterative data sampling, and adaptation-free out-of-distribution generalization. Our results show that the hierarchical formulation better captures the intrinsic variability within the sets in the small data regime. This work generalizes deep latent variable approaches to few-shot learning, taking a step toward large-scale few-shot generation with a formulation that readily works with current state-of-the-art deep generative models.

A Joint Exponential Mechanism For Differentially Private Top-k

- Jennifer Gillenwater, Matthew Joseph, Andres Munoz, Monica Ribero Diaz
- abstract: We present a differentially private algorithm for releasing the sequence of k elements with the highest counts from a data domain of d elements. The algorithm is a "joint" instance of the exponential mechanism, and its output space consists of all $O(d^k)$ length-k sequences. Our main contribution is a method to sample this exponential mechanism in time $O(dk\log(k) + d\log(d))$ and space $O(dk)$. Experiments show that this approach outperforms existing pure differential privacy methods and improves upon even approximate differential privacy methods for moderate k.

Neuro-Symbolic Hierarchical Rule Induction

- Claire Glanois, Zhaohui Jiang, Xuening Feng, Paul Weng, Matthieu Zimmer, Dong Li, Wulong Liu, Jianye Hao
- abstract: We propose Neuro-Symbolic Hierarchical Rule Induction, an efficient interpretable neuro-symbolic model, to solve Inductive Logic Programming (ILP) problems. In this model, which is built from a pre-defined set of meta-rules organized in a hierarchical structure, first-order rules are invented by learning embeddings to match facts and body predicates of a meta-rule. To instantiate, we specifically design an expressive set of generic meta-rules, and demonstrate they generate a consequent fragment of Horn clauses. As a differentiable model, HRI can be trained both via supervised learning and reinforcement learning. To converge to interpretable rules, we inject a controlled noise to avoid local optima and employ an interpretability-regularization term. We empirically validate our model on various tasks (ILP, visual genome, reinforcement learning) against relevant state-of-the-art methods, including traditional ILP methods and neuro-symbolic models.

It’s Raw! Audio Generation with State-Space Models

- Karan Goel, Albert Gu, Chris Donahue, Christopher Re
- abstract: Developing architectures suitable for modeling raw audio is a challenging problem due to the high sampling rates of audio waveforms. Standard sequence modeling approaches like RNNs and CNNs have previously been tailored to fit the demands of audio, but the resultant architectures make undesirable computational tradeoffs and struggle to model waveforms effectively. We propose SaShiMi, a new multi-scale architecture for waveform modeling built around the recently introduced S4 model for long sequence modeling. We identify that S4 can be unstable during autoregressive generation, and provide a simple improvement to its parameterization by drawing connections to Hurwitz matrices. SaShiMi yields state-of-the-art performance for unconditional waveform generation in the autoregressive setting. Additionally, SaShiMi improves non-autoregressive generation performance when used as the backbone architecture for a diffusion model. Compared to prior architectures in the autoregressive generation setting, SaShiMi generates piano and speech waveforms which humans find more musical and coherent respectively, e.g. 2{\texttimes} better mean opinion scores

than WaveNet on an unconditional speech generation task. On a music generation task, SaShiMi outperforms WaveNet on density estimation and speed at both training and inference even when using 3{\texttimes} fewer parameters

RankSim: Ranking Similarity Regularization for Deep Imbalanced Regression

- Yu Gong, Greg Mori, Fred Tung
- abstract: Data imbalance, in which a plurality of the data samples come from a small proportion of labels, poses a challenge in training deep neural networks. Unlike classification, in regression the labels are continuous, potentially boundless, and form a natural ordering. These distinct features of regression call for new techniques that leverage the additional information encoded in label-space relationships. This paper presents the RankSim (ranking similarity) regularizer for deep imbalanced regression, which encodes an inductive bias that samples that are closer in label space should also be closer in feature space. In contrast to recent distribution smoothing based approaches, RankSim captures both nearby and distant relationships: for a given data sample, RankSim encourages the sorted list of its neighbors in label space to match the sorted list of its neighbors in feature space. RankSim is complementary to conventional imbalanced learning techniques, including re-weighting, two-stage training, and distribution smoothing, and lifts the state-of-the-art performance on three imbalanced regression benchmarks: IMDB-WIKI-DIR, AgeDB-DIR, and STS-B-DIR.

How to Fill the Optimum Set? Population Gradient Descent with Harmless Diversity

- Chengyue Gong, Lemeng Wu, Qiang Liu
- abstract: Although traditional optimization methods focus on finding a single optimal solution, most objective functions in modern machine learning problems, especially those in deep learning, often have multiple or infinite number of optimal points. Therefore, it is useful to consider the problem of finding a set of diverse points in the optimum set of an objective function. In this work, we frame this problem as a bi-level optimization problem of maximizing a diversity score inside the optimum set of the main loss function, and solve it with a simple population gradient descent framework that iteratively updates the points to maximize the diversity score in a fashion that does not hurt the optimization of the main loss. We demonstrate that our method can efficiently generate diverse solutions on multiple applications, e.g. text-to-image generation, text-to-mesh generation, molecular conformation generation and ensemble neural network training.

Partial Label Learning via Label Influence Function

- Xiuwen Gong, Dong Yuan, Wei Bao
- abstract: To deal with ambiguities in partial label learning (PLL), state-of-the-art strategies implement disambiguations by identifying the ground-truth label directly from the candidate label set. However, these approaches usually take the label that incurs a minimal loss as the ground-truth label or use the weight to represent which label has a high likelihood to be the ground-truth label. Little work has been done to investigate from the perspective of how a candidate label changing a predictive model. In this paper, inspired by influence function, we develop a novel PLL framework called Partial Label Learning via Label Influence Function (PLL-IF). Moreover, we implement the framework with two specific representative models, an SVM model and a neural network model, which are called PLL-IF+SVM and PLL-IF+NN method respectively. Extensive experiments conducted on various datasets demonstrate the superiorities of the proposed methods in terms of prediction accuracy, which in turn validates the effectiveness of the proposed PLL-IF framework.

Secure Distributed Training at Scale

- Eduard Gorbunov, Alexander Borzunov, Michael Diskin, Max Ryabinin
- abstract: Many areas of deep learning benefit from using increasingly larger neural networks trained on public data, as is the case for pre-trained models for NLP and computer vision. Training such models requires a lot of computational resources (e.g., HPC clusters) that are not available to small research groups and independent researchers. One way to address it is for several smaller groups to pool their computational resources together and train a model that benefits all participants. Unfortunately, in this case, any participant can jeopardize the entire training run by sending incorrect updates, deliberately or by mistake. Training in presence of such peers requires specialized distributed training algorithms with Byzantine tolerance. These algorithms often sacrifice efficiency by introducing redundant communication or passing all updates through a trusted server, making it infeasible to apply them to large-scale deep learning, where models can have billions of parameters. In this work, we propose a novel protocol for secure (Byzantine-tolerant) decentralized training that emphasizes communication efficiency.

Retrieval-Augmented Reinforcement Learning

- Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adrià Puigdomènech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, Michal Valko, Simon Osindero, Timothy Lillicrap, Nicolas Heess, Charles Blundell
- abstract: Most deep reinforcement learning (RL) algorithms distill experience into parametric behavior policies or value functions via gradient updates. While effective, this approach has several disadvantages: (1) it is computationally expensive, (2) it can take many updates to integrate experiences into the parametric model, (3) experiences that are not fully integrated do not appropriately influence the agent's behavior, and (4) behavior is limited by the capacity of the model. In this paper we explore an alternative paradigm in which we train a network to map a dataset of past experiences to optimal behavior. Specifically, we augment an RL agent with a retrieval process (parameterized as a neural network) that has direct access to a dataset of experiences. This dataset can come from the agent's past experiences, expert demonstrations, or any other relevant source. The retrieval process is trained to retrieve information from the dataset that may be useful in the current context, to help the agent achieve its goal faster and more efficiently. The proposed method facilitates learning agents that at test time can condition their behavior on the entire dataset and not only the current state, or current trajectory. We integrate our method into two different RL agents: an offline DQN agent and an online R2D2 agent. In offline multi-task problems, we show that the retrieval-augmented DQN agent avoids task interference and learns faster than the baseline DQN agent. On Atari, we show that retrieval-augmented R2D2 learns significantly faster than the baseline R2D2 agent and achieves higher scores. We run extensive ablations to measure the contributions of the components of our proposed method.

The State of Sparse Training in Deep Reinforcement Learning

- Laura Graesser, Utku Evci, Erich Elsen, Pablo Samuel Castro
- abstract: The use of sparse neural networks has seen rapid growth in recent years, particularly in computer vision. Their appeal stems largely from the reduced number of parameters required to train and store, as well as in an increase in learning efficiency. Somewhat surprisingly, there have been very few efforts exploring their use in Deep Reinforcement Learning (DRL). In this work we perform a systematic investigation into applying a number of existing sparse training techniques on a variety of DRL agents and environments. Our results corroborate the findings from sparse training in the computer vision domain {–}sparse networks perform better than dense networks for the same parameter count{–} in the DRL domain. We provide detailed analyses on how the various components in DRL are affected by the use of sparse networks and conclude by suggesting promising avenues for improving the effectiveness of sparse training methods, as well as for advancing their use in DRL.

Causal Inference Through the Structural Causal Marginal Problem

- Luigi Gresele, Julius Von Kügelgen, Jonas Kübler, Elke Kirschbaum, Bernhard Schölkopf, Dominik Janzing

- abstract: We introduce an approach to counterfactual inference based on merging information from multiple datasets. We consider a causal reformulation of the statistical marginal problem: given a collection of marginal structural causal models (SCMs) over distinct but overlapping sets of variables, determine the set of joint SCMs that are counterfactually consistent with the marginal ones. We formalise this approach for categorical SCMs using the response function formulation and show that it reduces the space of allowed marginal and joint SCMs. Our work thus highlights a new mode of falsifiability through additional variables, in contrast to the statistical one via additional data.

Mirror Learning: A Unifying Framework of Policy Optimisation

- Jakub Grudzien, Christian A Schroeder De Witt, Jakob Foerster
- abstract: Modern deep reinforcement learning (RL) algorithms are motivated by either the general policy improvement (GPI) or trust-region learning (TRL) frameworks. However, algorithms that strictly respect these theoretical frameworks have proven unscalable. Surprisingly, the only known scalable algorithms violate the GPI/TRL assumptions, e.g. due to required regularisation or other heuristics. The current explanation of their empirical success is essentially “by analogy”: they are deemed approximate adaptations of theoretically sound methods. Unfortunately, studies have shown that in practice these algorithms differ greatly from their conceptual ancestors. In contrast, in this paper, we introduce a novel theoretical framework, named Mirror Learning, which provides theoretical guarantees to a large class of algorithms, including TRPO and PPO. While the latter two exploit the flexibility of our framework, GPI and TRL fit in merely as pathologically restrictive corner cases thereof. This suggests that the empirical performance of state-of-the-art methods is a direct consequence of their theoretical properties, rather than of aforementioned approximate analogies. Mirror learning sets us free to boldly explore novel, theoretically sound RL algorithms, a thus far uncharted wonderland.

Adapting k-means Algorithms for Outliers

- Christoph Grunau, Václav Rozhoň
- abstract: This paper shows how to adapt several simple and classical sampling-based algorithms for the k-means problem to the setting with outliers. Recently, Bhaskara et al. (NeurIPS 2019) showed how to adapt the classical k-means++ algorithm to the setting with outliers. However, their algorithm needs to output O(log(k)$\cdot$z) outliers, where z is the number of true outliers, to match the O(log k)-approximation guarantee of k-means++. In this paper, we build on their ideas and show how to adapt several sequential and distributed k-means algorithms to the setting with outliers, but with substantially stronger theoretical guarantees: our algorithms output (1 + ϵ)z outliers while achieving an O(1/ϵ)-approximation to the objective function. In the sequential world, we achieve this by adapting a recent algorithm of Lattanzi and Sohler (ICML 2019). In the distributed setting, we adapt a simple algorithm of Guha et al. (IEEE Trans. Know. and Data Engineering 2003) and the popular k-means\|{of} Bahmani et al. (PVLDB2012). A theoretical application of our techniques is an algorithm with running time O(nk^2/z) that achieves an O(1)-approximation to the objective function while outputting O(z) outliers, assuming k << z << n. This is complemented with a matching lower bound of Ω(nk^2/z) for this problem in the oracle model.

Variational Mixtures of ODEs for Inferring Cellular Gene Expression Dynamics

- Yichen Gu, David T Blaauw, Joshua Welch
- abstract: A key problem in computational biology is discovering the gene expression changes that regulate cell fate transitions, in which one cell type turns into another. However, each individual cell cannot be tracked longitudinally, and cells at the same point in real time may be at different stages of the transition process. This can be viewed as a problem of learning the behavior of a dynamical system from observations whose times are unknown. Additionally, a single progenitor cell type often bifurcates into multiple child cell types, further complicating the problem of modeling the dynamics. To address this problem, we developed an approach called variational mixtures of ordinary differential equations. By using a simple family of ODEs informed by the biochemistry of gene expression to constrain the likelihood of a deep generative model, we can simultaneously infer the latent time and latent state of each cell and predict its future gene expression state. The model can be interpreted as a mixture of ODEs whose parameters vary continuously across a latent space of cell states. Our approach dramatically improves data fit, latent time inference, and future cell state estimation of single-cell gene expression data compared to previous approaches.

Learning Pseudometric-based Action Representations for Offline Reinforcement Learning

- Pengjie Gu, Mengchen Zhao, Chen Chen, Dong Li, Jianye Hao, Bo An
- abstract: Offline reinforcement learning is a promising approach for practical applications since it does not require interactions with real-world environments. However, existing offline RL methods only work well in environments with continuous or small discrete action spaces. In environments with large and discrete action spaces, such as recommender systems and dialogue systems, the performance of existing methods decreases drastically because they suffer from inaccurate value estimation for a large proportion of out-of-distribution (o.o.d.) actions. While recent works have demonstrated that online RL benefits from incorporating semantic information in action representations, unfortunately, they fail to learn reasonable relative distances between action representations, which is key to offline RL to reduce the influence of o.o.d. actions. This paper proposes an action representation learning framework for offline RL based on a pseudometric, which measures both the behavioral relation and the data-distributional relation between actions. We provide theoretical analysis on the continuity of the expected Q-values and the offline policy improvement using the learned action representations. Experimental results show that our methods significantly improve the performance of two typical offline RL methods in environments with large and discrete action spaces.

NeuroFluid: Fluid Dynamics Grounding with Particle-Driven Neural Radiance Fields

- Shanyan Guan, Huayu Deng, Yunbo Wang, Xiaokang Yang
- abstract: Deep learning has shown great potential for modeling the physical dynamics of complex particle systems such as fluids. Existing approaches, however, require the supervision of consecutive particle properties, including positions and velocities. In this paper, we consider a partially observable scenario known as fluid dynamics grounding, that is, inferring the state transitions and interactions within the fluid particle systems from sequential visual observations of the fluid surface. We propose a differentiable two-stage network named NeuroFluid. Our approach consists of (i) a particle-driven neural renderer, which involves fluid physical properties into the volume rendering function, and (ii) a particle transition model optimized to reduce the differences between the rendered and the observed images. NeuroFluid provides the first solution to unsupervised learning of particle-based fluid dynamics by training these two models jointly. It is shown to reasonably estimate the underlying physics of fluids with different initial shapes, viscosity, and densities.

Fast-Rate PAC-Bayesian Generalization Bounds for Meta-Learning

- Jiechao Guan, Zhiwu Lu
- abstract: PAC-Bayesian error bounds provide a theoretical guarantee on the generalization abilities of meta-learning from training tasks to unseen tasks. However, it is still unclear how tight PAC-Bayesian bounds we can achieve for meta-learning. In this work, we propose a general PAC-Bayesian framework to cope with single-task learning and meta-learning uniformly. With this framework, we generalize the two tightest PAC-Bayesian bounds (i.e., kl-bound and Catoni-bound) from single-task learning to standard meta-learning, resulting in fast convergence rates for PAC-Bayesian meta-learners. By minimizing the derived two bounds, we develop two meta-learning algorithms for classification problems with deep neural networks. For regression problems, by setting Gibbs optimal posterior for each training task, we obtain the closed-form formula of the minimizer of our Catoni-bound, leading to an efficient Gibbs meta-learning algorithm. Although minimizing our kl-bound can not yield a closed-form solution, we show that it can be extended for

analyzing the more challenging meta-learning setting where samples from different training tasks exhibit interdependencies. Experiments empirically show that our proposed meta-learning algorithms achieve competitive results with respect to latest works.

Leveraging Approximate Symbolic Models for Reinforcement Learning via Skill Diversity

- Lin Guan, Sarath Sreedharan, Subbarao Kambhampati
- abstract: Creating reinforcement learning (RL) agents that are capable of accepting and leveraging task-specific knowledge from humans has been long identified as a possible strategy for developing scalable approaches for solving long-horizon problems. While previous works have looked at the possibility of using symbolic models along with RL approaches, they tend to assume that the high-level action models are executable at low level and the fluents can exclusively characterize all desirable MDP states. Symbolic models of real world tasks are however often incomplete. To this end, we introduce Approximate Symbolic-Model Guided Reinforcement Learning, wherein we will formalize the relationship between the symbolic model and the underlying MDP that will allow us to characterize the incompleteness of the symbolic model. We will use these models to extract high-level landmarks that will be used to decompose the task. At the low level, we learn a set of diverse policies for each possible task subgoal identified by the landmark, which are then stitched together. We evaluate our system by testing on three different benchmark domains and show how even with incomplete symbolic model information, our approach is able to discover the task structure and efficiently guide the RL agent towards the goal.

Large-Scale Graph Neural Architecture Search

- Chaoyu Guan, Xin Wang, Hong Chen, Ziwei Zhang, Wenwu Zhu
- abstract: Graph Neural Architecture Search (GNAS) has become a powerful method in automatically discovering suitable Graph Neural Network (GNN) architectures for different tasks. However, existing approaches fail to handle large-scale graphs because current performance estimation strategies in GNAS are computationally expensive for large-scale graphs and suffer from consistency collapse issues. To tackle these problems, we propose the Graph ArchitectUre Search at Scale (GAUSS) method that can handle large-scale graphs by designing an efficient light-weight supernet and the joint architecture-graph sampling. In particular, a graph sampling-based single-path one-shot supernet is proposed to reduce the computation burden. To address the consistency collapse issues, we further explicitly consider the joint architecture-graph sampling through a novel architecture peer learning mechanism on the sampled sub-graphs and an architecture importance sampling algorithm. Our proposed framework is able to smooth the highly non-convex optimization objective and stabilize the architecture sampling process. We provide theoretical analyses on GAUSS and empirically evaluate it on five datasets whose vertex sizes range from 10^4 to 10^8. The experimental results demonstrate substantial improvements of GAUSS over other GNAS baselines on all datasets. To the best of our knowledge, the proposed GAUSS method is the first graph neural architecture search framework that can handle graphs with billions of edges within 1 GPU day.

Identifiability Conditions for Domain Adaptation

- Ishaan Gulrajani, Tatsunori Hashimoto
- abstract: Domain adaptation algorithms and theory have relied upon an assumption that the observed data uniquely specify the correct correspondence between the domains. Unfortunately, it is unclear under what conditions this identifiability assumption holds, even when restricting ourselves to the case where a correct bijective map between domains exists. We study this bijective domain mapping problem and provide several new sufficient conditions for the identifiability of linear domain maps. As a consequence of our analysis, we show that weak constraints on the third moment tensor suffice for identifiability, prove identifiability for common latent variable models such as topic models, and give a computationally tractable method for generating certificates for the identifiability of linear maps. Inspired by our certification method, we derive a new objective function for domain mapping that explicitly accounts for uncertainty over maps arising from unidentifiability. We demonstrate that our objective leads to improvements in uncertainty quantification and model performance estimation.

A Parametric Class of Approximate Gradient Updates for Policy Optimization

- Ramki Gummadi, Saurabh Kumar, Junfeng Wen, Dale Schuurmans
- abstract: Approaches to policy optimization have been motivated from diverse principles, based on how the parametric model is interpreted (e.g. value versus policy representation) or how the learning objective is formulated, yet they share a common goal of maximizing expected return. To better capture the commonalities and identify key differences between policy optimization methods, we develop a unified perspective that re-expresses the underlying updates in terms of a limited choice of gradient form and scaling function. In particular, we identify a parameterized space of approximate gradient updates for policy optimization that is highly structured, yet covers both classical and recent examples, including PPO. As a result, we obtain novel yet well motivated updates that generalize existing algorithms in a way that can deliver benefits both in terms of convergence speed and final result quality. An experimental investigation demonstrates that the additional degrees of freedom provided in the parameterized family of updates can be leveraged to obtain non-trivial improvements both in synthetic domains and on popular deep RL benchmarks.

Provably Efficient Offline Reinforcement Learning for Partially Observable Markov Decision Processes

- Hongyi Guo, Qi Cai, Yufeng Zhang, Zhuoran Yang, Zhaoran Wang
- abstract: We study offline reinforcement learning (RL) for partially observable Markov decision processes (POMDPs) with possibly infinite state and observation spaces. Under the undercompleteness assumption, the optimal policy in such POMDPs are characterized by a class of finite-memory Bellman operators. In the offline setting, estimating these operators directly is challenging due to (i) the large observation space and (ii) insufficient coverage of the offline dataset. To tackle these challenges, we propose a novel algorithm that constructs confidence regions for these Bellman operators via offline estimation of their RKHS embeddings, and returns the final policy via pessimistic planning within the confidence regions. We prove that the proposed algorithm attains an (\epsilon)-optimal policy using an offline dataset containing (\tilde\cO(1 / \epsilon^2)){episodes}, provided that the behavior policy has good coverage over the optimal trajectory. To our best knowledge, our algorithm is the first provably sample efficient offline algorithm for POMDPs without uniform coverage assumptions.

No-Regret Learning in Partially-Informed Auctions

- Wenshuo Guo, Michael Jordan, Ellen Vitercik
- abstract: Auctions with partially-revealed information about items are broadly employed in real-world applications, but the underlying mechanisms have limited theoretical support. In this work, we study a machine learning formulation of these types of mechanisms, presenting algorithms that are no-regret from the buyer’s perspective. Specifically, a buyer who wishes to maximize his utility interacts repeatedly with a platform over a series of T rounds. In each round, a new item is drawn from an unknown distribution and the platform publishes a price together with incomplete, “masked” information about the item. The buyer then decides whether to purchase the item. We formalize this problem as an online learning task where the goal is to have low regret with respect to a myopic oracle that has perfect knowledge of the distribution over items and the seller’s masking function. When the distribution over items is known to the buyer and the mask is a SimHash function mapping $\R^d$ to ${0,1}^{\ell}$, our algorithm has regret $\tilde \cO((Td\ell)^{\nicefrac{1}{2}})$. In a fully agnostic setting when the mask is an arbitrary function mapping to a set of size n and the prices are stochastic, our algorithm has regret $\tilde \cO((Tn)^{\nicefrac{1}{2}})$.

Bounding Training Data Reconstruction in Private (Deep) Learning

- Chuan Guo, Brian Karrer, Kamalika Chaudhuri, Laurens van der Maaten
- abstract: Differential privacy is widely accepted as the de facto method for preventing data leakage in ML, and conventional wisdom suggests that it offers strong protection against privacy attacks. However, existing semantic guarantees for DP focus on membership inference, which may overestimate the adversary's capabilities and is not applicable when membership status itself is non-sensitive. In this paper, we derive the first semantic guarantees for DP mechanisms against training data reconstruction attacks under a formal threat model. We show that two distinct privacy accounting methods—Renyi differential privacy and Fisher information leakage—both offer strong semantic protection against data reconstruction attacks.

Adversarially trained neural representations are already as robust as biological neural representations

- Chong Guo, Michael Lee, Guillaume Leclerc, Joel Dapello, Yug Rao, Aleksander Madry, James Dicarlo
- abstract: Visual systems of primates are the gold standard of robust perception. There is thus a general belief that mimicking the neural representations that underlie those systems will yield artificial visual systems that are adversarially robust. In this work, we develop a method for performing adversarial visual attacks directly on primate brain activity. We then leverage this method to demonstrate that the above-mentioned belief might not be well-founded. Specifically, we report that the biological neurons that make up visual systems of primates exhibit susceptibility to adversarial perturbations that is comparable in magnitude to existing (robustly trained) artificial neural networks.

Class-Imbalanced Semi-Supervised Learning with Adaptive Thresholding

- Lan-Zhe Guo, Yu-Feng Li
- abstract: Semi-supervised learning (SSL) has proven to be successful in overcoming labeling difficulties by leveraging unlabeled data. Previous SSL algorithms typically assume a balanced class distribution. However, real-world datasets are usually class-imbalanced, causing the performance of existing SSL algorithms to be seriously decreased. One essential reason is that pseudo-labels for unlabeled data are selected based on a fixed confidence threshold, resulting in low performance on minority classes. In this paper, we develop a simple yet effective framework, which only involves adaptive thresholding for different classes in SSL algorithms, and achieves remarkable performance improvement on more than twenty imbalance ratios. Specifically, we explicitly optimize the number of pseudo-labels for each class in the SSL objective, so as to simultaneously obtain adaptive thresholds and minimize empirical risk. Moreover, the determination of the adaptive threshold can be efficiently obtained by a closed-form solution. Extensive experimental results demonstrate the effectiveness of our proposed algorithms.

Deep Squared Euclidean Approximation to the Levenshtein Distance for DNA Storage

- Alan J.X. Guo, Cong Liang, Qing-Hu Hou
- abstract: Storing information in DNA molecules is of great interest because of its advantages in longevity, high storage density, and low maintenance cost. A key step in the DNA storage pipeline is to efficiently cluster the retrieved DNA sequences according to their similarities. Levenshtein distance is the most suitable metric on the similarity between two DNA sequences, but it is inferior in terms of computational complexity and less compatible with mature clustering algorithms. In this work, we propose a novel deep squared Euclidean embedding for DNA sequences using Siamese neural network, squared Euclidean embedding, and chi-squared regression. The Levenshtein distance is approximated by the squared Euclidean distance between the embedding vectors, which is fast calculated and clustering algorithm friendly. The proposed approach is analyzed theoretically and experimentally. The results show that the proposed embedding is efficient and robust.

Online Continual Learning through Mutual Information Maximization

- Yiduo Guo, Bing Liu, Dongyan Zhao
- abstract: This paper proposed a new online continual learning approach called OCM based on mutual information (MI) maximization. It achieves two objectives that are critical in dealing with catastrophic forgetting (CF). (1) It reduces feature bias caused by cross entropy (CE) as CE learns only discriminative features for each task, but these features may not be discriminative for another task. To learn a new task well, the network parameters learned before have to be modified, which causes CF. The new approach encourages the learning of each task to make use of the full features of the task training data. (2) It encourages preservation of the previously learned knowledge when training a new batch of incrementally arriving data. Empirical evaluation shows that OCM substantially outperforms the latest online CL baselines. For example, for CIFAR10, OCM improves the accuracy of the best baseline by 13.1% from 64.1% (baseline) to 77.2% (OCM).The code is publicly available at https://github.com/gydpku/OCM.

Fast Provably Robust Decision Trees and Boosting

- Jun-Qi Guo, Ming-Zhuo Teng, Wei Gao, Zhi-Hua Zhou
- abstract: Learning with adversarial robustness has been a challenge in contemporary machine learning, and recent years have witnessed increasing attention on robust decision trees and ensembles, mostly working with high computational complexity or without guarantees of provable robustness. This work proposes the Fast Provably Robust Decision Tree (FPRDT) with the smallest computational complexity O(n log n), a tradeoff between global and local optimizations over the adversarial 0/1 loss. We further develop the Provably Robust AdaBoost (PRAdaBoost) according to our robust decision trees, and present convergence analysis for training adversarial 0/1 loss. We conduct extensive experiments to support our approaches; in particular, our approaches are superior to those unprovably robust methods, and achieve better or comparable performance to those provably robust methods yet with the smallest running time.

Understanding and Improving Knowledge Graph Embedding for Entity Alignment

- Lingbing Guo, Qiang Zhang, Zequn Sun, Mingyang Chen, Wei Hu, Huajun Chen
- abstract: Embedding-based entity alignment (EEA) has recently received great attention. Despite significant performance improvement, few efforts have been paid to facilitate understanding of EEA methods. Most existing studies rest on the assumption that a small number of pre-aligned entities can serve as anchors connecting the embedding spaces of two KGs. Nevertheless, no one has investigated the rationality of such an assumption. To fill the research gap, we define a typical paradigm abstracted from existing EEA methods and analyze how the embedding discrepancy between two potentially aligned entities is implicitly bounded by a predefined margin in the score function. Further, we find that such a bound cannot guarantee to be tight enough for alignment learning. We mitigate this problem by proposing a new approach, named NeoEA, to explicitly learn KG-invariant and principled entity embeddings. In this sense, an EEA model not only pursues the closeness of aligned entities based on geometric distance, but also aligns the neural ontologies of two KGs by eliminating the discrepancy in embedding distribution and underlying ontology knowledge. Our experiments demonstrate consistent and significant performance improvement against the best-performing EEA methods.

NISPA: Neuro-Inspired Stability-Plasticity Adaptation for Continual Learning in Sparse Networks

- Mustafa B Gurbuz, Constantine Dovrolis
- abstract: The goal of continual learning (CL) is to learn different tasks over time. The main desiderata associated with CL are to maintain performance on older tasks, leverage the latter to improve learning of future tasks, and to introduce minimal overhead in the training process (for instance, to not require a growing model or retraining). We propose the Neuro-Inspired Stability-Plasticity Adaptation (NISPA) architecture that addresses these desiderata through a sparse neural network with fixed density. NISPA forms stable paths to preserve learned knowledge from older tasks. Also, NISPA uses connection

rewiring to create new plastic paths that reuse existing knowledge on novel tasks. Our extensive evaluation on EMNIST, FashionMNIST, CIFAR10, and CIFAR100 datasets shows that NISPA significantly outperforms representative state-of-the-art continual learning baselines, and it uses up to ten times fewer learnable parameters compared to baselines. We also make the case that sparsity is an essential ingredient for continual learning. The NISPA code is available at https://github.com/BurakGurbuz97/NISPA.

Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets

- Guy Hacohen, Avihu Dekel, Daphna Weinshall
- abstract: Investigating active learning, we focus on the relation between the number of labeled examples (budget size), and suitable querying strategies. Our theoretical analysis shows a behavior reminiscent of phase transition: typical examples are best queried when the budget is low, while unrepresentative examples are best queried when the budget is large. Combined evidence shows that a similar phenomenon occurs in common classification models. Accordingly, we propose TypiClust – a deep active learning strategy suited for low budgets. In a comparative empirical investigation of supervised learning, using a variety of architectures and image datasets, TypiClust outperforms all other active learning strategies in the low-budget regime. Using TypiClust in the semi-supervised framework, performance gets an even more significant boost. In particular, state-of-the-art semi-supervised methods trained on CIFAR-10 with 10 labeled examples selected by TypiClust, reach 93.2% accuracy – an improvement of 39.4% over random selection. Code is available at https://github.com/avihu111/TypiClust.

You Only Cut Once: Boosting Data Augmentation with a Single Cut

- Junlin Han, Pengfei Fang, Weihao Li, Jie Hong, Mohammad Ali Armin, Ian Reid, Lars Petersson, Hongdong Li
- abstract: We present You Only Cut Once (YOCO) for performing data augmentations. YOCO cuts one image into two pieces and performs data augmentations individually within each piece. Applying YOCO improves the diversity of the augmentation per sample and encourages neural networks to recognize objects from partial information. YOCO enjoys the properties of parameter-free, easy usage, and boosting almost all augmentations for free. Thorough experiments are conducted to evaluate its effectiveness. We first demonstrate that YOCO can be seamlessly applied to varying data augmentations, neural network architectures, and brings performance gains on CIFAR and ImageNet classification tasks, sometimes surpassing conventional image-level augmentation by large margins. Moreover, we show YOCO benefits contrastive pre-training toward a more powerful representation that can be better transferred to multiple downstream tasks. Finally, we study a number of variants of YOCO and empirically analyze the performance for respective settings.

Scalable MCMC Sampling for Nonsymmetric Determinantal Point Processes

- Insu Han, Mike Gartrell, Elvis Dohmatob, Amin Karbasi
- abstract: A determinantal point process (DPP) is an elegant model that assigns a probability to every subset of a collection of n items. While conventionally a DPP is parameterized by a symmetric kernel matrix, removing this symmetry constraint, resulting in nonsymmetric DPPs (NDPPs), leads to significant improvements in modeling power and predictive performance. Recent work has studied an approximate Markov chain Monte Carlo (MCMC) sampling algorithm for NDPPs restricted to size-k subsets (called k-NDPPs). However, the runtime of this approach is quadratic in n, making it infeasible for large-scale settings. In this work, we develop a scalable MCMC sampling algorithm for k-NDPPs with low-rank kernels, thus enabling runtime that is sublinear in n. Our method is based on a state-of-the-art NDPP rejection sampling algorithm, which we enhance with a novel approach for efficiently constructing the proposal distribution. Furthermore, we extend our scalable k-NDPP sampling algorithm to NDPPs without size constraints. Our resulting sampling method has polynomial time complexity in the rank of the kernel, while the existing approach has runtime that is exponential in the rank. With both a theoretical analysis and experiments on real-world datasets, we verify that our scalable approximate sampling algorithms are orders of magnitude faster than existing sampling approaches for k-NDPPs and NDPPs.

G-Mixup: Graph Data Augmentation for Graph Classification

- Xiaotian Han, Zhimeng Jiang, Ninghao Liu, Xia Hu
- abstract: This work develops mixup for graph data. Mixup has shown superiority in improving the generalization and robustness of neural networks by interpolating features and labels between two random samples. Traditionally, Mixup can work on regular, grid-like, and Euclidean data such as image or tabular data. However, it is challenging to directly adopt Mixup to augment graph data because different graphs typically: 1) have different numbers of nodes; 2) are not readily aligned; and 3) have unique typologies in non-Euclidean space. To this end, we propose G-Mixup to augment graphs for graph classification by interpolating the generator (i.e., graphon) of different classes of graphs. Specifically, we first use graphs within the same class to estimate a graphon. Then, instead of directly manipulating graphs, we interpolate graphons of different classes in the Euclidean space to get mixed graphons, where the synthetic graphs are generated through sampling based on the mixed graphons. Extensive experiments show that G-Mixup substantially improves the generalization and robustness of GNNs.

Private Streaming SCO in ℓ_p geometry with Applications in High Dimensional Online Decision Making

- Yuxuan Han, Zhicong Liang, Zhipeng Liang, Yang Wang, Yuan Yao, Jiheng Zhang
- abstract: Differentially private (DP) stochastic convex optimization (SCO) is ubiquitous in trustworthy machine learning algorithm design. This paper studies the DP-SCO problem with streaming data sampled from a distribution and arrives sequentially. We also consider the continual release model where parameters related to private information are updated and released upon each new data. Numerous algorithms have been developed to achieve optimal excess risks in different ℓ_p norm geometries, but none of the existing ones can be adapted to the streaming and continual release setting. We propose a private variant of the Frank-Wolfe algorithm with recursive gradients for variance reduction to update and reveal the parameters upon each data. Combined with the adaptive DP analysis, our algorithm achieves the first optimal excess risk in linear time in the case $1 Cite this Paper

BibTeX

@InProceedings{pmlr-v162-han22d, title = {Private Streaming {SCO} in ℓ_p geometry with Applications in High Dimensional Online Decision Making}, author = {Han, Yuxuan and Liang, Zhicong and Liang, Zhipeng and Wang, Yang and Yao, Yuan and Zhang, Jiheng}, booktitle = {Proceedings of the 39th International Conference on Machine Learning}, pages = {8249--8279}, year = {2022}, editor = {Chaudhuri, Kamalika and Jegelka, Stefanie and Song, Le and Szepesvari, Csaba and Niu, Gang and Sabato, Sivan}, volume = {162}, series = {Proceedings of Machine Learning Research}, month = {17--23 Jul}, publisher = {PMLR}, pdf = {https://proceedings.mlr.press/v162/han22d/han22d.pdf}, url = {https://proceedings.mlr.press/v162/han22d.html}, abstract = {Differentially private (DP) stochastic convex optimization (SCO) is ubiquitous in trustworthy machine learning algorithm design. This paper studies the DP-SCO problem with streaming data sampled from a distribution and arrives sequentially. We also consider the continual release model where parameters related to private information are updated and released upon each new data. Numerous algorithms have been developed to achieve optimal excess risks in different ℓ_p norm geometries, but none of the existing ones can be adapted to the streaming and continual release setting. We propose a private variant of the Frank-Wolfe algorithm with recursive gradients for variance reduction to update and reveal the parameters upon each data. Combined with the adaptive DP analysis, our algorithm achieves the first optimal excess risk in linear time in the case $1 Copy to ClipboardDownload Endnote %0 Conference Paper %T Private Streaming SCO in ℓ_p geometry with Applications in High Dimensional Online Decision Making %A Yuxuan Han %A Zhicong Liang %A Zhipeng Liang %A Yang Wang %A Yuan Yao %A Jiheng Zhang %B Proceedings of the 39th International Conference on Machine Learning %C Proceedings of Machine Learning Research %D 2022 %E Kamalika Chaudhuri %E Stefanie Jegelka %E Le Song %E Csaba Szepesvari %E Gang Niu %E Sivan Sabato %F pmlr-v162-han22d %I PMLR %P 8249--8279 %U https://proceedings.mlr.press/v162/han22d.html %V 162 %X Differentially private (DP) stochastic convex optimization (SCO) is

ubiquitous in trustworthy machine learning algorithm design. This paper studies the DP-SCO problem with streaming data sampled from a distribution and arrives sequentially. We also consider the continual release model where parameters related to private information are updated and released upon each new data. Numerous algorithms have been developed to achieve optimal excess risks in different ℓ_p norm geometries, but none of the existing ones can be adapted to the streaming and continual release setting. We propose a private variant of the Frank-Wolfe algorithm with recursive gradients for variance reduction to update and reveal the parameters upon each data. Combined with the adaptive DP analysis, our algorithm achieves the first optimal excess risk in linear time in the case $1 Copy to ClipboardDownload APA

Han, Y., Liang, Z., Liang, Z., Wang, Y., Yao, Y. & Zhang, J.. (2022). Private Streaming SCO in ℓ_p geometry with Applications in High Dimensional Online Decision Making. Proceedings of the 39th International Conference on Machine Learning, in Proceedings of Machine Learning Research 162:8249-8279 Available from https://proceedings.mlr.press/v162/han22d.html.

Copy to ClipboardDownload

Related Material

Download PDF

Off-Policy Reinforcement Learning with Delayed Rewards

- Beining Han, Zhizhou Ren, Zuofan Wu, Yuan Zhou, Jian Peng
- abstract: We study deep reinforcement learning (RL) algorithms with delayed rewards. In many real-world tasks, instant rewards are often not readily accessible or even defined immediately after the agent performs actions. In this work, we first formally define the environment with delayed rewards and discuss the challenges raised due to the non-Markovian nature of such environments. Then, we introduce a general off-policy RL framework with a new Q-function formulation that can handle the delayed rewards with theoretical convergence guarantees. For practical tasks with high dimensional state spaces, we further introduce the HC-decomposition rule of the Q-function in our framework which naturally leads to an approximation scheme that helps boost the training efficiency and stability. We finally conduct extensive experiments to demonstrate the superior performance of our algorithms over the existing work and their variants.

Adversarial Attacks on Gaussian Process Bandits

- Eric Han, Jonathan Scarlett
- abstract: Gaussian processes (GP) are a widely-adopted tool used to sequentially optimize black-box functions, where evaluations are costly and potentially noisy. Recent works on GP bandits have proposed to move beyond random noise and devise algorithms robust to adversarial attacks. This paper studies this problem from the attacker’s perspective, proposing various adversarial attack methods with differing assumptions on the attacker’s strength and prior information. Our goal is to understand adversarial attacks on GP bandits from theoretical and practical perspectives. We focus primarily on targeted attacks on the popular GP-UCB algorithm and a related elimination-based algorithm, based on adversarially perturbing the function f to produce another function f whose optima are in some target region. Based on our theoretical analysis, we devise both white-box attacks (known f) and black-box attacks (unknown f), with the former including a Subtraction attack and Clipping attack, and the latter including an Aggressive subtraction attack. We demonstrate that adversarial attacks on GP bandits can succeed in forcing the algorithm towards the target region even with a low attack budget, and we test our attacks’ effectiveness on a diverse range of objective functions.

Random Gegenbauer Features for Scalable Kernel Methods

- Insu Han, Amir Zandieh, Haim Avron
- abstract: We propose efficient random features for approximating a new and rich class of kernel functions that we refer to as Generalized Zonal Kernels (GZK). Our proposed GZK family, generalizes the zonal kernels (i.e., dot-product kernels on the unit sphere) by introducing radial factors in the Gegenbauer series expansion of these kernel functions. The GZK class of kernels includes a wide range of ubiquitous kernel functions such as the entirety of dot-product kernels as well as the Gaussian and the recently introduced Neural Tangent kernels. Interestingly, by exploiting the reproducing property of the Gegenbauer (Zonal) Harmonics, we can construct efficient random features for the GZK family based on randomly oriented Gegenbauer harmonics. We prove subspace embedding guarantees for our Gegenbauer features which ensures that our features can be used for approximately solving learning problems such as kernel k-means clustering, kernel ridge regression, etc. Empirical results show that our proposed features outperform recent kernel approximation methods.

Stochastic Reweighted Gradient Descent

- Ayoub El Hanchi, David Stephens, Chris Maddison
- abstract: Importance sampling is a promising strategy for improving the convergence rate of stochastic gradient methods. It is typically used to precondition the optimization problem, but it can also be used to reduce the variance of the gradient estimator. Unfortunately, this latter point of view has yet to lead to practical methods that provably improve the asymptotic error of stochastic gradient methods. In this work, we propose stochastic reweighted gradient descent (SRG), a stochastic gradient method based solely on importance sampling that can reduce the variance of the gradient estimator and improve on the asymptotic error of stochastic gradient descent (SGD) in the strongly convex and smooth case. We show that SRG can be extended to combine the benefits of both importance-sampling-based preconditioning and variance reduction. When compared to SGD, the resulting algorithm can simultaneously reduce the condition number and the asymptotic error, both by up to a factor equal to the number of component functions. We demonstrate improved convergence in practice on regularized logistic regression problems.

Dual Perspective of Label-Specific Feature Learning for Multi-Label Classification

- Jun-Yi Hang, Min-Ling Zhang
- abstract: Label-specific features serve as an effective strategy to facilitate multi-label classification, which account for the distinct discriminative properties of each class label via tailoring its own features. Existing approaches implement this strategy in a quite straightforward way, i.e. finding the most pertinent and discriminative features for each class label and directly inducing classifiers on constructed label-specific features. In this paper, we propose a dual perspective for label-specific feature learning, where label-specific discriminative properties are considered by identifying each label’s own non-informative features and making the discrimination process immutable to variations of these features. To instantiate it, we present a perturbation-based approach DELA to provide classifiers with label-specific immutability on simultaneously identified non-informative features, which is optimized towards a probabilistically-relaxed expected risk minimization problem. Comprehensive experiments on 10 benchmark data sets show that our approach outperforms the state-of-the-art counterparts.

Temporal Difference Learning for Model Predictive Control

- Nicklas A Hansen, Hao Su, Xiaolong Wang
- abstract: Data-driven model predictive control has two key advantages over model-free methods: a potential for improved sample efficiency through model learning, and better performance as computational budget for planning increases. However, it is both costly to plan over long horizons and

challenging to obtain an accurate model of the environment. In this work, we combine the strengths of model-free and model-based methods. We use a learned task-oriented latent dynamics model for local trajectory optimization over a short horizon, and use a learned terminal value function to estimate long-term return, both of which are learned jointly by temporal difference learning. Our method, TD-MPC, achieves superior sample efficiency and asymptotic performance over prior work on both state and image-based continuous control tasks from DMControl and Meta-World. Code and videos are available at https://nicklashansen.github.io/td-mpc.

Bisimulation Makes Analogies in Goal-Conditioned Reinforcement Learning

- Philippe Hansen-Estruch, Amy Zhang, Ashvin Nair, Patrick Yin, Sergey Levine
- abstract: Building generalizable goal-conditioned agents from rich observations is a key to reinforcement learning (RL) solving real world problems. Traditionally in goal-conditioned RL, an agent is provided with the exact goal they intend to reach. However, it is often not realistic to know the configuration of the goal before performing a task. A more scalable framework would allow us to provide the agent with an example of an analogous task, and have the agent then infer what the goal should be for its current state. We propose a new form of state abstraction called goal-conditioned bisimulation that captures functional equivariance, allowing for the reuse of skills to achieve new goals. We learn this representation using a metric form of this abstraction, and show its ability to generalize to new goals in real world manipulation tasks. Further, we prove that this learned representation is sufficient not only for goal-conditioned tasks, but is amenable to any downstream task described by a state-only reward function.

TURF: Two-Factor, Universal, Robust, Fast Distribution Learning Algorithm

- Yi Hao, Ayush Jain, Alon Orlitsky, Vaishakh Ravindrakumar
- abstract: Approximating distributions from their samples is a canonical statistical-learning problem. One of its most powerful and successful modalities approximates every distribution to an ℓ_1 distance essentially at most a constant times larger than its closest t-piece degree-d polynomial, where $t\ge1$ and $d\ge0$. Letting $c_{t,d}$ denote the smallest such factor, clearly $c_{1,0}=1$, and it can be shown that $c_{t,d}\ge 2$ for all other t and d. Yet current computationally efficient algorithms show only $c_{t,1}\le 2.25$ and the bound rises quickly to $c_{t,d}\le 3$ for $d\ge 9$. We derive a near-linear-time and essentially sample-optimal estimator that establishes $c_{t,d}=2$ for all $(t,d)\ne(1,0)$. Additionally, for many practical distributions, the lowest approximation distance is achieved by polynomials with vastly varying number of pieces. We provide a method that estimates this number near-optimally, hence helps approach the best possible approximation. Experiments combining the two techniques confirm improved performance over existing methodologies.

Contextual Information-Directed Sampling

- Botao Hao, Tor Lattimore, Chao Qin
- abstract: Information-directed sampling (IDS) has recently demonstrated its potential as a data-efficient reinforcement learning algorithm. However, it is still unclear what is the right form of information ratio to optimize when contextual information is available. We investigate the IDS design through two contextual bandit problems: contextual bandits with graph feedback and sparse linear contextual bandits. We provably demonstrate the advantage of contextual IDS over conditional IDS and emphasize the importance of considering the context distribution. The main message is that an intelligent agent should invest more on the actions that are beneficial for the future unseen contexts while the conditional IDS can be myopic. We further propose a computationally-efficient version of contextual IDS based on Actor-Critic and evaluate it empirically on a neural network contextual bandit.

GSmooth: Certified Robustness against Semantic Transformations via Generalized Randomized Smoothing

- Zhongkai Hao, Chengyang Ying, Yinpeng Dong, Hang Su, Jian Song, Jun Zhu
- abstract: Certified defenses such as randomized smoothing have shown promise towards building reliable machine learning systems against ℓ_p norm bounded attacks. However, existing methods are insufficient or unable to provably defend against semantic transformations, especially those without closed-form expressions (such as defocus blur and pixelate), which are more common in practice and often unrestricted. To fill up this gap, we propose generalized randomized smoothing (GSmooth), a unified theoretical framework for certifying robustness against general semantic transformations via a novel dimension augmentation strategy. Under the GSmooth framework, we present a scalable algorithm that uses a surrogate image-to-image network to approximate the complex transformation. The surrogate model provides a powerful tool for studying the properties of semantic transformations and certifying robustness. Experimental results on several datasets demonstrate the effectiveness of our approach for robustness certification against multiple kinds of semantic transformations and corruptions, which is not achievable by the alternative baselines.

Implicit Regularization with Polynomial Growth in Deep Tensor Factorization

- Kais Hariz, Hachem Kadri, Stephane Ayache, Maher Moakher, Thierry Artieres
- abstract: We study the implicit regularization effects of deep learning in tensor factorization. While implicit regularization in deep matrix and 'shallow' tensor factorization via linear and certain type of non-linear neural networks promotes low-rank solutions with at most quadratic growth, we show that its effect in deep tensor factorization grows polynomially with the depth of the network. This provides a remarkably faithful description of the observed experimental behaviour. Using numerical experiments, we demonstrate the benefits of this implicit regularization in yielding a more accurate estimation and better convergence properties.

Strategic Instrumental Variable Regression: Recovering Causal Relationships From Strategic Responses

- Keegan Harris, Dung Daniel T Ngo, Logan Stapleton, Hoda Heidari, Steven Wu
- abstract: In settings where Machine Learning (ML) algorithms automate or inform consequential decisions about people, individual decision subjects are often incentivized to strategically modify their observable attributes to receive more favorable predictions. As a result, the distribution the assessment rule is trained on may differ from the one it operates on in deployment. While such distribution shifts, in general, can hinder accurate predictions, our work identifies a unique opportunity associated with shifts due to strategic responses: We show that we can use strategic responses effectively to recover causal relationships between the observable features and outcomes we wish to predict, even under the presence of unobserved confounding variables. Specifically, our work establishes a novel connection between strategic responses to ML models and instrumental variable (IV) regression by observing that the sequence of deployed models can be viewed as an instrument that affects agents' observable features but does not directly influence their outcomes. We show that our causal recovery method can be utilized to improve decision-making across several important criteria: individual fairness, agent outcomes, and predictive risk. In particular, we show that if decision subjects differ in their ability to modify non-causal attributes, any decision rule deviating from the causal coefficients can lead to (potentially unbounded) individual-level unfairness. .

C-algebra Net: A New Approach Generalizing Neural Network Parameters to C-algebra

- Yuka Hashimoto, Zhao Wang, Tomoko Matsui
- abstract: We propose a new framework that generalizes the parameters of neural network models to $C^$-algebra-valued ones. $C^$-algebra is a generalization of the space of complex numbers. A typical example is the space of continuous functions on a compact space. This generalization enables us to combine multiple models continuously and use tools for functions such as regression and integration. Consequently, we can learn features of data efficiently and adapt the models to problems continuously. We apply our framework to practical problems such as density estimation and few-shot

learning and show that our framework enables us to learn features of data even with a limited number of samples. Our new framework highlights the potential possibility of applying the theory of C^*-algebra to general neural network models.

General-purpose, long-context autoregressive modeling with Perceiver AR

- Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, Hannah Sheahan, Neil Zeghidour, Jean-Baptiste Alayrac, Joao Carreira, Jesse Engel
- abstract: Real-world data is high-dimensional: a book, image, or musical performance can easily contain hundreds of thousands of elements even after compression. However, the most commonly used autoregressive models, Transformers, are prohibitively expensive to scale to the number of inputs and layers needed to capture this long-range structure. We develop Perceiver AR, an autoregressive, modality-agnostic architecture which uses cross-attention to map long-range inputs to a small number of latents while also maintaining end-to-end causal masking. Perceiver AR can directly attend to over a hundred thousand tokens, enabling practical long-context density estimation without the need for hand-crafted sparsity patterns or memory mechanisms. When trained on images or music, Perceiver AR generates outputs with clear long-term coherence and structure. Our architecture also obtains state-of-the-art likelihood on long-sequence benchmarks, including 64x64 ImageNet images and PG-19 books.

On Distribution Shift in Learning-based Bug Detectors

- Jingxuan He, Luca Beurer-Kellner, Martin Vechev
- abstract: Deep learning has recently achieved initial success in program analysis tasks such as bug detection. Lacking real bugs, most existing works construct training and test data by injecting synthetic bugs into correct programs. Despite achieving high test accuracy (e.g., >90%), the resulting bug detectors are found to be surprisingly unusable in practice, i.e., <10% precision when used to scan real software repositories. In this work, we argue that this massive performance difference is caused by a distribution shift, i.e., a fundamental mismatch between the real bug distribution and the synthetic bug distribution used to train and evaluate the detectors. To address this key challenge, we propose to train a bug detector in two phases, first on a synthetic bug distribution to adapt the model to the bug detection domain, and then on a real bug distribution to drive the model towards the real distribution. During these two phases, we leverage a multi-task hierarchy, focal loss, and contrastive learning to further boost performance. We evaluate our approach extensively on three widely studied bug types, for which we construct new datasets carefully designed to capture the real bug distribution. The results demonstrate that our approach is practically effective and successfully mitigates the distribution shift: our learned detectors are highly performant on both our test set and the latest version of open source repositories. Our code, datasets, and models are publicly available at https://github.com/eth-sri/learning-real-bug-detector.

GNNRank: Learning Global Rankings from Pairwise Comparisons via Directed Graph Neural Networks

- Yixuan He, Quan Gan, David Wipf, Gesine D Reinert, Junchi Yan, Mihai Cucuringu
- abstract: Recovering global rankings from pairwise comparisons has wide applications from time synchronization to sports team ranking. Pairwise comparisons corresponding to matches in a competition can be construed as edges in a directed graph (digraph), whose nodes represent e.g. competitors with an unknown rank. In this paper, we introduce neural networks into the ranking recovery problem by proposing the so-called GNNRank, a trainable GNN-based framework with digraph embedding. Moreover, new objectives are devised to encode ranking upsets/violations. The framework involves a ranking score estimation approach, and adds an inductive bias by unfolding the Fiedler vector computation of the graph constructed from a learnable similarity matrix. Experimental results on extensive data sets show that our methods attain competitive and often superior performance against baselines, as well as showing promising transfer ability. Codes and preprocessed data are at: \url{https://github.com/SherylHYX/GNNRank}.

Exploring the Gap between Collapsed & Whitened Features in Self-Supervised Learning

- Bobby He, Mete Ozay
- abstract: Avoiding feature collapse, when a Neural Network (NN) encoder maps all inputs to a constant vector, is a shared implicit desideratum of various methodological advances in self-supervised learning (SSL). To that end, whitened features have been proposed as an explicit objective to ensure uncollapsed features \cite{zbontar2021barlow,ermolov2021whitening,hua2021feature,bardes2022vicreg}. We identify power law behaviour in eigenvalue decay, parameterised by exponent $\beta{\geq}0$, as a spectrum that bridges between the collapsed & whitened feature extremes. We provide theoretical & empirical evidence highlighting the factors in SSL, like projection layers & regularisation strength, that influence eigenvalue decay rate, & demonstrate that the degree of feature whitening affects generalisation, particularly in label scarce regimes. We use our insights to motivate a novel method, PMP (PostMan-Pat), which efficiently post-processes a pretrained encoder to enforce eigenvalue decay rate with power law exponent β, & find that PostMan-Pat delivers improved label efficiency and transferability across a range of SSL methods and encoder architectures.

Sparse Double Descent: Where Network Pruning Aggravates Overfitting

- Zheng He, Zeke Xie, Quanzhi Zhu, Zengchang Qin
- abstract: People usually believe that network pruning not only reduces the computational cost of deep networks, but also prevents overfitting by decreasing model capacity. However, our work surprisingly discovers that network pruning sometimes even aggravates overfitting. We report an unexpected sparse double descent phenomenon that, as we increase model sparsity via network pruning, test performance first gets worse (due to overfitting), then gets better (due to relieved overfitting), and gets worse at last (due to forgetting useful information). While recent studies focused on the deep double descent with respect to model overparameterization, they failed to recognize that sparsity may also cause double descent. In this paper, we have three main contributions. First, we report the novel sparse double descent phenomenon through extensive experiments. Second, for this phenomenon, we propose a novel learning distance interpretation that the curve of l2 learning distance of sparse models (from initialized parameters to final parameters) may correlate with the sparse double descent curve well and reflect generalization better than minima flatness. Third, in the context of sparse double descent, a winning ticket in the lottery ticket hypothesis surprisingly may not always win.

A Reduction from Linear Contextual Bandits Lower Bounds to Estimations Lower Bounds

- Jiahao He, Jiheng Zhang, Rachel Q. Zhang
- abstract: Linear contextual bandits and their variants are usually solved using algorithms guided by parameter estimation. Cauchy-Schwartz inequality established that estimation errors dominate algorithm regrets, and thus, accurate estimators suffice to guarantee algorithms with low regrets. In this paper, we complete the reverse direction by establishing the necessity. In particular, we provide a generic transformation from algorithms for linear contextual bandits to estimators for linear models, and show that algorithm regrets dominate estimation errors of their induced estimators, i.e., low-regret algorithms must imply accurate estimators. Moreover, our analysis reduces the regret lower bound to an estimation error, bridging the lower bound analysis in linear contextual bandit problems and linear regression.

HyperPrompt: Prompt-based Task-Conditioning of Transformers

- Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, Yaguang Li, Zhao Chen, Donald Metzler, Heng-Tze Cheng, Ed H. Chi
- abstract: Prompt-Tuning is a new paradigm for finetuning pre-trained language models in a parameter efficient way. Here, we explore the use of HyperNetworks to generate hyper-prompts: we propose HyperPrompt, a novel architecture for prompt-based task-conditioning of self-attention in

Transformers. The hyper-prompts are end-to-end learnable via generation by a HyperNetwork. HyperPrompt allows the network to learn task-specific feature maps where the hyper-prompts serve as task global memories for the queries to attend to, at the same time enabling flexible information sharing among tasks. We show that HyperPrompt is competitive against strong multi-task learning baselines with as few as 0.14% of additional task-conditioning parameters, achieving great parameter and computational efficiency. Through extensive empirical experiments, we demonstrate that HyperPrompt can achieve superior performances over strong T5 multi-task learning baselines and parameter-efficient adapter variants including Prompt-Tuning and HyperFormer++ on Natural Language Understanding benchmarks of GLUE and SuperGLUE across many model sizes.

Label-Descriptive Patterns and Their Application to Characterizing Classification Errors

- Michael A. Hedderich, Jonas Fischer, Dietrich Klakow, Jilles Vreeken
- abstract: State-of-the-art deep learning methods achieve human-like performance on many tasks, but make errors nevertheless. Characterizing these errors in easily interpretable terms gives insight into whether a classifier is prone to making systematic errors, but also gives a way to act and improve the classifier. We propose to discover those feature-value combinations (i.e., patterns) that strongly correlate with correct resp. erroneous predictions to obtain a global and interpretable description for arbitrary classifiers. We show this is an instance of the more general label description problem, which we formulate in terms of the Minimum Description Length principle. To discover a good pattern set, we develop the efficient Premise algorithm. Through an extensive set of experiments we show it performs very well in practice on both synthetic and real-world data. Unlike existing solutions, it ably recovers ground truth patterns, even on highly imbalanced data over many features. Through two case studies on Visual Question Answering and Named Entity Recognition, we confirm that Premise gives clear and actionable insight into the systematic errors made by modern NLP classifiers.

NOMU: Neural Optimization-based Model Uncertainty

- Jakob M Heiss, Jakob Weissteiner, Hanna S Wutte, Sven Seuken, Josef Teichmann
- abstract: We study methods for estimating model uncertainty for neural networks (NNs) in regression. To isolate the effect of model uncertainty, we focus on a noiseless setting with scarce training data. We introduce five important desiderata regarding model uncertainty that any method should satisfy. However, we find that established benchmarks often fail to reliably capture some of these desiderata, even those that are required by Bayesian theory. To address this, we introduce a new approach for capturing model uncertainty for NNs, which we call Neural Optimization-based Model Uncertainty (NOMU). The main idea of NOMU is to design a network architecture consisting of two connected sub-NNs, one for model prediction and one for model uncertainty, and to train it using a carefully-designed loss function. Importantly, our design enforces that NOMU satisfies our five desiderata. Due to its modular architecture, NOMU can provide model uncertainty for any given (previously trained) NN if given access to its training data. We evaluate NOMU in various regressions tasks and noiseless Bayesian optimization (BO) with costly evaluations. In regression, NOMU performs at least as well as state-of-the-art methods. In BO, NOMU even outperforms all considered benchmarks.

Scaling Out-of-Distribution Detection for Real-World Settings

- Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, Dawn Song
- abstract: Detecting out-of-distribution examples is important for safety-critical machine learning applications such as detecting novel biological phenomena and self-driving cars. However, existing research mainly focuses on simple small-scale settings. To set the stage for more realistic out-of-distribution detection, we depart from small-scale settings and explore large-scale multiclass and multi-label settings with high-resolution images and thousands of classes. To make future work in real-world settings possible, we create new benchmarks for three large-scale settings. To test ImageNet multiclass anomaly detectors, we introduce the Species dataset containing over 700,000 images and over a thousand anomalous species. We leverage ImageNet-21K to evaluate PASCAL VOC and COCO multilabel anomaly detectors. Third, we introduce a new benchmark for anomaly segmentation by introducing a segmentation benchmark with road anomalies. We conduct extensive experiments in these more realistic settings for out-of-distribution detection and find that a surprisingly simple detector based on the maximum logit outperforms prior methods in all the large-scale multi-class, multi-label, and segmentation tasks, establishing a simple new baseline for future work.

Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers

- Liam Hodgkinson, Umut Simsekli, Rajiv Khanna, Michael Mahoney
- abstract: Despite the ubiquitous use of stochastic optimization algorithms in machine learning, the precise impact of these algorithms and their dynamics on generalization performance in realistic non-convex settings is still poorly understood. While recent work has revealed connections between generalization and heavy-tailed behavior in stochastic optimization, they mainly relied on continuous-time approximations; and a rigorous treatment for the original discrete-time iterations is yet to be performed. To bridge this gap, we present novel bounds linking generalization to the lower tail exponent of the transition kernel associated with the optimizer around a local minimum, in both discrete- and continuous-time settings. To achieve this, we first prove a data- and algorithm-dependent generalization bound in terms of the celebrated Fernique-Talagrand functional applied to the trajectory of the optimizer. Then, we specialize this result by exploiting the Markovian structure of stochastic optimizers, and derive bounds in terms of their (data-dependent) transition kernels. We support our theory with empirical results from a variety of neural networks, showing correlations between generalization error and lower tail exponents.

Unsupervised Detection of Contextualized Embedding Bias with Application to Ideology

- Valentin Hofmann, Janet Pierrehumbert, Hinrich Schütze
- abstract: We propose a fully unsupervised method to detect bias in contextualized embeddings. The method leverages the assortative information latently encoded by social networks and combines orthogonality regularization, structured sparsity learning, and graph neural networks to find the embedding subspace capturing this information. As a concrete example, we focus on the phenomenon of ideological bias: we introduce the concept of an ideological subspace, show how it can be found by applying our method to online discussion forums, and present techniques to probe it. Our experiments suggest that the ideological subspace encodes abstract evaluative semantics and reflects changes in the political left-right spectrum during the presidency of Donald Trump.

Neural Laplace: Learning diverse classes of differential equations in the Laplace domain

- Samuel I Holt, Zhaozhi Qian, Mihaela van der Schaar
- abstract: Neural Ordinary Differential Equations model dynamical systems with ODEs learned by neural networks. However, ODEs are fundamentally inadequate to model systems with long-range dependencies or discontinuities, which are common in engineering and biological systems. Broader classes of differential equations (DE) have been proposed as remedies, including delay differential equations and integro-differential equations. Furthermore, Neural ODE suffers from numerical instability when modelling stiff ODEs and ODEs with piecewise forcing functions. In this work, we propose Neural Laplace, a unifying framework for learning diverse classes of DEs including all the aforementioned ones. Instead of modelling the dynamics in the time domain, we model it in the Laplace domain, where the history-dependencies and discontinuities in time can be represented as summations of complex exponentials. To make learning more efficient, we use the geometrical stereographic map of a Riemann sphere to induce more smoothness in the Laplace domain. In the experiments, Neural Laplace shows superior performance in modelling and extrapolating the trajectories of diverse classes of DEs, including the ones with complex history dependency and abrupt changes.

Deep Hierarchy in Bandits

- Joey Hong, Branislav Kveton, Sumeet Katariya, Manzil Zaheer, Mohammad Ghavamzadeh
- abstract: Mean rewards of actions are often correlated. The form of these correlations may be complex and unknown a priori, such as the preferences of users for recommended products and their categories. To maximize statistical efficiency, it is important to leverage these correlations when learning. We formulate a bandit variant of this problem where the correlations of mean action rewards are represented by a hierarchical Bayesian model with latent variables. Since the hierarchy can have multiple layers, we call it deep. We propose a hierarchical Thompson sampling algorithm (HierTS) for this problem and show how to implement it efficiently for Gaussian hierarchies. The efficient implementation is possible due to a novel exact hierarchical representation of the posterior, which itself is of independent interest. We use this exact posterior to analyze the Bayes regret of HierTS. Our regret bounds reflect the structure of the problem, that the regret decreases with more informative priors, and can be recast to highlight reduced dependence on the number of actions. We confirm these theoretical findings empirically, in both synthetic and real-world experiments.

DAdaQuant: Doubly-adaptive quantization for communication-efficient Federated Learning

- Robert Hönig, Yiren Zhao, Robert Mullins
- abstract: Federated Learning (FL) is a powerful technique to train a model on a server with data from several clients in a privacy-preserving manner. FL incurs significant communication costs because it repeatedly transmits the model between the server and clients. Recently proposed algorithms quantize the model parameters to efficiently compress FL communication. We find that dynamic adaptations of the quantization level can boost compression without sacrificing model quality. We introduce DAdaQuant as a doubly-adaptive quantization algorithm that dynamically changes the quantization level across time and different clients. Our experiments show that DAdaQuant consistently improves client$\rightarrow$server compression, outperforming the strongest non-adaptive baselines by up to $2.8\times$.

Equivariant Diffusion for Molecule Generation in 3D

- Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, Max Welling
- abstract: This work introduces a diffusion model for molecule generation in 3D that is equivariant to Euclidean transformations. Our E(3) Equivariant Diffusion Model (EDM) learns to denoise a diffusion process with an equivariant network that jointly operates on both continuous (atom coordinates) and categorical features (atom types). In addition, we provide a probabilistic analysis which admits likelihood computation of molecules using our model. Experimentally, the proposed method significantly outperforms previous 3D molecular generative methods regarding the quality of generated samples and the efficiency at training time.

Conditional GANs with Auxiliary Discriminative Classifier

- Liang Hou, Qi Cao, Huawei Shen, Siyuan Pan, Xiaoshuang Li, Xueqi Cheng
- abstract: Conditional generative models aim to learn the underlying joint distribution of data and labels to achieve conditional data generation. Among them, the auxiliary classifier generative adversarial network (AC-GAN) has been widely used, but suffers from the problem of low intra-class diversity of the generated samples. The fundamental reason pointed out in this paper is that the classifier of AC-GAN is generator-agnostic, which therefore cannot provide informative guidance for the generator to approach the joint distribution, resulting in a minimization of the conditional entropy that decreases the intra-class diversity. Motivated by this understanding, we propose a novel conditional GAN with an auxiliary discriminative classifier (ADC-GAN) to resolve the above problem. Specifically, the proposed auxiliary discriminative classifier becomes generator-aware by recognizing the class-labels of the real data and the generated data discriminatively. Our theoretical analysis reveals that the generator can faithfully learn the joint distribution even without the original discriminator, making the proposed ADC-GAN robust to the value of the coefficient hyperparameter and the selection of the GAN loss, and stable during training. Extensive experimental results on synthetic and real-world datasets demonstrate the superiority of ADC-GAN in conditional generative modeling compared to state-of-the-art classifier-based and projection-based conditional GANs.

AdAUC: End-to-end Adversarial AUC Optimization Against Long-tail Problems

- Wenzheng Hou, Qianqian Xu, Zhiyong Yang, Shilong Bao, Yuan He, Qingming Huang
- abstract: It is well-known that deep learning models are vulnerable to adversarial examples. Existing studies of adversarial training have made great progress against this challenge. As a typical trait, they often assume that the class distribution is overall balanced. However, long-tail datasets are ubiquitous in a wide spectrum of applications, where the amount of head class instances is significantly larger than the tail classes. Under such a scenario, AUC is a much more reasonable metric than accuracy since it is insensitive toward class distribution. Motivated by this, we present an early trial to explore adversarial training methods to optimize AUC. The main challenge lies in that the positive and negative examples are tightly coupled in the objective function. As a direct result, one cannot generate adversarial examples without a full scan of the dataset. To address this issue, based on a concavity regularization scheme, we reformulate the AUC optimization problem as a saddle point problem, where the objective becomes an instance-wise function. This leads to an end-to-end training protocol. Furthermore, we provide a convergence guarantee of the proposed training algorithm. Our analysis differs from the existing studies since the algorithm is asked to generate adversarial examples by calculating the gradient of a min-max problem. Finally, the extensive experimental results show the performance and robustness of our algorithm in three long-tail datasets.

Wide Bayesian neural networks have a simple weight posterior: theory and accelerated sampling

- Jiri Hron, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein
- abstract: We introduce repriorisation, a data-dependent reparameterisation which transforms a Bayesian neural network (BNN) posterior to a distribution whose KL divergence to the BNN prior vanishes as layer widths grow. The repriorisation map acts directly on parameters, and its analytic simplicity complements the known neural network Gaussian process (NNGP) behaviour of wide BNNs in function space. Exploiting the repriorisation, we develop a Markov chain Monte Carlo (MCMC) posterior sampling algorithm which mixes faster the wider the BNN. This contrasts with the typically poor performance of MCMC in high dimensions. We observe up to 50x higher effective sample size relative to no reparametrisation for both fully-connected and residual networks. Improvements are achieved at all widths, with the margin between reparametrised and standard BNNs growing with layer width.

Learning inverse folding from millions of predicted structures

- Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, Alexander Rives
- abstract: We consider the problem of predicting a protein sequence from its backbone atom coordinates. Machine learning approaches to this problem to date have been limited by the number of available experimentally determined protein structures. We augment training data by nearly three orders of magnitude by predicting structures for 12M protein sequences using AlphaFold2. Trained with this additional data, a sequence-to-sequence transformer with invariant geometric input processing layers achieves 51% native sequence recovery on structurally held-out backbones with 72% recovery for buried residues, an overall improvement of almost 10 percentage points over existing methods. The model generalizes to a variety of more complex tasks including design of protein complexes, partially masked structures, binding interfaces, and multiple states.

Nearly Minimax Optimal Reinforcement Learning with Linear Function Approximation

- Pihe Hu, Yu Chen, Longbo Huang
- abstract: We study reinforcement learning with linear function approximation where the transition probability and reward functions are linear with respect to a feature mapping $\boldsymbol{\phi}(s,a)$. Specifically, we consider the episodic inhomogeneous linear Markov Decision Process (MDP), and propose a novel computation-efficient algorithm, LSVI-UCB$^+$, which achieves an $\widetilde{O}(Hd\sqrt{T})$ regret bound where H is the episode length, d is the feature dimension, and T is the number of steps. LSVI-UCB$^+$ builds on weighted ridge regression and upper confidence value iteration with a Bernstein-type exploration bonus. Our statistical results are obtained with novel analytical tools, including a new Bernstein self-normalized bound with conservatism on elliptical potentials, and refined analysis of the correction term. To the best of our knowledge, this is the first minimax optimal algorithm for linear MDPs up to logarithmic factors, which closes the $\sqrt{Hd}$ gap between the best known upper bound of $\widetilde{O}(\sqrt{H^3d^3T})$ in \cite{jin2020provably} and lower bound of $\Omega(Hd\sqrt{T})$ for linear MDPs.

Neuron Dependency Graphs: A Causal Abstraction of Neural Networks

- Yaojie Hu, Jin Tian
- abstract: We discover that neural networks exhibit approximate logical dependencies among neurons, and we introduce Neuron Dependency Graphs (NDG) that extract and present them as directed graphs. In an NDG, each node corresponds to the boolean activation value of a neuron, and each edge models an approximate logical implication from one node to another. We show that the logical dependencies extracted from the training dataset generalize well to the test set. In addition to providing symbolic explanations to the neural network's internal structure, NDGs can represent a Structural Causal Model. We empirically show that an NDG is a causal abstraction of the corresponding neural network that "unfolds" the same way under causal interventions using the theory by Geiger et al. (2021). Code is available at https://github.com/phimachine/ndg.

Policy Diagnosis via Measuring Role Diversity in Cooperative Multi-agent RL

- Siyi Hu, Chuanlong Xie, Xiaodan Liang, Xiaojun Chang
- abstract: Cooperative multi-agent reinforcement learning (MARL) is making rapid progress for solving tasks in a grid world and real-world scenarios, in which agents are given different attributes and goals, resulting in different behavior through the whole multi-agent task. In this study, we quantify the agent's behavior difference and build its relationship with the policy performance via {\bf Role Diversity}, a metric to measure the characteristics of MARL tasks. We define role diversity from three perspectives: action-based, trajectory-based, and contribution-based to fully measure a multi-agent task. Through theoretical analysis, we find that the error bound in MARL can be decomposed into three parts that have a strong relation to the role diversity. The decomposed factors can significantly impact policy optimization in three popular directions including parameter sharing, communication mechanism, and credit assignment. The main experimental platforms are based on {\bf Multiagent Particle Environment (MPE) }and {\bf The StarCraft Multi-Agent Challenge (SMAC)}. Extensive experiments clearly show that role diversity can serve as a robust measurement for the characteristics of a multi-agent cooperation task and help diagnose whether the policy fits the current multi-agent system for better policy performance.

On the Role of Discount Factor in Offline Reinforcement Learning

- Hao Hu, Yiqin Yang, Qianchuan Zhao, Chongjie Zhang
- abstract: Offline reinforcement learning (RL) enables effective learning from previously collected data without exploration, which shows great promise in real-world applications when exploration is expensive or even infeasible. The discount factor, γ, plays a vital role in improving online RL sample efficiency and estimation accuracy, but the role of the discount factor in offline RL is not well explored. This paper examines two distinct effects of γ in offline RL with theoretical analysis, namely the regularization effect and the pessimism effect. On the one hand, γ is a regulator to trade-off optimality with sample efficiency upon existing offline techniques. On the other hand, lower guidance γ can also be seen as a way of pessimism where we optimize the policy's performance in the worst possible models. We empirically verify the above theoretical observation with tabular MDPs and standard D4RL tasks. The results show that the discount factor plays an essential role in the performance of offline RL algorithms, both under small data regimes upon existing offline methods and in large data regimes without other conservative methods.

Transformer Quality in Linear Time

- Weizhe Hua, Zihang Dai, Hanxiao Liu, Quoc Le
- abstract: We revisit the design choices in Transformers, and propose methods to address their weaknesses in handling long sequences. First, we propose a simple layer named gated attention unit, which allows the use of a weaker single-head attention with minimal quality loss. We then propose a linear approximation method complementary to this new layer, which is accelerator-friendly and highly competitive in quality. The resulting model, named FLASH, matches the perplexity of improved Transformers over both short (512) and long (8K) context lengths, achieving training speedups of up to 4.9x on Wiki-40B and 12.1x on PG-19 for auto-regressive language modeling, and 4.8x on C4 for masked language modeling.

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

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

Forward Operator Estimation in Generative Models with Kernel Transfer Operators

- Zhichun Huang, Rudrasis Chakraborty, Vikas Singh
- abstract: Generative models which use explicit density modeling (e.g., variational autoencoders, flow-based generative models) involve finding a mapping from a known distribution, e.g. Gaussian, to the unknown input distribution. This often requires searching over a class of non-linear functions (e.g., representable by a deep neural network). While effective in practice, the associated runtime/memory costs can increase rapidly, usually as a function of the performance desired in an application. We propose a substantially cheaper (and simpler) forward operator estimation strategy based on adapting known results on kernel transfer operators. We show that our formulation enables highly efficient distribution approximation and sampling, and offers surprisingly good empirical performance that compares favorably with powerful baselines, but with significant runtime savings. We show that the algorithm also performs well in small sample size settings (in brain imaging).

Adaptive Best-of-Both-Worlds Algorithm for Heavy-Tailed Multi-Armed Bandits

- Jiatai Huang, Yan Dai, Longbo Huang

- abstract: In this paper, we generalize the concept of heavy-tailed multi-armed bandits to adversarial environments, and develop robust best-of-both-worlds algorithms for heavy-tailed multi-armed bandits (MAB), where losses have α-th ($1<\alpha\le 2$) moments bounded by σ^α, while the variances may not exist. Specifically, we design an algorithm \texttt{HTINF}, when the heavy-tail parameters α and σ are known to the agent, \texttt{HTINF} simultaneously achieves the optimal regret for both stochastic and adversarial environments, without knowing the actual environment type a-priori. When α,σ are unknown, \texttt{HTINF} achieves a $\log T$-style instance-dependent regret in stochastic cases and $o(T)$ no-regret guarantee in adversarial cases. We further develop an algorithm \texttt{AdaTINF}, achieving $\mathcal O(\sigma K^{1-\nicefrac 1\alpha}T^{\nicefrac{1}{\alpha}})$ minimax optimal regret even in adversarial settings, without prior knowledge on α and σ. This result matches the known regret lower-bound (Bubeck et al., 2013), which assumed a stochastic environment and α and σ are both known. To our knowledge, the proposed \texttt{HTINF} algorithm is the first to enjoy a best-of-both-worlds regret guarantee, and \texttt{AdaTINF} is the first algorithm that can adapt to both α and σ to achieve optimal gap-indepedent regret bound in classical heavy-tailed stochastic MAB setting and our novel adversarial formulation.

Frustratingly Easy Transferability Estimation

- Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, Ying Wei
- abstract: Transferability estimation has been an essential tool in selecting a pre-trained model and the layers in it for transfer learning, to transfer, so as to maximize the performance on a target task and prevent negative transfer. Existing estimation algorithms either require intensive training on target tasks or have difficulties in evaluating the transferability between layers. To this end, we propose a simple, efficient, and effective transferability measure named TransRate. Through a single pass over examples of a target task, TransRate measures the transferability as the mutual information between features of target examples extracted by a pre-trained model and their labels. We overcome the challenge of efficient mutual information estimation by resorting to coding rate that serves as an effective alternative to entropy. From the perspective of feature representation, the resulting TransRate evaluates both completeness (whether features contain sufficient information of a target task) and compactness (whether features of each class are compact enough for good generalization) of pre-trained features. Theoretically, we have analyzed the close connection of TransRate to the performance after transfer learning. Despite its extraordinary simplicity in 10 lines of codes, TransRate performs remarkably well in extensive evaluations on 35 pre-trained models and 16 downstream tasks.

Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably)

- Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, Longbo Huang
- abstract: Despite the remarkable success of deep multi-modal learning in practice, it has not been well-explained in theory. Recently, it has been observed that the best uni-modal network outperforms the jointly trained multi-modal network across different combinations of modalities on various tasks, which is counter-intuitive since multiple signals would bring more information (Wang et al., 2020). This work provides a theoretical explanation for the emergence of such performance gap in neural networks for the prevalent joint training framework. Based on a simplified data distribution that captures the realistic property of multi-modal data, we prove that for multi-modal late-fusion network with (smoothed) ReLU activation trained jointly by gradient descent, different modalities will compete with each other and only a subset of modalities will be learned by its corresponding encoder networks. We refer to this phenomenon as modality competition, and the losing modalities, which fail to be discovered, are the origins where the sub-optimality of joint training comes from. In contrast, for uni-modal networks with similar learning settings, we provably show that the networks will focus on learning modality-associated features. Experimentally, we illustrate that modality competition matches the intrinsic behavior of late-fusion joint training to supplement our theoretical results. To the best of our knowledge, our work is the first theoretical treatment towards the degenerating aspect of multi-modal learning in neural networks.

Action-Sufficient State Representation Learning for Control with Structural Constraints

- Biwei Huang, Chaochao Lu, Liu Leqi, Jose Miguel Hernandez-Lobato, Clark Glymour, Bernhard Schölkopf, Kun Zhang
- abstract: Perceived signals in real-world scenarios are usually high-dimensional and noisy, and finding and using their representation that contains essential and sufficient information required by downstream decision-making tasks will help improve computational efficiency and generalization ability in the tasks. In this paper, we focus on partially observable environments and propose to learn a minimal set of state representations that capture sufficient information for decision-making, termed Action-Sufficient state Representations (ASRs). We build a generative environment model for the structural relationships among variables in the system and present a principled way to characterize ASRs based on structural constraints and the goal of maximizing cumulative reward in policy learning. We then develop a structured sequential Variational Auto-Encoder to estimate the environment model and extract ASRs. Our empirical results on CarRacing and VizDoom demonstrate a clear advantage of learning and using ASRs for policy learning. Moreover, the estimated environment model and ASRs allow learning behaviors from imagined outcomes in the compact latent space to improve sample efficiency.

3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design

- Yinan Huang, Xingang Peng, Jianzhu Ma, Muhan Zhang
- abstract: Deep learning has achieved tremendous success in designing novel chemical compounds with desirable pharmaceutical properties. In this work, we focus on a new type of drug design problem — generating a small “linker” to physically attach two independent molecules with their distinct functions. The main computational challenges include: 1) the generation of linkers is conditional on the two given molecules, in contrast to generating complete molecules from scratch in previous works; 2) linkers heavily depend on the anchor atoms of the two molecules to be connected, which are not known beforehand; 3) 3D structures and orientations of the molecules need to be considered to avoid atom clashes, for which equivariance to E(3) group are necessary. To address these problems, we propose a conditional generative model, named 3DLinker, which is able to predict anchor atoms and jointly generate linker graphs and their 3D structures based on an E(3) equivariant graph variational autoencoder. So far as we know, no previous models could achieve this task. We compare our model with multiple conditional generative models modified from other molecular design tasks and find that our model has a significantly higher rate in recovering molecular graphs, and more importantly, accurately predicting the 3D coordinates of all the atoms.

SDQ: Stochastic Differentiable Quantization with Mixed Precision

- Xijie Huang, Zhiqiang Shen, Shichao Li, Zechun Liu, Hu Xianghong, Jeffry Wicaksana, Eric Xing, Kwang-Ting Cheng
- abstract: In order to deploy deep models in a computationally efficient manner, model quantization approaches have been frequently used. In addition, as new hardware that supports various-bit arithmetic operations, recent research on mixed precision quantization (MPQ) begins to fully leverage the capacity of representation by searching various bitwidths for different layers and modules in a network. However, previous studies mainly search the MPQ strategy in a costly scheme using reinforcement learning, neural architecture search, etc., or simply utilize partial prior knowledge for bitwidth distribution, which might be biased and sub-optimal. In this work, we present a novel Stochastic Differentiable Quantization (SDQ) method that can automatically learn the MPQ strategy in a more flexible and globally-optimized space with a smoother gradient approximation. Particularly, Differentiable Bitwidth Parameters (DBPs) are employed as the probability factors in stochastic quantization between adjacent bitwidth. After the optimal MPQ strategy is acquired, we further train our network with the entropy-aware bin regularization and knowledge distillation. We extensively evaluate our method on different networks, hardwares (GPUs and FPGA), and datasets. SDQ outperforms all other state-of-the-art mixed or single precision quantization with less bitwidth, and are even better than the original full-precision counterparts across various ResNet and MobileNet families, demonstrating the effectiveness and superiority of our method. Code will be publicly available.

Tackling Data Heterogeneity: A New Unified Framework for Decentralized SGD with Sample-induced Topology

- Yan Huang, Ying Sun, Zehan Zhu, Changzhi Yan, Jinming Xu
- abstract: We develop a general framework unifying several gradient-based stochastic optimization methods for empirical risk minimization problems both in centralized and distributed scenarios. The framework hinges on the introduction of an augmented graph consisting of nodes modeling the samples and edges modeling both the inter-device communication and intra-device stochastic gradient computation. By designing properly the topology of the augmented graph, we are able to recover as special cases the renowned Local-SGD and DSGD algorithms, and provide a unified perspective for variance-reduction (VR) and gradient-tracking (GT) methods such as SAGA, Local-SVRG and GT-SAGA. We also provide a unified convergence analysis for smooth and (strongly) convex objectives relying on a proper structured Lyapunov function, and the obtained rate can recover the best known results for many existing algorithms. The rate results further reveal that VR and GT methods can effectively eliminate data heterogeneity within and across devices, respectively, enabling the exact convergence of the algorithm to the optimal solution. Numerical experiments confirm the findings in this paper.

Efficient Representation Learning via Adaptive Context Pooling

- Chen Huang, Walter Talbott, Navdeep Jaitly, Joshua M Susskind
- abstract: Self-attention mechanisms model long-range context by using pairwise attention between all input tokens. In doing so, they assume a fixed attention granularity defined by the individual tokens (e.g., text characters or image pixels), which may not be optimal for modeling complex dependencies at higher levels. In this paper, we propose ContextPool to address this problem by adapting the attention granularity for each token. Inspired by the success of ConvNets that are combined with pooling to capture long-range dependencies, we learn to pool neighboring features for each token before computing attention in a given attention layer. The pooling weights and support size are adaptively determined, allowing the pooled features to encode meaningful context with varying scale. We show that ContextPool makes attention models more expressive, achieving strong performance often with fewer layers and thus significantly reduced cost. Experiments validate that our ContextPool module, when plugged into transformer models, matches or surpasses state-of-the-art performance using less compute on several language and image benchmarks, outperforms recent works with learned context sizes or sparse attention patterns, and is also applicable to ConvNets for efficient feature learning.

On the Learning of Non-Autoregressive Transformers

- Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, Minlie Huang
- abstract: Non-autoregressive Transformer (NAT) is a family of text generation models, which aims to reduce the decoding latency by predicting the whole sentences in parallel. However, such latency reduction sacrifices the ability to capture left-to-right dependencies, thereby making NAT learning very challenging. In this paper, we present theoretical and empirical analyses to reveal the challenges of NAT learning and propose a unified perspective to understand existing successes. First, we show that simply training NAT by maximizing the likelihood can lead to an approximation of marginal distributions but drops all dependencies between tokens, where the dropped information can be measured by the dataset's conditional total correlation. Second, we formalize many previous objectives in a unified framework and show that their success can be concluded as maximizing the likelihood on a proxy distribution, leading to a reduced information loss. Empirical studies show that our perspective can explain the phenomena in NAT learning and guide the design of new training methods.

Going Deeper into Permutation-Sensitive Graph Neural Networks

- Zhongyu Huang, Yingheng Wang, Chaozhuo Li, Huiguang He
- abstract: The invariance to permutations of the adjacency matrix, i.e., graph isomorphism, is an overarching requirement for Graph Neural Networks (GNNs). Conventionally, this prerequisite can be satisfied by the invariant operations over node permutations when aggregating messages. However, such an invariant manner may ignore the relationships among neighboring nodes, thereby hindering the expressivity of GNNs. In this work, we devise an efficient permutation-sensitive aggregation mechanism via permutation groups, capturing pairwise correlations between neighboring nodes. We prove that our approach is strictly more powerful than the 2-dimensional Weisfeiler-Lehman (2-WL) graph isomorphism test and not less powerful than the 3-WL test. Moreover, we prove that our approach achieves the linear sampling complexity. Comprehensive experiments on multiple synthetic and real-world datasets demonstrate the superiority of our model.

Directed Acyclic Transformer for Non-Autoregressive Machine Translation

- Fei Huang, Hao Zhou, Yang Liu, Hang Li, Minlie Huang
- abstract: Non-autoregressive Transformers (NATs) significantly reduce the decoding latency by generating all tokens in parallel. However, such independent predictions prevent NATs from capturing the dependencies between the tokens for generating multiple possible translations. In this paper, we propose Directed Acyclic Transfomer (DA-Transformer), which represents the hidden states in a Directed Acyclic Graph (DAG), where each path of the DAG corresponds to a specific translation. The whole DAG simultaneously captures multiple translations and facilitates fast predictions in a non-autoregressive fashion. Experiments on the raw training data of WMT benchmark show that DA-Transformer substantially outperforms previous NATs by about 3 BLEU on average, which is the first NAT model that achieves competitive results with autoregressive Transformers without relying on knowledge distillation.

Unsupervised Ground Metric Learning Using Wasserstein Singular Vectors

- Geert-Jan Huizing, Laura Cantini, Gabriel Peyré
- abstract: Defining meaningful distances between samples in a dataset is a fundamental problem in machine learning. Optimal Transport (OT) lifts a distance between features (the "ground metric") to a geometrically meaningful distance between samples. However, there is usually no straightforward choice of ground metric. Supervised ground metric learning approaches exist but require labeled data. In absence of labels, only ad-hoc ground metrics remain. Unsupervised ground metric learning is thus a fundamental problem to enable data-driven applications of OT. In this paper, we propose for the first time a canonical answer by simultaneously computing an OT distance between samples and between features of a dataset. These distance matrices emerge naturally as positive singular vectors of the function mapping ground metrics to OT distances. We provide criteria to ensure the existence and uniqueness of these singular vectors. We then introduce scalable computational methods to approximate them in high-dimensional settings, using stochastic approximation and entropic regularization. Finally, we showcase Wasserstein Singular Vectors on a single-cell RNA-sequencing dataset.

Robust Kernel Density Estimation with Median-of-Means principle

- Pierre Humbert, Batiste Le Bars, Ludovic Minvielle
- abstract: In this paper, we introduce a robust non-parametric density estimator combining the popular Kernel Density Estimation method and the Median-of-Means principle (MoM-KDE). This estimator is shown to achieve robustness for a large class of anomalous data, potentially adversarial. In particular, while previous works only prove consistency results under very specific contamination models, this work provides finite-sample high-probability error-bounds without any prior knowledge on the outliers. To highlight the robustness of our method, we introduce an influence function adapted to the considered OUI framework. Finally, we show that MoM-KDE achieves competitive results when compared with other robust kernel estimators, while having significantly lower computational complexity.

A data-driven approach for learning to control computers

- Peter C Humphreys, David Raposo, Tobias Pohlen, Gregory Thornton, Rachita Chhaparia, Alistair Muldal, Josh Abramson, Petko Georgiev, Adam Santoro, Timothy Lillicrap
- abstract: It would be useful for machines to use computers as humans do so that they can aid us in everyday tasks. This is a setting in which there is also the potential to leverage large-scale expert demonstrations and human judgements of interactive behaviour, which are two ingredients that have driven much recent success in AI. Here we investigate the setting of computer control using keyboard and mouse, with goals specified via natural language. Instead of focusing on hand-designed curricula and specialized action spaces, we focus on developing a scalable method centered on reinforcement learning combined with behavioural priors informed by actual human-computer interactions. We achieve state-of-the-art and human-level mean performance across all tasks within the MiniWob++ benchmark, a challenging suite of computer control problems, and find strong evidence of cross-task transfer. These results demonstrate the usefulness of a unified human-agent interface when training machines to use computers. Altogether our results suggest a formula for achieving competency beyond MiniWob++ and towards controlling computers, in general, as a human would.

Proximal Denoiser for Convergent Plug-and-Play Optimization with Nonconvex Regularization

- Samuel Hurault, Arthur Leclaire, Nicolas Papadakis
- abstract: Plug-and-Play (PnP) methods solve ill-posed inverse problems through iterative proximal algorithms by replacing a proximal operator by a denoising operation. When applied with deep neural network denoisers, these methods have shown state-of-the-art visual performance for image restoration problems. However, their theoretical convergence analysis is still incomplete. Most of the existing convergence results consider nonexpansive denoisers, which is non-realistic, or limit their analysis to strongly convex data-fidelity terms in the inverse problem to solve. Recently, it was proposed to train the denoiser as a gradient descent step on a functional parameterized by a deep neural network. Using such a denoiser guarantees the convergence of the PnP version of the Half-Quadratic-Splitting (PnP-HQS) iterative algorithm. In this paper, we show that this gradient denoiser can actually correspond to the proximal operator of another scalar function. Given this new result, we exploit the convergence theory of proximal algorithms in the nonconvex setting to obtain convergence results for PnP-PGD (Proximal Gradient Descent) and PnP-ADMM (Alternating Direction Method of Multipliers). When built on top of a smooth gradient denoiser, we show that PnP-PGD and PnP-ADMM are convergent and target stationary points of an explicit functional. These convergence results are confirmed with numerical experiments on deblurring, super-resolution and inpainting.

Inverse Contextual Bandits: Learning How Behavior Evolves over Time

- Alihan Hüyük, Daniel Jarrett, Mihaela van der Schaar
- abstract: Understanding a decision-maker’s priorities by observing their behavior is critical for transparency and accountability in decision processes{—}such as in healthcare. Though conventional approaches to policy learning almost invariably assume stationarity in behavior, this is hardly true in practice: Medical practice is constantly evolving as clinical professionals fine-tune their knowledge over time. For instance, as the medical community’s understanding of organ transplantations has progressed over the years, a pertinent question is: How have actual organ allocation policies been evolving? To give an answer, we desire a policy learning method that provides interpretable representations of decision-making, in particular capturing an agent’s non-stationary knowledge of the world, as well as operating in an offline manner. First, we model the evolving behavior of decision-makers in terms of contextual bandits, and formalize the problem of Inverse Contextual Bandits ("ICB"). Second, we propose two concrete algorithms as solutions, learning parametric and non-parametric representations of an agent’s behavior. Finally, using both real and simulated data for liver transplantations, we illustrate the applicability and explainability of our method, as well as benchmarking and validating the accuracy of our algorithms.

Datamodels: Understanding Predictions with Data and Data with Predictions

- Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, Aleksander Madry
- abstract: We present a conceptual framework, datamodeling, for analyzing the behavior of a model class in terms of the training data. For any fixed “target” example x, training set S, and learning algorithm, a datamodel is a parameterized function $2^S \to \mathbb{R}$ that for any subset of $S’ \subset S$—using only information about which examples of S are contained in $S’$—predicts the outcome of training a model on $S’$ and evaluating on x. Despite the complexity of the underlying process being approximated (e.g. end-to-end training and evaluation of deep neural networks), we show that even simple linear datamodels successfully predict model outputs. We then demonstrate that datamodels give rise to a variety of applications, such as: accurately predicting the effect of dataset counterfactuals; identifying brittle predictions; finding semantically similar examples; quantifying train-test leakage; and embedding data into a well-behaved and feature-rich representation space.

Parsimonious Learning-Augmented Caching

- Sungjin Im, Ravi Kumar, Aditya Petety, Manish Purohit
- abstract: Learning-augmented algorithms—in which, traditional algorithms are augmented with machine-learned predictions—have emerged as a framework to go beyond worst-case analysis. The overarching goal is to design algorithms that perform near-optimally when the predictions are accurate yet retain certain worst-case guarantees irrespective of the accuracy of the predictions. This framework has been successfully applied to online problems such as caching where the predictions can be used to alleviate uncertainties. In this paper we introduce and study the setting in which the learning-augmented algorithm can utilize the predictions parsimoniously. We consider the caching problem—which has been extensively studied in the learning-augmented setting—and show that one can achieve quantitatively similar results but only using a sublinear number of predictions.

Bayesian Optimization for Distributionally Robust Chance-constrained Problem

- Yu Inatsu, Shion Takeno, Masayuki Karasuyama, Ichiro Takeuchi
- abstract: In black-box function optimization, we need to consider not only controllable design variables but also uncontrollable stochastic environment variables. In such cases, it is necessary to solve the optimization problem by taking into account the uncertainty of the environmental variables. Chance-constrained (CC) problem, the problem of maximizing the expected value under a certain level of constraint satisfaction probability, is one of the practically important problems in the presence of environmental variables. In this study, we consider distributionally robust CC (DRCC) problem and propose a novel DRCC Bayesian optimization method for the case where the distribution of the environmental variables cannot be precisely specified. We show that the proposed method can find an arbitrary accurate solution with high probability in a finite number of trials, and confirm the usefulness of the proposed method through numerical experiments.

LeNSE: Learning To Navigate Subgraph Embeddings for Large-Scale Combinatorial Optimisation

- David Ireland, Giovanni Montana
- abstract: Combinatorial Optimisation problems arise in several application domains and are often formulated in terms of graphs. Many of these problems are NP-hard, but exact solutions are not always needed. Several heuristics have been developed to provide near-optimal solutions; however, they do not typically scale well with the size of the graph. We propose a low-complexity approach for identifying a (possibly much smaller) subgraph of the original graph where the heuristics can be run in reasonable time and with a high likelihood of finding a global near-optimal solution. The core component of our approach is LeNSE, a reinforcement learning algorithm that learns how to navigate the space of possible subgraphs using an Euclidean subgraph embedding as its map. To solve CO problems, LeNSE is provided with a discriminative embedding trained using any existing heuristics using only on a small portion of the original graph. When tested on three problems (vertex cover, max-cut and influence maximisation) using real graphs with up to 10 million edges, LeNSE identifies small subgraphs yielding solutions comparable to those found by running the heuristics on the entire graph, but at a fraction of the total run time. Code for the experiments is available in the public GitHub repo at https://github.com/davidireland3/LeNSE.

The Dual Form of Neural Networks Revisited: Connecting Test Time Predictions to Training Patterns via Spotlights of Attention

- Kazuki Irie, Róbert Csordás, Jürgen Schmidhuber
- abstract: Linear layers in neural networks (NNs) trained by gradient descent can be expressed as a key-value memory system which stores all training datapoints and the initial weights, and produces outputs using unnormalised dot attention over the entire training experience. While this has been technically known since the 1960s, no prior work has effectively studied the operations of NNs in such a form, presumably due to prohibitive time and space complexities and impractical model sizes, all of them growing linearly with the number of training patterns which may get very large. However, this dual formulation offers a possibility of directly visualising how an NN makes use of training patterns at test time, by examining the corresponding attention weights. We conduct experiments on small scale supervised image classification tasks in single-task, multi-task, and continual learning settings, as well as language modelling, and discuss potentials and limits of this view for better understanding and interpreting how NNs exploit training patterns. Our code is public.

A Modern Self-Referential Weight Matrix That Learns to Modify Itself

- Kazuki Irie, Imanol Schlag, Róbert Csordás, Jürgen Schmidhuber
- abstract: The weight matrix (WM) of a neural network (NN) is its program. The programs of many traditional NNs are learned through gradient descent in some error function, then remain fixed. The WM of a self-referential NN, however, can keep rapidly modifying all of itself during runtime. In principle, such NNs can meta-learn to learn, and meta-meta-learn to meta-learn to learn, and so on, in the sense of recursive self-improvement. While NN architectures potentially capable of implementing such behaviour have been proposed since the ’90s, there have been few if any practical studies. Here we revisit such NNs, building upon recent successes of fast weight programmers and closely related linear Transformers. We propose a scalable self-referential WM (SRWM) that learns to use outer products and the delta update rule to modify itself. We evaluate our SRWM in supervised few-shot learning and in multi-task reinforcement learning with procedurally generated game environments. Our experiments demonstrate both practical applicability and competitive performance of the proposed SRWM. Our code is public.

Revisiting Online Submodular Minimization: Gap-Dependent Regret Bounds, Best of Both Worlds and Adversarial Robustness

- Shinji Ito
- abstract: In this paper, we consider online decision problems with submodular loss functions. For such problems, existing studies have only dealt with worst-case analysis. This study goes beyond worst-case analysis to show instance-dependent regret bounds. More precisely, for each of the full-information and bandit-feedback settings, we propose an algorithm that achieves a gap-dependent O(log T)-regret bound in the stochastic environment and is comparable to the best existing algorithm in the adversarial environment. The proposed algorithms also work well in the stochastic environment with adversarial corruptions, which is an intermediate setting between the stochastic and adversarial environments.

Modeling Strong and Human-Like Gameplay with KL-Regularized Search

- Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, Noam Brown
- abstract: We consider the task of accurately modeling strong human policies in multi-agent decision-making problems, given examples of human behavior. Imitation learning is effective at predicting human actions but may not match the strength of expert humans (e.g., by sometimes committing blunders), while self-play learning and search techniques such as AlphaZero lead to strong performance but may produce policies that differ markedly from human behavior. In chess and Go, we show that regularized search algorithms that penalize KL divergence from an imitation-learned policy yield higher prediction accuracy of strong humans and better performance than imitation learning alone. We then introduce a novel regret minimization algorithm that is regularized based on the KL divergence from an imitation-learned policy, and show that using this algorithm for search in no-press Diplomacy yields a policy that matches the human prediction accuracy of imitation learning while being substantially stronger.

A deep convolutional neural network that is invariant to time rescaling

- Brandon G Jacques, Zoran Tiganj, Aakash Sarkar, Marc Howard, Per Sederberg
- abstract: Human learners can readily understand speech, or a melody, when it is presented slower or faster than usual. This paper presents a deep CNN (SITHCon) that uses a logarithmically compressed temporal representation at each level. Because rescaling the time of the input results in a translation of $\log$ time, and because the output of the convolution is invariant to translations, this network can generalize to out-of-sample data that are temporal rescalings of a learned pattern. We compare the performance of SITHCon to a Temporal Convolution Network (TCN) on classification and regression problems with both univariate and multivariate time series. We find that SITHCon, unlike TCN, generalizes robustly over rescalings of about an order of magnitude. Moreover, we show that the network can generalize over exponentially large scales without retraining the weights simply by extending the range of the logarithmically-compressed temporal memory.

Input Dependent Sparse Gaussian Processes

- Bahram Jafrasteh, Carlos Villacampa-Calvo, Daniel Hernandez-Lobato
- abstract: Gaussian Processes (GPs) are non-parametric models that provide accurate uncertainty estimates. Nevertheless, they have a cubic cost in the number of data instances N. To overcome this, sparse GP approximations are used, in which a set of $M \ll N$ inducing points is introduced. The location of the inducing points is learned by considering them parameters of an approximate posterior distribution q. Sparse GPs, combined with stochastic variational inference for inferring q have a cost per iteration in $\mathcal{O}(M^3)$. Critically, the inducing points determine the flexibility of the model and they are often located in regions where the latent function changes. A limitation is, however, that in some tasks a large number of inducing points may be required to obtain good results. To alleviate this, we propose here to amortize the computation of the inducing points locations, as well as the parameters of q. For this, we use a neural network that receives a data instance as an input and outputs the corresponding inducing points locations and the parameters of q. We evaluate our method in several experiments, showing that it performs similar or better than other state-of-the-art sparse variational GPs. However, in our method the number of inducing points is reduced drastically since they depend on the input data. This makes our method scale to larger datasets and have faster training and prediction times.

Regret Minimization with Performative Feedback

- Meena Jagadeesan, Tijana Zrnic, Celestine Mendler-Dünner
- abstract: In performative prediction, the deployment of a predictive model triggers a shift in the data distribution. As these shifts are typically unknown ahead of time, the learner needs to deploy a model to get feedback about the distribution it induces. We study the problem of finding near-optimal models under performativity while maintaining low regret. On the surface, this problem might seem equivalent to a bandit problem. However, it exhibits a fundamentally richer feedback structure that we refer to as performative feedback: after every deployment, the learner receives samples from the shifted distribution rather than bandit feedback about the reward. Our main contribution is regret bounds that scale only with the complexity of the distribution shifts and not that of the reward function. The key algorithmic idea is careful exploration of the distribution shifts that informs a novel construction of confidence bounds on the risk of unexplored models. The construction only relies on smoothness of the shifts and does not assume convexity. More broadly, our work establishes a conceptual approach for leveraging tools from the bandits literature for the purpose of regret minimization with performative feedback.

Biological Sequence Design with GFlowNets

- Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, Yoshua Bengio
- abstract: Design of de novo biological sequences with desired properties, like protein and DNA sequences, often involves an active loop with several rounds of molecule ideation and expensive wet-lab evaluations. These experiments can consist of multiple stages, with increasing levels of precision and cost of evaluation, where candidates are filtered. This makes the diversity of proposed candidates a key consideration in the ideation phase. In this work, we propose an active learning algorithm leveraging epistemic uncertainty estimation and the recently proposed GFlowNets as a generator of diverse candidate solutions, with the objective to obtain a diverse batch of useful (as defined by some utility function, for example, the predicted anti-microbial activity of a peptide) and informative candidates after each round. We also propose a scheme to incorporate existing labeled datasets of candidates, in addition to a reward function, to speed up learning in GFlowNets. We present empirical results on several biological sequence design tasks, and we find that our method generates more diverse and novel batches with high scoring candidates compared to existing approaches.

Combining Diverse Feature Priors

- Saachi Jain, Dimitris Tsipras, Aleksander Madry
- abstract: To improve model generalization, model designers often restrict the features that their models use, either implicitly or explicitly. In this work, we explore the design space of leveraging such feature priors by viewing them as distinct perspectives on the data. Specifically, we find that models trained with diverse sets of explicit feature priors have less overlapping failure modes, and can thus be combined more effectively. Moreover, we demonstrate that jointly training such models on additional (unlabeled) data allows them to correct each other’s mistakes, which, in turn, leads to better generalization and resilience to spurious correlations.

Training Your Sparse Neural Network Better with Any Mask

- Ajay Kumar Jaiswal, Haoyu Ma, Tianlong Chen, Ying Ding, Zhangyang Wang
- abstract: Pruning large neural networks to create high-quality, independently trainable sparse masks, which can maintain similar performance to their dense counterparts, is very desirable due to the reduced space and time complexity. As research effort is focused on increasingly sophisticated pruning methods that leads to sparse subnetworks trainable from the scratch, we argue for an orthogonal, under-explored theme: improving training techniques for pruned sub-networks, i.e. sparse training. Apart from the popular belief that only the quality of sparse masks matters for sparse training, in this paper we demonstrate an alternative opportunity: one can carefully customize the sparse training techniques to deviate from the default dense network training protocols, consisting of introducing “ghost" neurons and skip connections at the early stage of training, and strategically modifying the initialization as well as labels. Our new sparse training recipe is generally applicable to improving training from scratch with various sparse masks. By adopting our newly curated techniques, we demonstrate significant performance gains across various popular datasets (CIFAR-10, CIFAR-100, TinyImageNet), architectures (ResNet-18/32/104, Vgg16, MobileNet), and sparse mask options (lottery ticket, SNIP/GRASP, SynFlow, or even randomly pruning), compared to the default training protocols, especially at high sparsity levels. Codes will be publicly available.

Sequential Covariate Shift Detection Using Classifier Two-Sample Tests

- Sooyong Jang, Sangdon Park, Insup Lee, Osbert Bastani
- abstract: A standard assumption in supervised learning is that the training data and test data are from the same distribution. However, this assumption often fails to hold in practice, which can cause the learned model to perform poorly. We consider the problem of detecting covariate shift, where the covariate distribution shifts but the conditional distribution of labels given covariates remains the same. This problem can naturally be solved using a two-sample test{—}i.e., test whether the current test distribution of covariates equals the training distribution of covariates. Our algorithm builds on classifier tests, which train a discriminator to distinguish train and test covariates, and then use the accuracy of this discriminator as a test statistic. A key challenge is that classifier tests assume given a fixed set of test covariates. In practice, test covariates often arrive sequentially over time{—}e.g., a self-driving car observes a stream of images while driving. Furthermore, covariate shift can occur multiple times{—}i.e., shift and then shift back later or gradually shift over time. To address these challenges, our algorithm trains the discriminator online. Additionally, it evaluates test accuracy using each new covariate before taking a gradient step; this strategy avoids constructing a held-out test set, which can improve sample efficiency. We prove that this optimization preserves the correctness{—}i.e., our algorithm achieves a desired bound on the false positive rate. In our experiments, we show that our algorithm efficiently detects covariate shifts on multiple datasets{—}ImageNet, IWildCam, and Py150.

Surrogate Likelihoods for Variational Annealed Importance Sampling

- Martin Jankowiak, Du Phan
- abstract: Variational inference is a powerful paradigm for approximate Bayesian inference with a number of appealing properties, including support for model learning and data subsampling. By contrast MCMC methods like Hamiltonian Monte Carlo do not share these properties but remain attractive since, contrary to parametric methods, MCMC is asymptotically unbiased. For these reasons researchers have sought to combine the strengths of both classes of algorithms, with recent approaches coming closer to realizing this vision in practice. However, supporting data subsampling in these hybrid methods can be a challenge, a shortcoming that we address by introducing a surrogate likelihood that can be learned jointly with other variational parameters. We argue theoretically that the resulting algorithm allows an intuitive trade-off between inference fidelity and computational cost. In an extensive empirical comparison we show that our method performs well in practice and that it is well-suited for black-box inference in probabilistic programming frameworks.

Planning with Diffusion for Flexible Behavior Synthesis

- Michael Janner, Yilun Du, Joshua Tenenbaum, Sergey Levine
- abstract: Model-based reinforcement learning methods often use learning only for the purpose of recovering an approximate dynamics model, offloading the rest of the decision-making work to classical trajectory optimizers. While conceptually simple, this combination has a number of empirical shortcomings, suggesting that learned models may not be well-suited to standard trajectory optimization. In this paper, we consider what it would look like to fold as much of the trajectory optimization pipeline as possible into the modeling problem, such that sampling from the model and planning with it become nearly identical. The core of our technical approach lies in a diffusion probabilistic model that plans by iteratively denoising trajectories. We show how classifier-guided sampling and image inpainting can be reinterpreted as coherent planning strategies, explore the unusual and useful properties of diffusion-based planning methods, and demonstrate the effectiveness of our framework in control settings that emphasize long-horizon decision-making and test-time flexibility.

HyperImpute: Generalized Iterative Imputation with Automatic Model Selection

- Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, Mihaela van der Schaar
- abstract: Consider the problem of imputing missing values in a dataset. One the one hand, conventional approaches using iterative imputation benefit from the simplicity and customizability of learning conditional distributions directly, but suffer from the practical requirement for appropriate model specification of each and every variable. On the other hand, recent methods using deep generative modeling benefit from the capacity and efficiency of

learning with neural network function approximators, but are often difficult to optimize and rely on stronger data assumptions. In this work, we study an approach that marries the advantages of both: We propose *HyperImpute*, a generalized iterative imputation framework for adaptively and automatically configuring column-wise models and their hyperparameters. Practically, we provide a concrete implementation with out-of-the-box learners, optimizers, simulators, and extensible interfaces. Empirically, we investigate this framework via comprehensive experiments and sensitivities on a variety of public datasets, and demonstrate its ability to generate accurate imputations relative to a strong suite of benchmarks. Contrary to recent work, we believe our findings constitute a strong defense of the iterative imputation paradigm.

Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization

- Adrian Javaloy, Maryam Meghdadi, Isabel Valera
- abstract: A number of variational autoencoders (VAEs) have recently emerged with the aim of modeling multimodal data, e.g., to jointly model images and their corresponding captions. Still, multimodal VAEs tend to focus solely on a subset of the modalities, e.g., by fitting the image while neglecting the caption. We refer to this limitation as modality collapse. In this work, we argue that this effect is a consequence of conflicting gradients during multimodal VAE training. We show how to detect the sub-graphs in the computational graphs where gradients conflict (impartiality blocks), as well as how to leverage existing gradient-conflict solutions from multitask learning to mitigate modality collapse. That is, to ensure impartial optimization across modalities. We apply our training framework to several multimodal VAE models, losses and datasets from the literature, and empirically show that our framework significantly improves the reconstruction performance, conditional generation, and coherence of the latent space across modalities.

Towards understanding how momentum improves generalization in deep learning

- Samy Jelassi, Yuanzhi Li
- abstract: Stochastic gradient descent (SGD) with momentum is widely used for training modern deep learning architectures. While it is well-understood that using momentum can lead to faster convergence rate in various settings, it has also been observed that momentum yields higher generalization. Prior work argue that momentum stabilizes the SGD noise during training and this leads to higher generalization. In this paper, we adopt another perspective and first empirically show that gradient descent with momentum (GD+M) significantly improves generalization compared to gradient descent (GD) in some deep learning problems. From this observation, we formally study how momentum improves generalization. We devise a binary classification setting where a one-hidden layer (over-parameterized) convolutional neural network trained with GD+M provably generalizes better than the same network trained with GD, when both algorithms are similarly initialized. The key insight in our analysis is that momentum is beneficial in datasets where the examples share some feature but differ in their margin. Contrary to GD that memorizes the small margin data, GD+M still learns the feature in these data thanks to its historical gradients. Lastly, we empirically validate our theoretical findings.

MASER: Multi-Agent Reinforcement Learning with Subgoals Generated from Experience Replay Buffer

- Jeewon Jeon, Woojun Kim, Whiyoung Jung, Youngchul Sung
- abstract: In this paper, we consider cooperative multi-agent reinforcement learning (MARL) with sparse reward. To tackle this problem, we propose a novel method named MASER: MARL with subgoals generated from experience replay buffer. Under the widely-used assumption of centralized training with decentralized execution and consistent Q-value decomposition for MARL, MASER automatically generates proper subgoals for multiple agents from the experience replay buffer by considering both individual Q-value and total Q-value. Then, MASER designs individual intrinsic reward for each agent based on actionable representation relevant to Q-learning so that the agents reach their subgoals while maximizing the joint action value. Numerical results show that MASER significantly outperforms StarCraft II micromanagement benchmark compared to other state-of-the-art MARL algorithms.

An Exact Symbolic Reduction of Linear Smart Predict+Optimize to Mixed Integer Linear Programming

- Jihwan Jeong, Parth Jaggi, Andrew Butler, Scott Sanner
- abstract: Predictive models are traditionally optimized independently of their use in downstream decision-based optimization. The ‘smart, predict then optimize’ (SPO) framework addresses this shortcoming by optimizing predictive models in order to minimize the final downstream decision loss. To date, several local first-order methods and convex approximations have been proposed. These methods have proven to be effective in practice, however, it remains generally unclear as to how close these local solutions are to global optimality. In this paper, we cast the SPO problem as a bi-level program and apply Symbolic Variable Elimination (SVE) to analytically solve the lower optimization. The resulting program can then be formulated as a mixed-integer linear program (MILP) which is solved to global optimality using standard off-the-shelf solvers. To our knowledge, our framework is the first to provide a globally optimal solution to the linear SPO problem. Experimental results comparing with state-of-the-art local SPO solvers show that the globally optimal solution obtains up to two orders of magnitude reduction in decision regret.

Agnostic Learnability of Halfspaces via Logistic Loss

- Ziwei Ji, Kwangjun Ahn, Pranjal Awasthi, Satyen Kale, Stefani Karp
- abstract: We investigate approximation guarantees provided by logistic regression for the fundamental problem of agnostic learning of homogeneous halfspaces. Previously, for a certain broad class of “well-behaved” distributions on the examples, Diakonikolas et al. (2020) proved an tilde{Omega}(OPT) lower bound, while Frei et al. (2021) proved an tilde{O}(sqrt{OPT}) upper bound, where OPT denotes the best zero-one/misclassification risk of a homogeneous halfspace. In this paper, we close this gap by constructing a well-behaved distribution such that the global minimizer of the logistic risk over this distribution only achieves Omega(sqrt{OPT}) misclassification risk, matching the upper bound in (Frei et al., 2021). On the other hand, we also show that if we impose a radial-Lipschitzness condition in addition to well-behaved-ness on the distribution, logistic regression on a ball of bounded radius reaches tilde{O}(OPT) misclassification risk. Our techniques also show for any well-behaved distribution, regardless of radial Lipschitzness, we can overcome the Omega(sqrt{OPT}) lower bound for logistic loss simply at the cost of one additional convex optimization step involving the hinge loss and attain tilde{O}(OPT) misclassification risk. This two-step convex optimization algorithm is simpler than previous methods obtaining this guarantee, all of which require solving O(log(1/OPT)) minimization problems.

Improving Policy Optimization with Generalist-Specialist Learning

- Zhiwei Jia, Xuanlin Li, Zhan Ling, Shuang Liu, Yiran Wu, Hao Su
- abstract: Generalization in deep reinforcement learning over unseen environment variations usually requires policy learning over a large set of diverse training variations. We empirically observe that an agent trained on many variations (a generalist) tends to learn faster at the beginning, yet its performance plateaus at a less optimal level for a long time. In contrast, an agent trained only on a few variations (a specialist) can often achieve high returns under a limited computational budget. To have the best of both worlds, we propose a novel generalist-specialist training framework. Specifically, we first train a generalist on all environment variations; when it fails to improve, we launch a large population of specialists with weights cloned from the generalist, each trained to master a selected small subset of variations. We finally resume the training of the generalist with auxiliary rewards induced by demonstrations of all specialists. In particular, we investigate the timing to start specialist training and compare strategies to learn generalists with assistance from specialists. We show that this framework pushes the envelope of policy learning on several challenging and popular benchmarks including Procgen, Meta-World and ManiSkill.

Translatotron 2: High-quality direct speech-to-speech translation with voice preservation

- Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, Roi Pomerantz
- abstract: We present Translatotron 2, a neural direct speech-to-speech translation model that can be trained end-to-end. Translatotron 2 consists of a speech encoder, a linguistic decoder, an acoustic synthesizer, and a single attention module that connects them together. Experimental results on three datasets consistently show that Translatotron 2 outperforms the original Translatotron by a large margin on both translation quality (up to +15.5 BLEU) and speech generation quality, and approaches the same of cascade systems. In addition, we propose a simple method for preserving speakers' voices from the source speech to the translation speech in a different language. Unlike existing approaches, the proposed method is able to preserve each speaker's voice on speaker turns without requiring for speaker segmentation. Furthermore, compared to existing approaches, it better preserves speaker's privacy and mitigates potential misuse of voice cloning for creating spoofing audio artifacts.

Online Learning and Pricing with Reusable Resources: Linear Bandits with Sub-Exponential Rewards

- Huiwen Jia, Cong Shi, Siqian Shen
- abstract: We consider a price-based revenue management problem with reusable resources over a finite time horizon T. The problem finds important applications in car/bicycle rental, ridesharing, cloud computing, and hospitality management. Customers arrive following a price-dependent Poisson process and each customer requests one unit of c homogeneous reusable resources. If there is an available unit, the customer gets served within a price-dependent exponentially distributed service time; otherwise, she waits in a queue until the next available unit. The decision maker assumes that the inter-arrival and service intervals have an unknown linear dependence on a d_f-dimensional feature vector associated with the posted price. We propose a rate-optimal online learning and pricing algorithm, termed Batch Linear Confidence Bound (BLinUCB), and prove that the cumulative regret is $\tilde{O} (d_f \sqrt{T })$. In establishing the regret, we bound the transient system performance upon price changes via a coupling argument, and also generalize linear bandits to accommodate sub-exponential rewards.

The Role of Deconfounding in Meta-learning

- Yinjie Jiang, Zhengyu Chen, Kun Kuang, Luotian Yuan, Xinhai Ye, Zhihua Wang, Fei Wu, Ying Wei
- abstract: Meta-learning has emerged as a potent paradigm for quick learning of few-shot tasks, by leveraging the meta-knowledge learned from meta-training tasks. Well-generalized meta-knowledge that facilitates fast adaptation in each task is preferred; however, recent evidence suggests the undesirable memorization effect where the meta-knowledge simply memorizing all meta-training tasks discourages task-specific adaptation and poorly generalizes. There have been several solutions to mitigating the effect, including both regularizer-based and augmentation-based methods, while a systematic understanding of these methods in a single framework is still lacking. In this paper, we offer a novel causal perspective of meta-learning. Through the lens of causality, we conclude the universal label space as a confounder to be the causing factor of memorization and frame the two lines of prevailing methods as different deconfounder approaches. Remarkably, derived from the causal inference principle of front-door adjustment, we propose two frustratingly easy but effective deconfounder algorithms, i.e., sampling multiple versions of the meta-knowledge via Dropout and grouping the meta-knowledge into multiple bins. The proposed causal perspective not only brings in the two deconfounder algorithms that surpass previous works in four benchmark datasets towards combating memorization, but also opens a promising direction for meta-learning.

Subspace Learning for Effective Meta-Learning

- Weisen Jiang, James Kwok, Yu Zhang
- abstract: Meta-learning aims to extract meta-knowledge from historical tasks to accelerate learning on new tasks. Typical meta-learning algorithms like MAML learn a globally-shared meta-model for all tasks. However, when the task environments are complex, task model parameters are diverse and a common meta-model is insufficient to capture all the meta-knowledge. To address this challenge, in this paper, task model parameters are structured into multiple subspaces, and each subspace represents one type of meta-knowledge. We propose an algorithm to learn the meta-parameters (\ie, subspace bases). We theoretically study the generalization properties of the learned subspaces. Experiments on regression and classification meta-learning datasets verify the effectiveness of the proposed algorithm.

Optimal Algorithms for Stochastic Multi-Level Compositional Optimization

- Wei Jiang, Bokun Wang, Yibo Wang, Lijun Zhang, Tianbao Yang
- abstract: In this paper, we investigate the problem of stochastic multi-level compositional optimization, where the objective function is a composition of multiple smooth but possibly non-convex functions. Existing methods for solving this problem either suffer from sub-optimal sample complexities or need a huge batch size. To address this limitation, we propose a Stochastic Multi-level Variance Reduction method (SMVR), which achieves the optimal sample complexity of $\mathcal{O}\left(1 / \epsilon^{3}\right)$ to find an ϵ-stationary point for non-convex objectives. Furthermore, when the objective function satisfies the convexity or Polyak-{Ł}ojasiewicz (PL) condition, we propose a stage-wise variant of SMVR and improve the sample complexity to $\mathcal{O}\left(1 / \epsilon^{2}\right)$ for convex functions or $\mathcal{O}\left(1 /(\mu\epsilon)\right)$ for non-convex functions satisfying the μ-PL condition. The latter result implies the same complexity for μ-strongly convex functions. To make use of adaptive learning rates, we also develop Adaptive SMVR, which achieves the same optimal complexities but converges faster in practice. All our complexities match the lower bounds not only in terms of ϵ but also in terms of μ (for PL or strongly convex functions), without using a large batch size in each iteration.

Antibody-Antigen Docking and Design via Hierarchical Structure Refinement

- Wengong Jin, Dr.Regina Barzilay, Tommi Jaakkola
- abstract: Computational antibody design seeks to automatically create an antibody that binds to an antigen. The binding affinity is governed by the 3D binding interface where antibody residues (paratope) closely interact with antigen residues (epitope). Thus, the key question of antibody design is how to predict the 3D paratope-epitope complex (i.e., docking) for paratope generation. In this paper, we propose a new model called Hierarchical Structure Refinement Network (HSRN) for paratope docking and design. During docking, HSRN employs a hierarchical message passing network to predict atomic forces and use them to refine a binding complex in an iterative, equivariant manner. During generation, its autoregressive decoder progressively docks generated paratopes and builds a geometric representation of the binding interface to guide the next residue choice. Our results show that HSRN significantly outperforms prior state-of-the-art on paratope docking and design benchmarks.

Sharpened Quasi-Newton Methods: Faster Superlinear Rate and Larger Local Convergence Neighborhood

- Qiujiang Jin, Alec Koppel, Ketan Rajawat, Aryan Mokhtari
- abstract: Non-asymptotic analysis of quasi-Newton methods have received a lot of attention recently. In particular, several works have established a non-asymptotic superlinear rate of $$\mathcal{O}((1/\sqrt{t})^t)$$ for the (classic) BFGS method by exploiting the fact that its error of Newton direction approximation approaches zero. Moreover, a greedy variant of the BFGS method was recently proposed which accelerates the convergence of BFGS by directly approximating the Hessian matrix, instead of Newton direction, and achieves a fast local quadratic convergence rate. Alas, the local quadratic convergence of Greedy-BFGS requires way more updates compared to the number of iterations that BFGS requires for a local superlinear rate. This is due to the fact that in Greedy-BFGS the Hessian is directly approximated and the Newton direction approximation may not be as accurate as the one for BFGS. In this paper, we close this gap and present a novel BFGS method that has the best of two worlds. More precisely, it leverages the approximation ideas of both BFGS and Greedy-BFGS to properly approximate both the Newton direction and the Hessian matrix. Our theoretical results show that our

method out-performs both BFGS and Greedy-BFGS in terms of convergence rate, while it reaches its quadratic convergence rate with fewer steps compared to Greedy-BFGS. Numerical experiments on various datasets also confirm our theoretical findings.

The Power of Exploiter: Provable Multi-Agent RL in Large State Spaces

- Chi Jin, Qinghua Liu, Tiancheng Yu
- abstract: Modern reinforcement learning (RL) commonly engages practical problems with large state spaces, where function approximation must be deployed to approximate either the value function or the policy. While recent progresses in RL theory address a rich set of RL problems with general function approximation, such successes are mostly restricted to the single-agent setting. It remains elusive how to extend these results to multi-agent RL, especially in the face of new game-theoretical challenges. This paper considers two-player zero-sum Markov Games (MGs). We propose a new algorithm that can provably find the Nash equilibrium policy using a polynomial number of samples, for any MG with low multi-agent Bellman-Eluder dimension—a new complexity measure adapted from its single-agent version (Jin et al., 2021). A key component of our new algorithm is the exploiter, which facilitates the learning of the main player by deliberately exploiting her weakness. Our theoretical framework is generic, which applies to a wide range of models including but not limited to tabular MGs, MGs with linear or kernel function approximation, and MGs with rich observations.

Domain Adaptation for Time Series Forecasting via Attention Sharing

- Xiaoyong Jin, Youngsuk Park, Danielle Maddix, Hao Wang, Yuyang Wang
- abstract: Recently, deep neural networks have gained increasing popularity in the field of time series forecasting. A primary reason for their success is their ability to effectively capture complex temporal dynamics across multiple related time series. The advantages of these deep forecasters only start to emerge in the presence of a sufficient amount of data. This poses a challenge for typical forecasting problems in practice, where there is a limited number of time series or observations per time series, or both. To cope with this data scarcity issue, we propose a novel domain adaptation framework, Domain Adaptation Forecaster (DAF). DAF leverages statistical strengths from a relevant domain with abundant data samples (source) to improve the performance on the domain of interest with limited data (target). In particular, we use an attention-based shared module with a domain discriminator across domains and private modules for individual domains. We induce domain-invariant latent features (queries and keys) and retrain domain-specific features (values) simultaneously to enable joint training of forecasters on source and target domains. A main insight is that our design of aligning keys allows the target domain to leverage source time series even with different characteristics. Extensive experiments on various domains demonstrate that our proposed method outperforms state-of-the-art baselines on synthetic and real-world datasets, and ablation studies verify the effectiveness of our design choices.

Accelerated Federated Learning with Decoupled Adaptive Optimization

- Jiayin Jin, Jiaxiang Ren, Yang Zhou, Lingjuan Lyu, Ji Liu, Dejing Dou
- abstract: The federated learning (FL) framework enables edge clients to collaboratively learn a shared inference model while keeping privacy of training data on clients. Recently, many heuristics efforts have been made to generalize centralized adaptive optimization methods, such as SGDM, Adam, AdaGrad, etc., to federated settings for improving convergence and accuracy. However, there is still a paucity of theoretical principles on where to and how to design and utilize adaptive optimization methods in federated settings. This work aims to develop novel adaptive optimization methods for FL from the perspective of dynamics of ordinary differential equations (ODEs). First, an analytic framework is established to build a connection between federated optimization methods and decompositions of ODEs of corresponding centralized optimizers. Second, based on this analytic framework, a momentum decoupling adaptive optimization method, FedDA, is developed to fully utilize the global momentum on each local iteration and accelerate the training convergence. Last but not least, full batch gradients are utilized to mimic centralized optimization in the end of the training process to ensure the convergence and overcome the possible inconsistency caused by adaptive optimization methods.

Supervised Off-Policy Ranking

- Yue Jin, Yue Zhang, Tao Qin, Xudong Zhang, Jian Yuan, Houqiang Li, Tie-Yan Liu
- abstract: Off-policy evaluation (OPE) is to evaluate a target policy with data generated by other policies. Most previous OPE methods focus on precisely estimating the true performance of a policy. We observe that in many applications, (1) the end goal of OPE is to compare two or multiple candidate policies and choose a good one, which is a much simpler task than precisely evaluating their true performance; and (2) there are usually multiple policies that have been deployed to serve users in real-world systems and thus the true performance of these policies can be known. Inspired by the two observations, in this work, we study a new problem, supervised off-policy ranking (SOPR), which aims to rank a set of target policies based on supervised learning by leveraging off-policy data and policies with known performance. We propose a method to solve SOPR, which learns a policy scoring model by minimizing a ranking loss of the training policies rather than estimating the precise policy performance. The scoring model in our method, a hierarchical Transformer based model, maps a set of state-action pairs to a score, where the state of each pair comes from the off-policy data and the action is taken by a target policy on the state in an offline manner. Extensive experiments on public datasets show that our method outperforms baseline methods in terms of rank correlation, regret value, and stability. Our code is publicly available at GitHub.

Input-agnostic Certified Group Fairness via Gaussian Parameter Smoothing

- Jiayin Jin, Zeru Zhang, Yang Zhou, Lingfei Wu
- abstract: Only recently, researchers attempt to provide classification algorithms with provable group fairness guarantees. Most of these algorithms suffer from harassment caused by the requirement that the training and deployment data follow the same distribution. This paper proposes an input-agnostic certified group fairness algorithm, FairSmooth, for improving the fairness of classification models while maintaining the remarkable prediction accuracy. A Gaussian parameter smoothing method is developed to transform base classifiers into their smooth versions. An optimal individual smooth classifier is learnt for each group with only the data regarding the group and an overall smooth classifier for all groups is generated by averaging the parameters of all the individual smooth ones. By leveraging the theory of nonlinear functional analysis, the smooth classifiers are reformulated as output functions of a Nemytskii operator. Theoretical analysis is conducted to derive that the Nemytskii operator is smooth and induces a Frechet differentiable smooth manifold. We theoretically demonstrate that the smooth manifold has a global Lipschitz constant that is independent of the domain of the input data, which derives the input-agnostic certified group fairness.

Score-based Generative Modeling of Graphs via the System of Stochastic Differential Equations

- Jaehyeong Jo, Seul Lee, Sung Ju Hwang
- abstract: Generating graph-structured data requires learning the underlying distribution of graphs. Yet, this is a challenging problem, and the previous graph generative methods either fail to capture the permutation-invariance property of graphs or cannot sufficiently model the complex dependency between nodes and edges, which is crucial for generating real-world graphs such as molecules. To overcome such limitations, we propose a novel score-based generative model for graphs with a continuous-time framework. Specifically, we propose a new graph diffusion process that models the joint distribution of the nodes and edges through a system of stochastic differential equations (SDEs). Then, we derive novel score matching objectives tailored for the proposed diffusion process to estimate the gradient of the joint log-density with respect to each component, and introduce a new solver for the system of SDEs to efficiently sample from the reverse diffusion process. We validate our graph generation method on diverse datasets, on which it either achieves significantly superior or competitive performance to the baselines. Further analysis shows that our method is able to generate molecules that lie

close to the training distribution yet do not violate the chemical valency rule, demonstrating the effectiveness of the system of SDEs in modeling the node-edge relationships.

Choosing Answers in Epsilon-Best-Answer Identification for Linear Bandits

- Marc Jourdan, Rémy Degenne
- abstract: In pure-exploration problems, information is gathered sequentially to answer a question on the stochastic environment. While best-arm identification for linear bandits has been extensively studied in recent years, few works have been dedicated to identifying one arm that is ε-close to the best one (and not exactly the best one). In this problem with several correct answers, an identification algorithm should focus on one candidate among those answers and verify that it is correct. We demonstrate that picking the answer with highest mean does not allow an algorithm to reach asymptotic optimality in terms of expected sample complexity. Instead, a furthest answer should be identified. Using that insight to choose the candidate answer carefully, we develop a simple procedure to adapt best-arm identification algorithms to tackle ε-best-answer identification in transductive linear stochastic bandits. Finally, we propose an asymptotically optimal algorithm for this setting, which is shown to achieve competitive empirical performance against existing modified best-arm identification algorithms.

Robust Fine-Tuning of Deep Neural Networks with Hessian-based Generalization Guarantees

- Haotian Ju, Dongyue Li, Hongyang R Zhang
- abstract: We consider transfer learning approaches that fine-tune a pretrained deep neural network on a target task. We investigate generalization properties of fine-tuning to understand the problem of overfitting, which often happens in practice. Previous works have shown that constraining the distance from the initialization of fine-tuning improves generalization. Using a PAC-Bayesian analysis, we observe that besides distance from initialization, Hessians affect generalization through the noise stability of deep neural networks against noise injections. Motivated by the observation, we develop Hessian distance-based generalization bounds for a wide range of fine-tuning methods. Next, we investigate the robustness of fine-tuning with noisy labels. We design an algorithm that incorporates consistent losses and distance-based regularization for fine-tuning. Additionally, we prove a generalization error bound of our algorithm under class conditional independent noise in the training dataset labels. We perform a detailed empirical study of our algorithm on various noisy environments and architectures. For example, on six image classification tasks whose training labels are generated with programmatic labeling, we show a 3.26% accuracy improvement over prior methods. Meanwhile, the Hessian distance measure of the fine-tuned network using our algorithm decreases by six times more than existing approaches.

Robust alignment of cross-session recordings of neural population activity by behaviour via unsupervised domain adaptation

- Justin Jude, Matthew Perich, Lee Miller, Matthias Hennig
- abstract: Neural population activity relating to behaviour is assumed to be inherently low-dimensional despite the observed high dimensionality of data recorded using multi-electrode arrays. Therefore, predicting behaviour from neural population recordings has been shown to be most effective when using latent variable models. Over time however, the activity of single neurons can drift, and different neurons will be recorded due to movement of implanted neural probes. This means that a decoder trained to predict behaviour on one day performs worse when tested on a different day. On the other hand, evidence suggests that the latent dynamics underlying behaviour may be stable even over months and years. Based on this idea, we introduce a model capable of inferring behaviourally relevant latent dynamics from previously unseen data recorded from the same animal, without any need for decoder recalibration. We show that unsupervised domain adaptation combined with a sequential variational autoencoder, trained on several sessions, can achieve good generalisation to unseen data and correctly predict behaviour where conventional methods fail. Our results further support the hypothesis that behaviour-related neural dynamics are low-dimensional and stable over time, and will enable more effective and flexible use of brain computer interface technologies.

On Measuring Causal Contributions via do-interventions

- Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Bloebaum, Elias Bareinboim
- abstract: Causal contributions measure the strengths of different causes to a target quantity. Understanding causal contributions is important in empirical sciences and data-driven disciplines since it allows to answer practical queries like “what are the contributions of each cause to the effect?” In this paper, we develop a principled method for quantifying causal contributions. First, we provide desiderata of properties axioms that causal contribution measures should satisfy and propose the do-Shapley values (inspired by do-interventions [Pearl, 2000]) as a unique method satisfying these properties. Next, we develop a criterion under which the do-Shapley values can be efficiently inferred from non-experimental data. Finally, we provide do-Shapley estimators exhibiting consistency, computational feasibility, and statistical robustness. Simulation results corroborate with the theory.

Efficient Approximate Inference for Stationary Kernel on Frequency Domain

- Yohan Jung, Kyungwoo Song, Jinkyoo Park
- abstract: Based on the Fourier duality between a stationary kernel and its spectral density, modeling the spectral density using a Gaussian mixture density enables one to construct a flexible kernel, known as a Spectral Mixture kernel, that can model any stationary kernel. However, despite its expressive power, training this kernel is typically difficult because scalability and overfitting issues often arise due to a large number of training parameters. To resolve these issues, we propose an approximate inference method for estimating the Spectral mixture kernel hyperparameters. Specifically, we approximate this kernel by using the finite random spectral points based on Random Fourier Feature and optimize the parameters for the distribution of spectral points by sampling-based variational inference. To improve this inference procedure, we analyze the training loss and propose two special methods: a sampling method of spectral points to reduce the error of the approximate kernel in training, and an approximate natural gradient to accelerate the convergence of parameter inference.

Sketching Algorithms and Lower Bounds for Ridge Regression

- Praneeth Kacham, David Woodruff
- abstract: We give a sketching-based iterative algorithm that computes a $1+\varepsilon$ approximate solution for the ridge regression problem $\min_x \|Ax-b\|_2^2 +\lambda\|x\|_2^2$ where $A \in R^{n \times d}$ with $d \ge n$. Our algorithm, for a constant number of iterations (requiring a constant number of passes over the input), improves upon earlier work (Chowdhury et al.) by requiring that the sketching matrix only has a weaker Approximate Matrix Multiplication (AMM) guarantee that depends on ε, along with a constant subspace embedding guarantee. The earlier work instead requires that the sketching matrix has a subspace embedding guarantee that depends on ε. For example, to produce a $1+\varepsilon$ approximate solution in 1 iteration, which requires 2 passes over the input, our algorithm requires the OSNAP embedding to have $m= O(n\sigma^2/\lambda\varepsilon)$ rows with a sparsity parameter $s = O(\log(n))$, whereas the earlier algorithm of Chowdhury et al. with the same number of rows of OSNAP requires a sparsity $s = O(\sqrt{\sigma^2/\lambda\varepsilon} \cdot \log(n))$, where $\sigma = \opnorm{A}$ is the spectral norm of the matrix A. We also show that this algorithm can be used to give faster algorithms for kernel ridge regression. Finally, we show that the sketch size required for our algorithm is essentially optimal for a natural framework of algorithms for ridge regression by proving lower bounds on oblivious sketching matrices for AMM. The sketch size lower bounds for AMM may be of independent interest.

Flashlight: Enabling Innovation in Tools for Machine Learning

- Jacob D Kahn, Vineel Pratap, Tatiana Likhomanenko, Qiantong Xu, Awni Hannun, Jeff Cai, Paden Tomasello, Ann Lee, Edouard Grave, Gilad Avidov, Benoit Steiner, Vitaliy Liptchinsky, Gabriel Synnaeve, Ronan Collobert
- abstract: As the computational requirements for machine learning systems and the size and complexity of machine learning frameworks increases, essential framework innovation has become challenging. While computational needs have driven recent compiler, networking, and hardware advancements, utilization of those advancements by machine learning tools is occurring at a slower pace. This is in part due to the difficulties involved in prototyping new computational paradigms with existing frameworks. Large frameworks prioritize machine learning researchers and practitioners as end users and pay comparatively little attention to systems researchers who can push frameworks forward — we argue that both are equally important stakeholders. We introduce Flashlight, an open-source library built to spur innovation in machine learning tools and systems by prioritizing open, modular, customizable internals and state-of-the-art, research-ready models and training setups across a variety of domains. Flashlight allows systems researchers to rapidly prototype and experiment with novel ideas in machine learning computation and has low overhead, competing with and often outperforming other popular machine learning frameworks. We see Flashlight as a tool enabling research that can benefit widely used libraries downstream and bring machine learning and systems researchers closer together.

Learning-based Optimisation of Particle Accelerators Under Partial Observability Without Real-World Training

- Jan Kaiser, Oliver Stein, Annika Eichler
- abstract: In recent work, it has been shown that reinforcement learning (RL) is capable of solving a variety of problems at sometimes super-human performance levels. But despite continued advances in the field, applying RL to complex real-world control and optimisation problems has proven difficult. In this contribution, we demonstrate how to successfully apply RL to the optimisation of a highly complex real-world machine {–} specifically a linear particle accelerator {–} in an only partially observable setting and without requiring training on the real machine. Our method outperforms conventional optimisation algorithms in both the achieved result and time taken as well as already achieving close to human-level performance. We expect that such automation of machine optimisation will push the limits of operability, increase machine availability and lead to a paradigm shift in how such machines are operated, ultimately facilitating advances in a variety of fields, such as science and medicine among many others.

Stochastic Deep Networks with Linear Competing Units for Model-Agnostic Meta-Learning

- Konstantinos Kalais, Sotirios Chatzis
- abstract: This work addresses meta-learning (ML) by considering deep networks with stochastic local winner-takes-all (LWTA) activations. This type of network units results in sparse representations from each model layer, as the units are organized into blocks where only one unit generates a non-zero output. The main operating principle of the introduced units rely on stochastic principles, as the network performs posterior sampling over competing units to select the winner. Therefore, the proposed networks are explicitly designed to extract input data representations of sparse stochastic nature, as opposed to the currently standard deterministic representation paradigm. Our approach produces state-of-the-art predictive accuracy on few-shot image classification and regression experiments, as well as reduced predictive error on an active learning setting; these improvements come with an immensely reduced computational cost. Code is available at: https://github.com/Kkalais/StochLWTA-ML

Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning

- Nathan Kallus, Xiaojie Mao, Kaiwen Wang, Zhengyuan Zhou
- abstract: Off-policy evaluation and learning (OPE/L) use offline observational data to make better decisions, which is crucial in applications where online experimentation is limited. However, depending entirely on logged data, OPE/L is sensitive to environment distribution shifts — discrepancies between the data-generating environment and that where policies are deployed. Si et al., (2020) proposed distributionally robust OPE/L (DROPE/L) to address this, but the proposal relies on inverse-propensity weighting, whose estimation error and regret will deteriorate if propensities are nonparametrically estimated and whose variance is suboptimal even if not. For standard, non-robust, OPE/L, this is solved by doubly robust (DR) methods, but they do not naturally extend to the more complex DROPE/L, which involves a worst-case expectation. In this paper, we propose the first DR algorithms for DROPE/L with KL-divergence uncertainty sets. For evaluation, we propose Localized Doubly Robust DROPE (LDR^{2}OPE) and show that it achieves semiparametric efficiency under weak product rates conditions. Thanks to a localization technique, LDR^{2}OPE only requires fitting a small number of regressions, just like DR methods for standard OPE. For learning, we propose Continuum Doubly Robust DROPL (CDR^{2}OPL) and show that, under a product rate condition involving a continuum of regressions, it enjoys a fast regret rate of $O(N^{-1/2})$ even when unknown propensities are nonparametrically estimated. We empirically validate our algorithms in simulations and further extend our results to general f-divergence uncertainty sets.

Improved Rates for Differentially Private Stochastic Convex Optimization with Heavy-Tailed Data

- Gautam Kamath, Xingtu Liu, Huanyu Zhang
- abstract: We study stochastic convex optimization with heavy-tailed data under the constraint of differential privacy (DP). Most prior work on this problem is restricted to the case where the loss function is Lipschitz. Instead, as introduced by Wang, Xiao, Devadas, and Xu \cite{WangXDX20}, we study general convex loss functions with the assumption that the distribution of gradients has bounded k-th moments. We provide improved upper bounds on the excess population risk under concentrated DP for convex and strongly convex loss functions. Along the way, we derive new algorithms for private mean estimation of heavy-tailed distributions, under both pure and concentrated DP. Finally, we prove nearly-matching lower bounds for private stochastic convex optimization with strongly convex losses and mean estimation, showing new separations between pure and concentrated DP.

Comprehensive Analysis of Negative Sampling in Knowledge Graph Representation Learning

- Hidetaka Kamigaito, Katsuhiko Hayashi
- abstract: Negative sampling (NS) loss plays an important role in learning knowledge graph embedding (KGE) to handle a huge number of entities. However, the performance of KGE degrades without hyperparameters such as the margin term and number of negative samples in NS loss being appropriately selected. Currently, empirical hyperparameter tuning addresses this problem at the cost of computational time. To solve this problem, we theoretically analyzed NS loss to assist hyperparameter tuning and understand the better use of the NS loss in KGE learning. Our theoretical analysis showed that scoring methods with restricted value ranges, such as TransE and RotatE, require appropriate adjustment of the margin term or the number of negative samples different from those without restricted value ranges, such as RESCAL, ComplEx, and DistMult. We also propose subsampling methods specialized for the NS loss in KGE studied from a theoretical aspect. Our empirical analysis on the FB15k-237, WN18RR, and YAGO3-10 datasets showed that the results of actually trained models agree with our theoretical findings.

Matching Learned Causal Effects of Neural Networks with Domain Priors

- Sai Srinivas Kancheti, Abbavaram Gowtham Reddy, Vineeth N Balasubramanian, Amit Sharma
- abstract: A trained neural network can be interpreted as a structural causal model (SCM) that provides the effect of changing input variables on the model's output. However, if training data contains both causal and correlational relationships, a model that optimizes prediction accuracy may not necessarily learn the true causal relationships between input and output variables. On the other hand, expert users often have prior knowledge of the causal relationship between certain input variables and output from domain knowledge. Therefore, we propose a regularization method that aligns the learned causal effects of a neural network with domain priors, including both direct and total causal effects. We show that this approach can generalize to different kinds of domain priors, including monotonicity of causal effect of an input variable on output or zero causal effect of a variable on output for purposes of

fairness. Our experiments on twelve benchmark datasets show its utility in regularizing a neural network model to maintain desired causal effects, without compromising on accuracy. Importantly, we also show that a model thus trained is robust and gets improved accuracy on noisy inputs.

Deduplicating Training Data Mitigates Privacy Risks in Language Models

- Nikhil Kandpal, Eric Wallace, Colin Raffel
- abstract: Past work has shown that large language models are susceptible to privacy attacks, where adversaries generate sequences from a trained model and detect which sequences are memorized from the training set. In this work, we show that the success of these attacks is largely due to duplication in commonly used web-scraped training sets. We first show that the rate at which language models regenerate training sequences is superlinearly related to a sequence's count in the training set. For instance, a sequence that is present 10 times in the training data is on average generated 1000x more often than a sequence that is present only once. We next show that existing methods for detecting memorized sequences have near-chance accuracy on non-duplicated training sequences. Finally, we find that after applying methods to deduplicate training data, language models are considerably more secure against these types of privacy attacks. Taken together, our results motivate an increased focus on deduplication in privacy-sensitive applications and a reevaluation of the practicality of existing privacy attacks.

Lyapunov Density Models: Constraining Distribution Shift in Learning-Based Control

- Katie Kang, Paula Gradu, Jason J Choi, Michael Janner, Claire Tomlin, Sergey Levine
- abstract: Learned models and policies can generalize effectively when evaluated within the distribution of the training data, but can produce unpredictable and erroneous outputs on out-of-distribution inputs. In order to avoid distribution shift when deploying learning-based control algorithms, we seek a mechanism to constrain the agent to states and actions that resemble those that the method was trained on. In control theory, Lyapunov stability and control-invariant sets allow us to make guarantees about controllers that stabilize the system around specific states, while in machine learning, density models allow us to estimate the training data distribution. Can we combine these two concepts, producing learning-based control algorithms that constrain the system to in-distribution states using only in-distribution actions? In this paper, we propose to do this by combining concepts from Lyapunov stability and density estimation, introducing Lyapunov density models: a generalization of control Lyapunov functions and density models that provides guarantees about an agent's ability to stay in-distribution over its entire trajectory.

Forget-free Continual Learning with Winning Subnetworks

- Haeyong Kang, Rusty John Lloyd Mina, Sultan Rizky Hikmawan Madjid, Jaehong Yoon, Mark Hasegawa-Johnson, Sung Ju Hwang, Chang D. Yoo
- abstract: Inspired by Lottery Ticket Hypothesis that competitive subnetworks exist within a dense network, we propose a continual learning method referred to as Winning SubNetworks (WSN), which sequentially learns and selects an optimal subnetwork for each task. Specifically, WSN jointly learns the model weights and task-adaptive binary masks pertaining to subnetworks associated with each task whilst attempting to select a small set of weights to be activated (winning ticket) by reusing weights of the prior subnetworks. The proposed method is inherently immune to catastrophic forgetting as each selected subnetwork model does not infringe upon other subnetworks. Binary masks spawned per winning ticket are encoded into one N-bit binary digit mask, then compressed using Huffman coding for a sub-linear increase in network capacity with respect to the number of tasks.

Differentially Private Approximate Quantiles

- Haim Kaplan, Shachar Schnapp, Uri Stemmer
- abstract: In this work we study the problem of differentially private (DP) quantiles, in which given dataset X and quantiles $q_1, ..., q_m \in [0,1]$, we want to output m quantile estimations which are as close as possible to the true quantiles and preserve DP. We describe a simple recursive DP algorithm, which we call Approximate Quantiles (AQ), for this task. We give a worst case upper bound on its error, and show that its error is much lower than of previous implementations on several different datasets. Furthermore, it gets this low error while running time two orders of magnitude faster that the best previous implementation.

Simultaneous Graph Signal Clustering and Graph Learning

- Abdullah Karaaslanli, Selin Aviyente
- abstract: Graph learning (GL) aims to infer the topology of an unknown graph from a set of observations on its nodes, i.e., graph signals. While most of the existing GL approaches focus on homogeneous datasets, in many real world applications, data is heterogeneous, where graph signals are clustered and each cluster is associated with a different graph. In this paper, we address the problem of learning multiple graphs from heterogeneous data by formulating an optimization problem for joint graph signal clustering and graph topology inference. In particular, our approach extends spectral clustering by partitioning the graph signals not only based on their pairwise similarities but also their smoothness with respect to the graphs associated with the clusters. The proposed method also learns the representative graph for each cluster using the smoothness of the graph signals with respect to the graph topology. The resulting optimization problem is solved with an efficient block-coordinate descent algorithm and results on simulated and real data indicate the effectiveness of the proposed method.

Composing Partial Differential Equations with Physics-Aware Neural Networks

- Matthias Karlbauer, Timothy Praditia, Sebastian Otte, Sergey Oladyshkin, Wolfgang Nowak, Martin V. Butz
- abstract: We introduce a compositional physics-aware FInite volume Neural Network (FINN) for learning spatiotemporal advection-diffusion processes. FINN implements a new way of combining the learning abilities of artificial neural networks with physical and structural knowledge from numerical simulation by modeling the constituents of partial differential equations (PDEs) in a compositional manner. Results on both one- and two-dimensional PDEs (Burgers', diffusion-sorption, diffusion-reaction, Allen{–}Cahn) demonstrate FINN's superior modeling accuracy and excellent out-of-distribution generalization ability beyond initial and boundary conditions. With only one tenth of the number of parameters on average, FINN outperforms pure machine learning and other state-of-the-art physics-aware models in all cases{—}often even by multiple orders of magnitude. Moreover, FINN outperforms a calibrated physical model when approximating sparse real-world data in a diffusion-sorption scenario, confirming its generalization abilities and showing explanatory potential by revealing the unknown retardation factor of the observed process.

Meta-Learning Hypothesis Spaces for Sequential Decision-making

- Parnian Kassraie, Jonas Rothfuss, Andreas Krause
- abstract: Obtaining reliable, adaptive confidence sets for prediction functions (hypotheses) is a central challenge in sequential decision-making tasks, such as bandits and model-based reinforcement learning. These confidence sets typically rely on prior assumptions on the hypothesis space, e.g., the known kernel of a Reproducing Kernel Hilbert Space (RKHS). Hand-designing such kernels is error prone, and misspecification may lead to poor or unsafe performance. In this work, we propose to meta-learn a kernel from offline data (Meta-KeL). For the case where the unknown kernel is a combination of known base kernels, we develop an estimator based on structured sparsity. Under mild conditions, we guarantee that our estimated RKHS yields valid confidence sets that, with increasing amounts of offline data, become as tight as those given the true unknown kernel. We demonstrate our approach on the kernelized bandits problem (a.k.a. Bayesian optimization), where we establish regret bounds competitive with those given the true kernel. We also empirically evaluate the effectiveness of our approach on a Bayesian optimization task.

FOCUS: Familiar Objects in Common and Uncommon Settings

- Priyatham Kattakinda, Soheil Feizi
- abstract: Standard training datasets for deep learning often do not contain objects in uncommon and rare settings (e.g., “a plane on water”, “a car in snowy weather”). This can cause models trained on these datasets to incorrectly predict objects that are typical for the context in the image, rather than identifying the objects that are actually present. In this paper, we introduce FOCUS (Familiar Objects in Common and Uncommon Settings), a dataset for stress-testing the generalization power of deep image classifiers. By leveraging the power of modern search engines, we deliberately gather data containing objects in common and uncommon settings; in a wide range of locations, weather conditions, and time of day. We present a detailed analysis of the performance of various popular image classifiers on our dataset and demonstrate a clear drop in accuracy when classifying images in uncommon settings. We also show that finetuning a model on our dataset drastically improves its ability to focus on the object of interest leading to better generalization. Lastly, we leverage FOCUS to machine annotate additional visual attributes for the entirety of ImageNet. We believe that our dataset will aid researchers in understanding the inability of deep models to generalize well to uncommon settings and drive future work on improving their distributional robustness.

Training OOD Detectors in their Natural Habitats

- Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, Yixuan Li
- abstract: Out-of-distribution (OOD) detection is important for machine learning models deployed in the wild. Recent methods use auxiliary outlier data to regularize the model for improved OOD detection. However, these approaches make a strong distributional assumption that the auxiliary outlier data is completely separable from the in-distribution (ID) data. In this paper, we propose a novel framework that leverages wild mixture data—that naturally consists of both ID and OOD samples. Such wild data is abundant and arises freely upon deploying a machine learning classifier in their natural habitats. Our key idea is to formulate a constrained optimization problem and to show how to tractably solve it. Our learning objective maximizes the OOD detection rate, subject to constraints on the classification error of ID data and on the OOD error rate of ID examples. We extensively evaluate our approach on common OOD detection tasks and demonstrate superior performance. Code is available at https://github.com/jkatzsam/woods_ood.

Robustness Implies Generalization via Data-Dependent Generalization Bounds

- Kenji Kawaguchi, Zhun Deng, Kyle Luh, Jiaoyang Huang
- abstract: This paper proves that robustness implies generalization via data-dependent generalization bounds. As a result, robustness and generalization are shown to be connected closely in a data-dependent manner. Our bounds improve previous bounds in two directions, to solve an open problem that has seen little development since 2010. The first is to reduce the dependence on the covering number. The second is to remove the dependence on the hypothesis space. We present several examples, including ones for lasso and deep learning, in which our bounds are provably preferable. The experiments on real-world data and theoretical models demonstrate near-exponential improvements in various situations. To achieve these improvements, we do not require additional assumptions on the unknown distribution; instead, we only incorporate an observable and computable property of the training samples. A key technical innovation is an improved concentration bound for multinomial random variables that is of independent interest beyond robustness and generalization.

Generating Distributional Adversarial Examples to Evade Statistical Detectors

- Yigitcan Kaya, Muhammad Bilal Zafar, Sergul Aydore, Nathalie Rauschmayr, Krishnaram Kenthapadi
- abstract: Deep neural networks (DNNs) are known to be highly vulnerable to adversarial examples (AEs) that include malicious perturbations. Assumptions about the statistical differences between natural and adversarial inputs are commonplace in many detection techniques. As a best practice, AE detectors are evaluated against ’adaptive’ attackers who actively perturb their inputs to avoid detection. Due to the difficulties in designing adaptive attacks, however, recent work suggests that most detectors have incomplete evaluation. We aim to fill this gap by designing a generic adaptive attack against detectors: the ’statistical indistinguishability attack’ (SIA). SIA optimizes a novel objective to craft adversarial examples (AEs) that follow the same distribution as the natural inputs with respect to DNN representations. Our objective targets all DNN layers simultaneously as we show that AEs being indistinguishable at one layer might fail to be so at other layers. SIA is formulated around evading distributional detectors that inspect a set of AEs as a whole and is also effective against four individual AE detectors, two dataset shift detectors, and an out-of-distribution sample detector, curated from published works. This suggests that SIA can be a reliable tool for evaluating the security of a range of detectors.

Secure Quantized Training for Deep Learning

- Marcel Keller, Ke Sun
- abstract: We implement training of neural networks in secure multi-party computation (MPC) using quantization commonly used in said setting. We are the first to present an MNIST classifier purely trained in MPC that comes within 0.2 percent of the accuracy of the same convolutional neural network trained via plaintext computation. More concretely, we have trained a network with two convolutional and two dense layers to 99.2% accuracy in 3.5 hours (under one hour for 99% accuracy). We have also implemented AlexNet for CIFAR-10, which converges in a few hours. We develop novel protocols for exponentiation and inverse square root. Finally, we present experiments in a range of MPC security models for up to ten parties, both with honest and dishonest majority as well as semi-honest and malicious security.

A Convergent and Dimension-Independent Min-Max Optimization Algorithm

- Vijay Keswani, Oren Mangoubi, Sushant Sachdeva, Nisheeth K. Vishnoi
- abstract: We study a variant of a recently introduced min-max optimization framework where the max-player is constrained to update its parameters in a greedy manner until it reaches a first-order stationary point. Our equilibrium definition for this framework depends on a proposal distribution which the min-player uses to choose directions in which to update its parameters. We show that, given a smooth and bounded nonconvex-nonconcave objective function, access to any proposal distribution for the min-player’s updates, and stochastic gradient oracle for the max-player, our algorithm converges to the aforementioned approximate local equilibrium in a number of iterations that does not depend on the dimension. The equilibrium point found by our algorithm depends on the proposal distribution, and when applying our algorithm to train GANs we choose the proposal distribution to be a distribution of stochastic gradients. We empirically evaluate our algorithm on challenging nonconvex-nonconcave test-functions and loss functions arising in GAN training. Our algorithm converges on these test functions and, when used to train GANs, trains stably on synthetic and real-world datasets and avoids mode collapse.

Neural Network Poisson Models for Behavioural and Neural Spike Train Data

- Moein Khajehnejad, Forough Habibollahi, Richard Nock, Ehsan Arabzadeh, Peter Dayan, Amir Dezfouli
- abstract: One of the most important and challenging application areas for complex machine learning methods is to predict, characterize and model rich, multi-dimensional, neural data. Recent advances in neural recording techniques have made it possible to monitor the activity of a large number of neurons across different brain regions as animals perform behavioural tasks. This poses the critical challenge of establishing links between neural activity at a microscopic scale, which might for instance represent sensory input, and at a macroscopic scale, which then generates behaviour. Predominant modeling methods apply rather disjoint techniques to these scales; by contrast, we suggest an end-to-end model which exploits recent developments of flexible, but

tractable, neural network point-process models to characterize dependencies between stimuli, actions, and neural data. We apply this model to a public dataset collected using Neuropixel probes in mice performing a visually-guided behavioural task as well as a synthetic dataset produced from a hierarchical network model with reciprocally connected sensory and integration circuits intended to characterize animal behaviour in a fixed-duration motion discrimination task. We show that our model outperforms previous approaches and contributes novel insights into the relationships between neural activity and behaviour.

Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling

- Sajad Khodadadian, Pranay Sharma, Gauri Joshi, Siva Theja Maguluri
- abstract: Since reinforcement learning algorithms are notoriously data-intensive, the task of sampling observations from the environment is usually split across multiple agents. However, transferring these observations from the agents to a central location can be prohibitively expensive in terms of the communication cost, and it can also compromise the privacy of each agent's local behavior policy. In this paper, we consider a federated reinforcement learning framework where multiple agents collaboratively learn a global model, without sharing their individual data and policies. Each agent maintains a local copy of the model and updates it using locally sampled data. Although having N agents enables the sampling of N times more data, it is not clear if it leads to proportional convergence speedup. We propose federated versions of on-policy TD, off-policy TD and Q-learning, and analyze their convergence. For all these algorithms, to the best of our knowledge, we are the first to consider Markovian noise and multiple local updates, and prove a linear convergence speedup with respect to the number of agents. To obtain these results, we show that federated TD and Q-learning are special cases of a general framework for federated stochastic approximation with Markovian noise, and we leverage this framework to provide a unified convergence analysis that applies to all the algorithms.

Multi-Level Branched Regularization for Federated Learning

- Jinkyu Kim, Geeho Kim, Bohyung Han
- abstract: A critical challenge of federated learning is data heterogeneity and imbalance across clients, which leads to inconsistency between local networks and unstable convergence of global models. To alleviate the limitations, we propose a novel architectural regularization technique that constructs multiple auxiliary branches in each local model by grafting local and global subnetworks at several different levels and that learns the representations of the main pathway in the local model congruent to the auxiliary hybrid pathways via online knowledge distillation. The proposed technique is effective to robustify the global model even in the non-iid setting and is applicable to various federated learning frameworks conveniently without incurring extra communication costs. We perform comprehensive empirical studies and demonstrate remarkable performance gains in terms of accuracy and efficiency compared to existing methods. The source code is available at our project page.

Learning fair representation with a parametric integral probability metric

- Dongha Kim, Kunwoong Kim, Insung Kong, Ilsang Ohn, Yongdai Kim
- abstract: As they have a vital effect on social decision-making, AI algorithms should be not only accurate but also fair. Among various algorithms for fairness AI, learning fair representation (LFR), whose goal is to find a fair representation with respect to sensitive variables such as gender and race, has received much attention. For LFR, the adversarial training scheme is popularly employed as is done in the generative adversarial network type algorithms. The choice of a discriminator, however, is done heuristically without justification. In this paper, we propose a new adversarial training scheme for LFR, where the integral probability metric (IPM) with a specific parametric family of discriminators is used. The most notable result of the proposed LFR algorithm is its theoretical guarantee about the fairness of the final prediction model, which has not been considered yet. That is, we derive theoretical relations between the fairness of representation and the fairness of the prediction model built on the top of the representation (i.e., using the representation as the input). Moreover, by numerical experiments, we show that our proposed LFR algorithm is computationally lighter and more stable, and the final prediction model is competitive or superior to other LFR algorithms using more complex discriminators.

Dataset Condensation via Efficient Synthetic-Data Parameterization

- Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, Hyun Oh Song
- abstract: The great success of machine learning with massive amounts of data comes at a price of huge computation costs and storage for training and tuning. Recent studies on dataset condensation attempt to reduce the dependence on such massive data by synthesizing a compact training dataset. However, the existing approaches have fundamental limitations in optimization due to the limited representability of synthetic datasets without considering any data regularity characteristics. To this end, we propose a novel condensation framework that generates multiple synthetic data with a limited storage budget via efficient parameterization considering data regularity. We further analyze the shortcomings of the existing gradient matching-based condensation methods and develop an effective optimization technique for improving the condensation of training data information. We propose a unified algorithm that drastically improves the quality of condensed data against the current state-of-the-art on CIFAR-10, ImageNet, and Speech Commands.

Guided-TTS: A Diffusion Model for Text-to-Speech via Classifier Guidance

- Heeseung Kim, Sungwon Kim, Sungroh Yoon
- abstract: We propose Guided-TTS, a high-quality text-to-speech (TTS) model that does not require any transcript of target speaker using classifier guidance. Guided-TTS combines an unconditional diffusion probabilistic model with a separately trained phoneme classifier for classifier guidance. Our unconditional diffusion model learns to generate speech without any context from untranscribed speech data. For TTS synthesis, we guide the generative process of the diffusion model with a phoneme classifier trained on a large-scale speech recognition dataset. We present a norm-based scaling method that reduces the pronunciation errors of classifier guidance in Guided-TTS. We show that Guided-TTS achieves a performance comparable to that of the state-of-the-art TTS model, Grad-TTS, without any transcript for LJSpeech. We further demonstrate that Guided-TTS performs well on diverse datasets including a long-form untranscribed dataset.

Variational On-the-Fly Personalization

- Jangho Kim, Jun-Tae Lee, Simyung Chang, Nojun Kwak
- abstract: With the development of deep learning (DL) technologies, the demand for DL-based services on personal devices, such as mobile phones, also increases rapidly. In this paper, we propose a novel personalization method, Variational On-the-Fly Personalization. Compared to the conventional personalization methods that require additional fine-tuning with personal data, the proposed method only requires forwarding a handful of personal data on-the-fly. Assuming even a single personal data can convey the characteristics of a target person, we develop the variational hyper-personalizer to capture the weight distribution of layers that fits the target person. In the testing phase, the hyper-personalizer estimates the model's weights on-the-fly based on personality by forwarding only a small amount of (even a single) personal enrollment data. Hence, the proposed method can perform the personalization without any training software platform and additional cost in the edge device. In experiments, we show our approach can effectively generate reliable personalized models via forwarding (not back-propagating) a handful of samples.

Fisher SAM: Information Geometry and Sharpness Aware Minimisation

- Minyoung Kim, Da Li, Shell X Hu, Timothy Hospedales
- abstract: Recent sharpness-aware minimisation (SAM) is known to find flat minima which is beneficial for better generalisation with improved robustness. SAM essentially modifies the loss function by the maximum loss value within the small neighborhood around the current iterate. However, it uses the Euclidean ball to define the neighborhood, which can be less accurate since loss functions for neural networks are typically defined over probability distributions (e.g., class predictive probabilities), rendering the parameter space no more Euclidean. In this paper we consider the information geometry of the model parameter space when defining the neighborhood, namely replacing SAM's Euclidean balls with ellipsoids induced by the Fisher information. Our approach, dubbed Fisher SAM, defines more accurate neighborhood structures that conform to the intrinsic metric of the underlying statistical manifold. For instance, SAM may probe the worst-case loss value at either a too nearby or inappropriately distant point due to the ignorance of the parameter space geometry, which is avoided by our Fisher SAM. Another recent Adaptive SAM approach that stretches/shrinks the Euclidean ball in accordance with the scales of the parameter magnitudes, might be dangerous, potentially destroying the neighborhood structure even severely. We demonstrate the improved performance of the proposed Fisher SAM on several benchmark datasets/tasks.

ViT-NeT: Interpretable Vision Transformers with Neural Tree Decoder

- Sangwon Kim, Jaeyeal Nam, Byoung Chul Ko
- abstract: Vision transformers (ViTs), which have demonstrated a state-of-the-art performance in image classification, can also visualize global interpretations through attention-based contributions. However, the complexity of the model makes it difficult to interpret the decision-making process, and the ambiguity of the attention maps can cause incorrect correlations between image patches. In this study, we propose a new ViT neural tree decoder (ViT-NeT). A ViT acts as a backbone, and to solve its limitations, the output contextual image patches are applied to the proposed NeT. The NeT aims to accurately classify fine-grained objects with similar inter-class correlations and different intra-class correlations. In addition, it describes the decision-making process through a tree structure and prototype and enables a visual interpretation of the results. The proposed ViT-NeT is designed to not only improve the classification performance but also provide a human-friendly interpretation, which is effective in resolving the trade-off between performance and interpretability. We compared the performance of ViT-NeT with other state-of-art methods using widely used fine-grained visual categorization benchmark datasets and experimentally proved that the proposed method is superior in terms of the classification performance and interpretability. The code and models are publicly available at https://github.com/jumpsnack/ViT-NeT.

Sanity Simulations for Saliency Methods

- Joon Sik Kim, Gregory Plumb, Ameet Talwalkar
- abstract: Saliency methods are a popular class of feature attribution explanation methods that aim to capture a model's predictive reasoning by identifying "important" pixels in an input image. However, the development and adoption of these methods are hindered by the lack of access to ground-truth model reasoning, which prevents accurate evaluation. In this work, we design a synthetic benchmarking framework, SMERF, that allows us to perform ground-truth-based evaluation while controlling the complexity of the model's reasoning. Experimentally, SMERF reveals significant limitations in existing saliency methods and, as a result, represents a useful tool for the development of new saliency methods.

Soft Truncation: A Universal Training Technique of Score-based Diffusion Model for High Precision Score Estimation

- Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, Il-Chul Moon
- abstract: Recent advances in diffusion models bring state-of-the-art performance on image generation tasks. However, empirical results from previous research in diffusion models imply an inverse correlation between density estimation and sample generation performances. This paper investigates with sufficient empirical evidence that such inverse correlation happens because density estimation is significantly contributed by small diffusion time, whereas sample generation mainly depends on large diffusion time. However, training a score network well across the entire diffusion time is demanding because the loss scale is significantly imbalanced at each diffusion time. For successful training, therefore, we introduce Soft Truncation, a universally applicable training technique for diffusion models, that softens the fixed and static truncation hyperparameter into a random variable. In experiments, Soft Truncation achieves state-of-the-art performance on CIFAR-10, CelebA, CelebA-HQ 256×256, and STL-10 datasets.

Rotting Infinitely Many-Armed Bandits

- Jung-Hun Kim, Milan Vojnovic, Se-Young Yun
- abstract: We consider the infinitely many-armed bandit problem with rotting rewards, where the mean reward of an arm decreases at each pull of the arm according to an arbitrary trend with maximum rotting rate $\varrho=o(1)$. We show that this learning problem has an $\Omega(\max{\varrho^{1/3}T, \sqrt{T}})$ worst-case regret lower bound where T is the time horizon. We show that a matching upper bound $\tilde{O}(\max{\varrho^{1/3}T, \sqrt{T}})$, up to a poly-logarithmic factor, can be achieved by an algorithm that uses a UCB index for each arm and a threshold value to decide whether to continue pulling an arm or remove the arm from further consideration, when the algorithm knows the value of the maximum rotting rate ϱ. We also show that an $\tilde{O}(\max{\varrho^{1/3}T, T^{3/4}})$ regret upper bound can be achieved by an algorithm that does not know the value of ϱ, by using an adaptive UCB index along with an adaptive threshold value.

Accelerated Gradient Methods for Geodesically Convex Optimization: Tractable Algorithms and Convergence Analysis

- Jungbin Kim, Insoon Yang
- abstract: We propose computationally tractable accelerated first-order methods for Riemannian optimization, extending the Nesterov accelerated gradient (NAG) method. For both geodesically convex and geodesically strongly convex objective functions, our algorithms are shown to have the same iteration complexities as those for the NAG method on Euclidean spaces, under only standard assumptions. To the best of our knowledge, the proposed scheme is the first fully accelerated method for geodesically convex optimization problems. Our convergence analysis makes use of novel metric distortion lemmas as well as carefully designed potential functions. A connection with the continuous-time dynamics for modeling Riemannian acceleration in (Alimisis et al., 2020) is also identified by letting the stepsize tend to zero. We validate our theoretical results through numerical experiments.

Generalizing to New Physical Systems via Context-Informed Dynamics Model

- Matthieu Kirchmeyer, Yuan Yin, Jeremie Dona, Nicolas Baskiotis, Alain Rakotomamonjy, Patrick Gallinari
- abstract: Data-driven approaches to modeling physical systems fail to generalize to unseen systems that share the same general dynamics with the learning domain, but correspond to different physical contexts. We propose a new framework for this key problem, context-informed dynamics adaptation (CoDA), which takes into account the distributional shift across systems for fast and efficient adaptation to new dynamics. CoDA leverages multiple environments, each associated to a different dynamic, and learns to condition the dynamics model on contextual parameters, specific to each environment. The conditioning is performed via a hypernetwork, learned jointly with a context vector from observed data. The proposed formulation constrains the search hypothesis space for fast adaptation and better generalization across environments with few samples. We theoretically motivate our approach and show state-of-the-art generalization results on a set of nonlinear dynamics, representative of a variety of application domains. We also show, on these systems, that new system parameters can be inferred from context vectors with minimal supervision.

SoQal: Selective Oracle Questioning for Consistency Based Active Learning of Cardiac Signals

- Dani Kiyasseh, Tingting Zhu, David A Clifton
- abstract: Clinical settings are often characterized by abundant unlabelled data and limited labelled data. This is typically driven by the high burden placed on oracles (e.g., physicians) to provide annotations. One way to mitigate this burden is via active learning (AL) which involves the (a) acquisition and (b) annotation of informative unlabelled instances. Whereas previous work addresses either one of these elements independently, we propose an AL framework that addresses both. For acquisition, we propose Bayesian Active Learning by Consistency (BALC), a sub-framework which perturbs both instances and network parameters and quantifies changes in the network output probability distribution. For annotation, we propose SoQal, a sub-framework that dynamically determines whether, for each acquired unlabelled instance, to request a label from an oracle or to pseudo-label it instead. We show that BALC can outperform start-of-the-art acquisition functions such as BALD, and SoQal outperforms baseline methods even in the presence of a noisy oracle.

Curriculum Reinforcement Learning via Constrained Optimal Transport

- Pascal Klink, Haoyi Yang, Carlo D'Eramo, Jan Peters, Joni Pajarinen
- abstract: Curriculum reinforcement learning (CRL) allows solving complex tasks by generating a tailored sequence of learning tasks, starting from easy ones and subsequently increasing their difficulty. Although the potential of curricula in RL has been clearly shown in a variety of works, it is less clear how to generate them for a given learning environment, resulting in a variety of methods aiming to automate this task. In this work, we focus on the idea of framing curricula as interpolations between task distributions, which has previously been shown to be a viable approach to CRL. Identifying key issues of existing methods, we frame the generation of a curriculum as a constrained optimal transport problem between task distributions. Benchmarks show that this way of curriculum generation can improve upon existing CRL methods, yielding high performance in a variety of tasks with different characteristics.

Exploiting Redundancy: Separable Group Convolutional Networks on Lie Groups

- David M. Knigge, David W Romero, Erik J Bekkers
- abstract: Group convolutional neural networks (G-CNNs) have been shown to increase parameter efficiency and model accuracy by incorporating geometric inductive biases. In this work, we investigate the properties of representations learned by regular G-CNNs, and show considerable parameter redundancy in group convolution kernels. This finding motivates further weight-tying by sharing convolution kernels over subgroups. To this end, we introduce convolution kernels that are separable over the subgroup and channel dimensions. In order to obtain equivariance to arbitrary affine Lie groups we provide a continuous parameterisation of separable convolution kernels. We evaluate our approach across several vision datasets, and show that our weight sharing leads to improved performance and computational efficiency. In many settings, separable G-CNNs outperform their non-separable counterpart, while only using a fraction of their training time. In addition, thanks to the increase in computational efficiency, we are able to implement G-CNNs equivariant to the $\mathrm{Sim(2)}$ group; the group of dilations, rotations and translations of the plane. $\mathrm{Sim(2)}$-equivariance further improves performance on all tasks considered, and achieves state-of-the-art performance on rotated MNIST.

Revisiting Contrastive Learning through the Lens of Neighborhood Component Analysis: an Integrated Framework

- Ching-Yun Ko, Jeet Mohapatra, Sijia Liu, Pin-Yu Chen, Luca Daniel, Lily Weng
- abstract: As a seminal tool in self-supervised representation learning, contrastive learning has gained unprecedented attention in recent years. In essence, contrastive learning aims to leverage pairs of positive and negative samples for representation learning, which relates to exploiting neighborhood information in a feature space. By investigating the connection between contrastive learning and neighborhood component analysis (NCA), we provide a novel stochastic nearest neighbor viewpoint of contrastive learning and subsequently propose a series of contrastive losses that outperform the existing ones. Under our proposed framework, we show a new methodology to design integrated contrastive losses that could simultaneously achieve good accuracy and robustness on downstream tasks. With the integrated framework, we achieve up to 6% improvement on the standard accuracy and 17% improvement on the robust accuracy.

Transfer Learning In Differential Privacy's Hybrid-Model

- Refael Kohen, Or Sheffet
- abstract: The hybrid-model (Avent et al 2017) in Differential Privacy is a an augmentation of the local-model where in addition to N local-agents we are assisted by one special agent who is in fact a curator holding the sensitive details of n additional individuals. Here we study the problem of machine learning in the hybrid-model where the n individuals in the curator's dataset are drawn from a different distribution than the one of the general population (the local-agents). We give a general scheme – Subsample-Test-Reweigh – for this transfer learning problem, which reduces any curator-model learner to a learner in the hybrid-model using iterative subsampling and reweighing of the n examples held by the curator based on a smooth variation (introduced by Bun et al 2020) of the Multiplicative-Weights algorithm. Our scheme has a sample complexity which relies on the χ^2-divergence between the two distributions. We give worst-case analysis bounds on the sample complexity required for our private reduction. Aiming to reduce said sample complexity, we give two specific instances our sample complexity can be drastically reduced (one instance is analyzed mathematically, while the other - empirically) and pose several directions for follow-up work.

Markov Chain Monte Carlo for Continuous-Time Switching Dynamical Systems

- Lukas Köhs, Bastian Alt, Heinz Koeppl
- abstract: Switching dynamical systems are an expressive model class for the analysis of time-series data. As in many fields within the natural and engineering sciences, the systems under study typically evolve continuously in time, it is natural to consider continuous-time model formulations consisting of switching stochastic differential equations governed by an underlying Markov jump process. Inference in these types of models is however notoriously difficult, and tractable computational schemes are rare. In this work, we propose a novel inference algorithm utilizing a Markov Chain Monte Carlo approach. The presented Gibbs sampler allows to efficiently obtain samples from the exact continuous-time posterior processes. Our framework naturally enables Bayesian parameter estimation, and we also include an estimate for the diffusion covariance, which is oftentimes assumed fixed in stochastic differential equations models. We evaluate our framework under the modeling assumption and compare it against an existing variational inference approach.

Partial disentanglement for domain adaptation

- Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, Kun Zhang
- abstract: Unsupervised domain adaptation is critical to many real-world applications where label information is unavailable in the target domain. In general, without further assumptions, the joint distribution of the features and the label is not identifiable in the target domain. To address this issue, we rely on a property of minimal changes of causal mechanisms across domains to minimize unnecessary influences of domain shift. To encode this property, we first formulate the data generating process using a latent variable model with two partitioned latent subspaces: invariant components whose distributions stay the same across domains, and sparse changing components that vary across domains. We further constrain the domain shift to have a restrictive influence on the changing components. Under mild conditions, we show that the latent variables are partially identifiable, from which it follows that the joint distribution of data and labels in the target domain is also identifiable. Given the theoretical insights, we propose a practical domain adaptation framework, called iMSDA. Extensive experimental results reveal that iMSDA outperforms state-of-the-art domain adaptation algorithms on benchmark datasets, demonstrating the effectiveness of our framework.

Simultaneously Learning Stochastic and Adversarial Bandits with General Graph Feedback

- Fang Kong, Yichi Zhou, Shuai Li
- abstract: The problem of online learning with graph feedback has been extensively studied in the literature due to its generality and potential to model various learning tasks. Existing works mainly study the adversarial and stochastic feedback separately. If the prior knowledge of the feedback mechanism is unavailable or wrong, such specially designed algorithms could suffer great loss. To avoid this problem, \citet{erez2021towards} try to optimize for both environments. However, they assume the feedback graphs are undirected and each vertex has a self-loop, which compromises the generality of the framework and may not be satisfied in applications. With a general feedback graph, the observation of an arm may not be available when this arm is pulled, which makes the exploration more expensive and the algorithms more challenging to perform optimally in both environments. In this work, we overcome this difficulty by a new trade-off mechanism with a carefully-designed proportion for exploration and exploitation. We prove the proposed algorithm simultaneously achieves $\mathrm{poly} \log T$ regret in the stochastic setting and minimax-optimal regret of $\tilde{O}(T^{2/3})$ in the adversarial setting where T is the horizon and $\tilde{O}$ hides parameters independent of T as well as logarithmic terms. To our knowledge, this is the first best-of-both-worlds result for general feedback graphs.

Adaptive Data Analysis with Correlated Observations

- Aryeh Kontorovich, Menachem Sadigurschi, Uri Stemmer
- abstract: The vast majority of the work on adaptive data analysis focuses on the case where the samples in the dataset are independent. Several approaches and tools have been successfully applied in this context, such as differential privacy, max-information, compression arguments, and more. The situation is far less well-understood without the independence assumption. We embark on a systematic study of the possibilities of adaptive data analysis with correlated observations. First, we show that, in some cases, differential privacy guarantees generalization even when there are dependencies within the sample, which we quantify using a notion we call Gibbs-dependence. We complement this result with a tight negative example. % Second, we show that the connection between transcript-compression and adaptive data analysis can be extended to the non-iid setting.

Controlling Conditional Language Models without Catastrophic Forgetting

- Tomasz Korbak, Hady Elsahar, German Kruszewski, Marc Dymetman
- abstract: Machine learning is shifting towards general-purpose pretrained generative models, trained in a self-supervised manner on large amounts of data, which can then be applied to solve a large number of tasks. However, due to their generic training methodology, these models often fail to meet some of the downstream requirements (e.g., hallucinations in abstractive summarization or style violations in code generation). This raises the important question of how to adapt pre-trained generative models to meet all requirements without destroying their general capabilities ("catastrophic forgetting"). Recent work has proposed to solve this problem by representing task-specific requirements through energy-based models (EBMs) and approximating these EBMs using distributional policy gradients (DPG). Despite its effectiveness, this approach is however limited to unconditional distributions. In this paper, we extend DPG to conditional tasks by proposing Conditional DPG (CDPG). We evaluate CDPG on four different control objectives across three tasks (translation, summarization and code generation) and two pretrained models (T5 and GPT-Neo). Our results show that fine-tuning using CDPG robustly moves these pretrained models closer towards meeting control objectives and — in contrast with baseline approaches — does not result in catastrophic forgetting.

Batch Greenkhorn Algorithm for Entropic-Regularized Multimarginal Optimal Transport: Linear Rate of Convergence and Iteration Complexity

- Vladimir R. Kostic, Saverio Salzo, Massimiliano Pontil
- abstract: In this work we propose a batch multimarginal version of the Greenkhorn algorithm for the entropic-regularized optimal transport problem. This framework is general enough to cover, as particular cases, existing Sinkhorn and Greenkhorn algorithms for the bi-marginal setting, and greedy MultiSinkhorn for the general multimarginal case. We provide a comprehensive convergence analysis based on the properties of the iterative Bregman projections method with greedy control. Linear rate of convergence as well as explicit bounds on the iteration complexity are obtained. When specialized to the above mentioned algorithms, our results give new convergence rates or provide key improvements over the state-of-the-art rates. We present numerical experiments showing that the flexibility of the batch can be exploited to improve performance of Sinkhorn algorithm both in bi-marginal and multimarginal settings.

Certified Adversarial Robustness Under the Bounded Support Set

- Yiwen Kou, Qinyuan Zheng, Yisen Wang
- abstract: Deep neural networks (DNNs) have revealed severe vulnerability to adversarial perturbations, beside empirical adversarial training for robustness, the design of provably robust classifiers attracts more and more attention. Randomized smoothing methods provide the certified robustness with agnostic architecture, which is further extended to a provable robustness framework using f-divergence. While these methods cannot be applied to smoothing measures with bounded support set such as uniform probability measure due to the use of likelihood ratio in their certification methods. In this paper, we generalize the f-divergence-based framework to a Wasserstein-distance-based and total-variation-distance-based framework that is first able to analyze robustness properties of bounded support set smoothing measures both theoretically and experimentally. By applying our methodology to uniform probability measures with support set $l_p (p=1,2,\infty\text{ and general})$ ball, we prove negative certified robustness properties with respect to $l_q (q=1, 2, \infty)$ perturbations and present experimental results on CIFAR-10 dataset with ResNet to validate our theory. And it is also worth mentioning that our certification procedure only costs constant computation time.

Exact Learning of Preference Structure: Single-peaked Preferences and Beyond

- Sonja Kraiczy, Edith Elkind
- abstract: We consider the setting where the members of a society (voters) have preferences over candidates, and the candidates can be ordered on an axis so that the voters' preferences are single-peaked on this axis. We ask whether this axis can be identified by sampling the voters' preferences. For several natural distributions, we obtain tight bounds on the number of samples required and show that, surprisingly, the bounds are independent of the number of candidates. We extend our results to the case where voters' preferences are sampled from two different axes over the same candidate set (one of which may be known). We also consider two alternative models of learning: (1) sampling pairwise comparisons rather than entire votes, and (2) learning from equivalence queries.

Reconstructing Nonlinear Dynamical Systems from Multi-Modal Time Series

- Daniel Kramer, Philine L Bommer, Carlo Tombolini, Georgia Koppe, Daniel Durstewitz
- abstract: Empirically observed time series in physics, biology, or medicine, are commonly generated by some underlying dynamical system (DS) which is the target of scientific interest. There is an increasing interest to harvest machine learning methods to reconstruct this latent DS in a data-driven, unsupervised way. In many areas of science it is common to sample time series observations from many data modalities simultaneously, e.g. electrophysiological and behavioral time series in a typical neuroscience experiment. However, current machine learning tools for reconstructing DSs usually focus on just one data modality. Here we propose a general framework for multi-modal data integration for the purpose of nonlinear DS

reconstruction and the analysis of cross-modal relations. This framework is based on dynamically interpretable recurrent neural networks as general approximators of nonlinear DSs, coupled to sets of modality-specific decoder models from the class of generalized linear models. Both an expectation-maximization and a variational inference algorithm for model training are advanced and compared. We show on nonlinear DS benchmarks that our algorithms can efficiently compensate for too noisy or missing information in one data channel by exploiting other channels, and demonstrate on experimental neuroscience data how the algorithm learns to link different data domains to the underlying dynamics.

Probabilistic ODE Solutions in Millions of Dimensions

- Nicholas Krämer, Nathanael Bosch, Jonathan Schmidt, Philipp Hennig
- abstract: Probabilistic solvers for ordinary differential equations (ODEs) have emerged as an efficient framework for uncertainty quantification and inference on dynamical systems. In this work, we explain the mathematical assumptions and detailed implementation schemes behind solving high-dimensional ODEs with a probabilistic numerical algorithm. This has not been possible before due to matrix-matrix operations in each solver step, but is crucial for scientifically relevant problems—most importantly, the solution of discretised partial differential equations. In a nutshell, efficient high-dimensional probabilistic ODE solutions build either on independence assumptions or on Kronecker structure in the prior model. We evaluate the resulting efficiency on a range of problems, including the probabilistic numerical simulation of a differential equation with millions of dimensions.

Active Nearest Neighbor Regression Through Delaunay Refinement

- Alexander Kravberg, Giovanni Luca Marchetti, Vladislav Polianskii, Anastasiia Varava, Florian T. Pokorny, Danica Kragic
- abstract: We introduce an algorithm for active function approximation based on nearest neighbor regression. Our Active Nearest Neighbor Regressor (ANNR) relies on the Voronoi-Delaunay framework from computational geometry to subdivide the space into cells with constant estimated function value and select novel query points in a way that takes the geometry of the function graph into account. We consider the recent state-of-the-art active function approximator called DEFER, which is based on incremental rectangular partitioning of the space, as the main baseline. The ANNR addresses a number of limitations that arise from the space subdivision strategy used in DEFER. We provide a computationally efficient implementation of our method, as well as theoretical halting guarantees. Empirical results show that ANNR outperforms the baseline for both closed-form functions and real-world examples, such as gravitational wave parameter inference and exploration of the latent space of a generative model.

Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions

- Heiner Kremer, Jia-Jie Zhu, Krikamol Muandet, Bernhard Schölkopf
- abstract: Important problems in causal inference, economics, and, more generally, robust machine learning can be expressed as conditional moment restrictions, but estimation becomes challenging as it requires solving a continuum of unconditional moment restrictions. Previous works addressed this problem by extending the generalized method of moments (GMM) to continuum moment restrictions. In contrast, generalized empirical likelihood (GEL) provides a more general framework and has been shown to enjoy favorable small-sample properties compared to GMM-based estimators. To benefit from recent developments in machine learning, we provide a functional reformulation of GEL in which arbitrary models can be leveraged. Motivated by a dual formulation of the resulting infinite dimensional optimization problem, we devise a practical method and explore its asymptotic properties. Finally, we provide kernel- and neural network-based implementations of the estimator, which achieve state-of-the-art empirical performance on two conditional moment restriction problems.

Calibrated and Sharp Uncertainties in Deep Learning via Density Estimation

- Volodymyr Kuleshov, Shachi Deshpande
- abstract: Accurate probabilistic predictions can be characterized by two properties{—}calibration and sharpness. However, standard maximum likelihood training yields models that are poorly calibrated and thus inaccurate{—}a 90% confidence interval typically does not contain the true outcome 90% of the time. This paper argues that calibration is important in practice and is easy to maintain by performing low-dimensional density estimation. We introduce a simple training procedure based on recalibration that yields calibrated models without sacrificing overall performance; unlike previous approaches, ours ensures the most general property of distribution calibration and applies to any model, including neural networks. We formally prove the correctness of our procedure assuming that we can estimate densities in low dimensions and we establish uniform convergence bounds. Our results yield empirical performance improvements on linear and deep Bayesian models and suggest that calibration should be increasingly leveraged across machine learning.

ActiveHedge: Hedge meets Active Learning

- Bhuvesh Kumar, Jacob D Abernethy, Venkatesh Saligrama
- abstract: We consider the classical problem of multi-class prediction with expert advice, but with an active learning twist. In this new setting the learner will only query the labels of a small number of examples, but still aims to minimize regret to the best expert as usual; the learner is also allowed a very short "burn-in" phase where it can fast-forward and query certain highly-informative examples. We design an algorithm that utilizes Hedge (aka Exponential Weights) as a subroutine, and we show that under a very particular combinatorial constraint on the matrix of expert predictions we can obtain a very strong regret guarantee while querying very few labels. This constraint, which we refer to as ζ-compactness, or just compactness, can be viewed as a non-stochastic variant of the disagreement coefficient, another popular parameter used to reason about the sample complexity of active learning in the IID setting. We also give a polynomial-time algorithm to calculate the ζ-compactness of a matrix up to an approximation factor of 3.

Balancing Discriminability and Transferability for Source-Free Domain Adaptation

- Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, Venkatesh Babu Radhakrishnan
- abstract: Conventional domain adaptation (DA) techniques aim to improve domain transferability by learning domain-invariant representations; while concurrently preserving the task-discriminability knowledge gathered from the labeled source data. However, the requirement of simultaneous access to labeled source and unlabeled target renders them unsuitable for the challenging source-free DA setting. The trivial solution of realizing an effective original to generic domain mapping improves transferability but degrades task discriminability. Upon analyzing the hurdles from both theoretical and empirical standpoints, we derive novel insights to show that a mixup between original and corresponding translated generic samples enhances the discriminability-transferability trade-off while duly respecting the privacy-oriented source-free setting. A simple but effective realization of the proposed insights on top of the existing source-free DA approaches yields state-of-the-art performance with faster convergence. Beyond single-source, we also outperform multi-source prior-arts across both classification and semantic segmentation benchmarks.

Showing Your Offline Reinforcement Learning Work: Online Evaluation Budget Matters

- Vladislav Kurenkov, Sergey Kolesnikov
- abstract: In this work, we argue for the importance of an online evaluation budget for a reliable comparison of deep offline RL algorithms. First, we delineate that the online evaluation budget is problem-dependent, where some problems allow for less but others for more. And second, we demonstrate that the preference between algorithms is budget-dependent across a diverse range of decision-making domains such as Robotics, Finance, and Energy Management. Following the points above, we suggest reporting the performance of deep offline RL algorithms under varying online evaluation budgets. To facilitate this, we propose to use a reporting tool from the NLP field, Expected Validation Performance. This technique makes it possible to reliably

estimate expected maximum performance under different budgets while not requiring any additional computation beyond hyperparameter search. By employing this tool, we also show that Behavioral Cloning is often more favorable to offline RL algorithms when working within a limited budget.

Equivariant Priors for compressed sensing with unknown orientation

- Anna Kuzina, Kumar Pratik, Fabio Valerio Massoli, Arash Behboodi
- abstract: In compressed sensing, the goal is to reconstruct the signal from an underdetermined system of linear measurements. Thus, prior knowledge about the signal of interest and its structure is required. Additionally, in many scenarios, the signal has an unknown orientation prior to measurements. To address such recovery problems, we propose using equivariant generative models as a prior, which encapsulate orientation information in their latent space. Thereby, we show that signals with unknown orientations can be recovered with iterative gradient descent on the latent space of these models and provide additional theoretical recovery guarantees. We construct an equivariant variational autoencoder and use the decoder as generative prior for compressed sensing. We discuss additional potential gains of the proposed approach in terms of convergence and latency.

Coordinated Attacks against Contextual Bandits: Fundamental Limits and Defense Mechanisms

- Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, Shie Mannor
- abstract: Motivated by online recommendation systems, we propose the problem of finding the optimal policy in multitask contextual bandits when a small fraction $\alpha < 1/2$ of tasks (users) are arbitrary and adversarial. The remaining fraction of good users share the same instance of contextual bandits with S contexts and A actions (items). Naturally, whether a user is good or adversarial is not known in advance. The goal is to robustly learn the policy that maximizes rewards for good users with as few user interactions as possible. Without adversarial users, established results in collaborative filtering show that $O(1/\epsilon^2)$ per-user interactions suffice to learn a good policy, precisely because information can be shared across users. This parallelization gain is fundamentally altered by the presence of adversarial users: unless there are super-polynomial number of users, we show a lower bound of $\tilde{\Omega}(\min(S,A) \cdot \alpha^2 / \epsilon^2)$ per-user interactions to learn an ϵ-optimal policy for the good users. We then show we can achieve an $\tilde{O}(\min(S,A)\cdot \alpha/\epsilon^2)$ upper-bound, by employing efficient robust mean estimators for both uni-variate and high-dimensional random variables. We also show that this can be improved depending on the distributions of contexts.

Large Batch Experience Replay

- Thibault Lahire, Matthieu Geist, Emmanuel Rachelson
- abstract: Several algorithms have been proposed to sample non-uniformly the replay buffer of deep Reinforcement Learning (RL) agents to speed-up learning, but very few theoretical foundations of these sampling schemes have been provided. Among others, Prioritized Experience Replay appears as a hyperparameter sensitive heuristic, even though it can provide good performance. In this work, we cast the replay buffer sampling problem as an importance sampling one for estimating the gradient. This allows deriving the theoretically optimal sampling distribution, yielding the best theoretical convergence speed. Elaborating on the knowledge of the ideal sampling scheme, we exhibit new theoretical foundations of Prioritized Experience Replay. The optimal sampling distribution being intractable, we make several approximations providing good results in practice and introduce, among others, LaBER (Large Batch Experience Replay), an easy-to-code and efficient method for sampling the replay buffer. LaBER, which can be combined with Deep Q-Networks, distributional RL agents or actor-critic methods, yields improved performance over a diverse range of Atari games and PyBullet environments, compared to the base agent it is implemented on and to other prioritization schemes.

FedScale: Benchmarking Model and System Performance of Federated Learning at Scale

- Fan Lai, Yinwei Dai, Sanjay Singapuram, Jiachen Liu, Xiangfeng Zhu, Harsha Madhyastha, Mosharaf Chowdhury
- abstract: We present FedScale, a federated learning (FL) benchmarking suite with realistic datasets and a scalable runtime to enable reproducible FL research. FedScale datasets encompass a wide range of critical FL tasks, ranging from image classification and object detection to language modeling and speech recognition. Each dataset comes with a unified evaluation protocol using real-world data splits and evaluation metrics. To reproduce realistic FL behavior, FedScale contains a scalable and extensible runtime. It provides high-level APIs to implement FL algorithms, deploy them at scale across diverse hardware and software backends, and evaluate them at scale, all with minimal developer efforts. We combine the two to perform systematic benchmarking experiments and highlight potential opportunities for heterogeneity-aware co-optimizations in FL. FedScale is open-source and actively maintained by contributors from different institutions at http://fedscale.ai. We welcome feedback and contributions from the community.

Smoothed Adaptive Weighting for Imbalanced Semi-Supervised Learning: Improve Reliability Against Unknown Distribution Data

- Zhengfeng Lai, Chao Wang, Henrry Gunawan, Sen-Ching S Cheung, Chen-Nee Chuah
- abstract: Despite recent promising results on semi-supervised learning (SSL), data imbalance, particularly in the unlabeled dataset, could significantly impact the training performance of a SSL algorithm if there is a mismatch between the expected and actual class distributions. The efforts on how to construct a robust SSL framework that can effectively learn from datasets with unknown distributions remain limited. We first investigate the feasibility of adding weights to the consistency loss and then we verify the necessity of smoothed weighting schemes. Based on this study, we propose a self-adaptive algorithm, named Smoothed Adaptive Weighting (SAW). SAW is designed to enhance the robustness of SSL by estimating the learning difficulty of each class and synthesizing the weights in the consistency loss based on such estimation. We show that SAW can complement recent consistency-based SSL algorithms and improve their reliability on various datasets including three standard datasets and one gigapixel medical imaging application without making any assumptions about the distribution of the unlabeled set.

Functional Output Regression with Infimal Convolution: Exploring the Huber and $ε$-insensitive Losses

- Alex Lambert, Dimitri Bouche, Zoltan Szabo, Florence D'Alché-Buc
- abstract: The focus of the paper is functional output regression (FOR) with convoluted losses. While most existing work consider the square loss setting, we leverage extensions of the Huber and the ϵ-insensitive loss (induced by infimal convolution) and propose a flexible framework capable of handling various forms of outliers and sparsity in the FOR family. We derive computationally tractable algorithms relying on duality to tackle the resulting tasks in the context of vector-valued reproducing kernel Hilbert spaces. The efficiency of the approach is demonstrated and contrasted with the classical squared loss setting on both synthetic and real-world benchmarks.

Tell me why! Explanations support learning relational and causal structure

- Andrew K Lampinen, Nicholas Roy, Ishita Dasgupta, Stephanie Cy Chan, Allison Tam, James Mcclelland, Chen Yan, Adam Santoro, Neil C Rabinowitz, Jane Wang, Felix Hill
- abstract: Inferring the abstract relational and causal structure of the world is a major challenge for reinforcement-learning (RL) agents. For humans, language{—}particularly in the form of explanations{—}plays a considerable role in overcoming this challenge. Here, we show that language can play a similar role for deep RL agents in complex environments. While agents typically struggle to acquire relational and causal knowledge, augmenting their experience by training them to predict language descriptions and explanations can overcome these limitations. We show that language can help agents learn challenging relational tasks, and examine which aspects of language contribute to its benefits. We then show that explanations can help agents to

infer not only relational but also causal structure. Language can shape the way that agents to generalize out-of-distribution from ambiguous, causally-confounded training, and explanations even allow agents to learn to perform experimental interventions to identify causal relationships. Our results suggest that language description and explanation may be powerful tools for improving agent learning and generalization.

Generative Cooperative Networks for Natural Language Generation

- Sylvain Lamprier, Thomas Scialom, Antoine Chaffin, Vincent Claveau, Ewa Kijak, Jacopo Staiano, Benjamin Piwowarski
- abstract: Generative Adversarial Networks (GANs) have known a tremendous success for many continuous generation tasks, especially in the field of image generation. However, for discrete outputs such as language, optimizing GANs remains an open problem with many instabilities, as no gradient can be properly back-propagated from the discriminator output to the generator parameters. An alternative is to learn the generator network via reinforcement learning, using the discriminator signal as a reward, but such a technique suffers from moving rewards and vanishing gradient problems. Finally, it often falls short compared to direct maximum-likelihood approaches. In this paper, we introduce Generative Cooperative Networks, in which the discriminator architecture is cooperatively used along with the generation policy to output samples of realistic texts for the task at hand. We give theoretical guarantees of convergence for our approach, and study various efficient decoding schemes to empirically achieve state-of-the-art results in two main NLG tasks.

DSTAGNN: Dynamic Spatial-Temporal Aware Graph Neural Network for Traffic Flow Forecasting

- Shiyong Lan, Yitong Ma, Weikang Huang, Wenwu Wang, Hongyu Yang, Pyang Li
- abstract: As a typical problem in time series analysis, traffic flow prediction is one of the most important application fields of machine learning. However, achieving highly accurate traffic flow prediction is a challenging task, due to the presence of complex dynamic spatial-temporal dependencies within a road network. This paper proposes a novel Dynamic Spatial-Temporal Aware Graph Neural Network (DSTAGNN) to model the complex spatial-temporal interaction in road network. First, considering the fact that historical data carries intrinsic dynamic information about the spatial structure of road networks, we propose a new dynamic spatial-temporal aware graph based on a data-driven strategy to replace the pre-defined static graph usually used in traditional graph convolution. Second, we design a novel graph neural network architecture, which can not only represent dynamic spatial relevance among nodes with an improved multi-head attention mechanism, but also acquire the wide range of dynamic temporal dependency from multi-receptive field features via multi-scale gated convolution. Extensive experiments on real-world data sets demonstrate that our proposed method significantly outperforms the state-of-the-art methods.

Cooperative Online Learning in Stochastic and Adversarial MDPs

- Tal Lancewicki, Aviv Rosenberg, Yishay Mansour
- abstract: We study cooperative online learning in stochastic and adversarial Markov decision process (MDP). That is, in each episode, m agents interact with an MDP simultaneously and share information in order to minimize their individual regret. We consider environments with two types of randomness: fresh – where each agent's trajectory is sampled i.i.d, and non-fresh – where the realization is shared by all agents (but each agent's trajectory is also affected by its own actions). More precisely, with non-fresh randomness the realization of every cost and transition is fixed at the start of each episode, and agents that take the same action in the same state at the same time observe the same cost and next state. We thoroughly analyze all relevant settings, highlight the challenges and differences between the models, and prove nearly-matching regret lower and upper bounds. To our knowledge, we are the first to consider cooperative reinforcement learning (RL) with either non-fresh randomness or in adversarial MDPs.

PINs: Progressive Implicit Networks for Multi-Scale Neural Representations

- Zoe Landgraf, Alexander Sorkine Hornung, Ricardo S Cabral
- abstract: Multi-layer perceptrons (MLP) have proven to be effective scene encoders when combined with higher-dimensional projections of the input, commonly referred to as positional encoding. However, scenes with a wide frequency spectrum remain a challenge: choosing high frequencies for positional encoding introduces noise in low structure areas, while low frequencies results in poor fitting of detailed regions. To address this, we propose a progressive positional encoding, exposing a hierarchical MLP structure to incremental sets of frequency encodings. Our model accurately reconstructs scenes with wide frequency bands and learns a scene representation at progressive level of detail without explicit per-level supervision. The architecture is modular: each level encodes a continuous implicit representation that can be leveraged separately for its respective resolution, meaning a smaller network for coarser reconstructions. Experiments on several 2D and 3D datasets shows improvements in reconstruction accuracy, representational capacity and training speed compared to baselines.

Co-training Improves Prompt-based Learning for Large Language Models

- Hunter Lang, Monica N Agrawal, Yoon Kim, David Sontag
- abstract: We demonstrate that co-training (Blum & Mitchell, 1998) can improve the performance of prompt-based learning by using unlabeled data. While prompting has emerged as a promising paradigm for few-shot and zero-shot learning, it is often brittle and requires much larger models compared to the standard supervised setup. We find that co-training makes it possible to improve the original prompt model and at the same time learn a smaller, downstream task-specific model. In the case where we only have partial access to a prompt model (e.g., output probabilities from GPT-3 (Brown et al., 2020)) we learn a calibration model over the prompt outputs. When we have full access to the prompt model's gradients but full finetuning remains prohibitively expensive (e.g., T0 (Sanh et al., 2021)), we learn a set of soft prompt continuous vectors to iteratively update the prompt model. We find that models trained in this manner can significantly improve performance on challenging datasets where there is currently a large gap between prompt-based learning and fully-supervised models.

Goal Misgeneralization in Deep Reinforcement Learning

- Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, David Krueger
- abstract: We study goal misgeneralization, a type of out-of-distribution robustness failure in reinforcement learning (RL). Goal misgeneralization occurs when an RL agent retains its capabilities out-of-distribution yet pursues the wrong goal. For instance, an agent might continue to competently avoid obstacles, but navigate to the wrong place. In contrast, previous works have typically focused on capability generalization failures, where an agent fails to do anything sensible at test time.We provide the first explicit empirical demonstrations of goal misgeneralization and present a partial characterization of its causes.

Marginal Tail-Adaptive Normalizing Flows

- Mike Laszkiewicz, Johannes Lederer, Asja Fischer
- abstract: Learning the tail behavior of a distribution is a notoriously difficult problem. By definition, the number of samples from the tail is small, and deep generative models, such as normalizing flows, tend to concentrate on learning the body of the distribution. In this paper, we focus on improving the ability of normalizing flows to correctly capture the tail behavior and, thus, form more accurate models. We prove that the marginal tailedness of an autoregressive flow can be controlled via the tailedness of the marginals of its base distribution. This theoretical insight leads us to a novel type of flows based on flexible base distributions and data-driven linear layers. An empirical analysis shows that the proposed method improves on the

accuracy{—}especially on the tails of the distribution{—}and is able to generate heavy-tailed data. We demonstrate its application on a weather and climate example, in which capturing the tail behavior is essential.

Bregman Proximal Langevin Monte Carlo via Bregman-Moreau Envelopes

- Tim Tsz-Kit Lau, Han Liu
- abstract: We propose efficient Langevin Monte Carlo algorithms for sampling distributions with nonsmooth convex composite potentials, which is the sum of a continuously differentiable function and a possibly nonsmooth function. We devise such algorithms leveraging recent advances in convex analysis and optimization methods involving Bregman divergences, namely the Bregman–Moreau envelopes and the Bregman proximity operators, and in the Langevin Monte Carlo algorithms reminiscent of mirror descent. The proposed algorithms extend existing Langevin Monte Carlo algorithms in two aspects—the ability to sample nonsmooth distributions with mirror descent-like algorithms, and the use of the more general Bregman–Moreau envelope in place of the Moreau envelope as a smooth approximation of the nonsmooth part of the potential. A particular case of the proposed scheme is reminiscent of the Bregman proximal gradient algorithm. The efficiency of the proposed methodology is illustrated with various sampling tasks at which existing Langevin Monte Carlo methods are known to perform poorly.

Scalable Deep Reinforcement Learning Algorithms for Mean Field Games

- Mathieu Lauriere, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Perolat, Romuald Elie, Olivier Pietquin, Matthieu Geist
- abstract: Mean Field Games (MFGs) have been introduced to efficiently approximate games with very large populations of strategic agents. Recently, the question of learning equilibria in MFGs has gained momentum, particularly using model-free reinforcement learning (RL) methods. One limiting factor to further scale up using RL is that existing algorithms to solve MFGs require the mixing of approximated quantities such as strategies or q-values. This is far from being trivial in the case of non-linear function approximation that enjoy good generalization properties, e.g. neural networks. We propose two methods to address this shortcoming. The first one learns a mixed strategy from distillation of historical data into a neural network and is applied to the Fictitious Play algorithm. The second one is an online mixing method based on regularization that does not require memorizing historical data or previous estimates. It is used to extend Online Mirror Descent. We demonstrate numerically that these methods efficiently enable the use of Deep RL algorithms to solve various MFGs. In addition, we show that these methods outperform SotA baselines from the literature.

Implicit Bias of Linear Equivariant Networks

- Hannah Lawrence, Kristian Georgiev, Andrew Dienes, Bobak T. Kiani
- abstract: Group equivariant convolutional neural networks (G-CNNs) are generalizations of convolutional neural networks (CNNs) which excel in a wide range of technical applications by explicitly encoding symmetries, such as rotations and permutations, in their architectures. Although the success of G-CNNs is driven by their explicit symmetry bias, a recent line of work has proposed that the implicit bias of training algorithms on particular architectures is key to understanding generalization for overparameterized neural nets. In this context, we show that L-layer full-width linear G-CNNs trained via gradient descent for binary classification converge to solutions with low-rank Fourier matrix coefficients, regularized by the 2/L-Schatten matrix norm. Our work strictly generalizes previous analysis on the implicit bias of linear CNNs to linear G-CNNs over all finite groups, including the challenging setting of non-commutative groups (such as permutations), as well as band-limited G-CNNs over infinite groups. We validate our theorems via experiments on a variety of groups, and empirically explore more realistic nonlinear networks, which locally capture similar regularization patterns. Finally, we provide intuitive interpretations of our Fourier space implicit regularization results in real space via uncertainty principles.

Differentially Private Maximal Information Coefficients

- John Lazarsfeld, Aaron Johnson, Emmanuel Adeniran
- abstract: The Maximal Information Coefficient (MIC) is a powerful statistic to identify dependencies between variables. However, it may be applied to sensitive data, and publishing it could leak private information. As a solution, we present algorithms to approximate MIC in a way that provides differential privacy. We show that the natural application of the classic Laplace mechanism yields insufficient accuracy. We therefore introduce the MICr statistic, which is a new MIC approximation that is more compatible with differential privacy. We prove MICr is a consistent estimator for MIC, and we provide two differentially private versions of it. We perform experiments on a variety of real and synthetic datasets. The results show that the private MICr statistics significantly outperform direct application of the Laplace mechanism. Moreover, experiments on real-world datasets show accuracy that is usable when the sample size is at least moderately large.

Entropic Gromov-Wasserstein between Gaussian Distributions

- Khang Le, Dung Q Le, Huy Nguyen, Dat Do, Tung Pham, Nhat Ho
- abstract: We study the entropic Gromov-Wasserstein and its unbalanced version between (unbalanced) Gaussian distributions with different dimensions. When the metric is the inner product, which we refer to as inner product Gromov-Wasserstein (IGW), we demonstrate that the optimal transportation plans of entropic IGW and its unbalanced variant are (unbalanced) Gaussian distributions. Via an application of von Neumann’s trace inequality, we obtain closed-form expressions for the entropic IGW between these Gaussian distributions. Finally, we consider an entropic inner product Gromov-Wasserstein barycenter of multiple Gaussian distributions. We prove that the barycenter is a Gaussian distribution when the entropic regularization parameter is small. We further derive a closed-form expression for the covariance matrix of the barycenter.

Neurocoder: General-Purpose Computation Using Stored Neural Programs

- Hung Le, Svetha Venkatesh
- abstract: Artificial Neural Networks are functionally equivalent to special-purpose computers. Their inter-neuronal connection weights represent the learnt Neural Program that instructs the networks on how to compute the data. However, without storing Neural Programs, they are restricted to only one, overwriting learnt programs when trained on new data. Here we design Neurocoder, a new class of general-purpose neural networks in which the neural network “codes” itself in a data-responsive way by composing relevant programs from a set of shareable, modular programs stored in external memory. This time, a Neural Program is efficiently treated as data in memory. Integrating Neurocoder into current neural architectures, we demonstrate new capacity to learn modular programs, reuse simple programs to build complex ones, handle pattern shifts and remember old programs as new ones are learnt, and show substantial performance improvement in solving object recognition, playing video games and continual learning tasks.

Convergence of Policy Gradient for Entropy Regularized MDPs with Neural Network Approximation in the Mean-Field Regime

- James-Michael Leahy, Bekzhan Kerimkulov, David Siska, Lukasz Szpruch
- abstract: We study the global convergence of policy gradient for infinite-horizon, continuous state and action space, and entropy-regularized Markov decision processes (MDPs). We consider a softmax policy with (one-hidden layer) neural network approximation in a mean-field regime. Additional entropic regularization in the associated mean-field probability measure is added, and the corresponding gradient flow is studied in the 2-Wasserstein metric. We show that the objective function is increasing along the gradient flow. Further, we prove that if the regularization in terms of the mean-field measure is sufficient, the gradient flow converges exponentially fast to the unique stationary solution, which is the unique maximizer of the regularized

MDP objective. Lastly, we study the sensitivity of the value function along the gradient flow with respect to regularization parameters and the initial condition. Our results rely on the careful analysis of the non-linear Fokker–Planck–Kolmogorov equation and extend the pioneering work of \cite{mei2020global} and \cite{agarwal2020optimality}, which quantify the global convergence rate of policy gradient for entropy-regularized MDPs in the tabular setting.

A Random Matrix Analysis of Data Stream Clustering: Coping With Limited Memory Resources

- Hugo Lebeau, Romain Couillet, Florent Chatelain
- abstract: This article introduces a random matrix framework for the analysis of clustering on high-dimensional data streams, a particularly relevant setting for a more sober processing of large amounts of data with limited memory and energy resources. Assuming data $\mathbf{x}*1, \mathbf{x}_2, \dots$ arrives as a continuous flow and a small number L of them can be kept in the learning pipeline, one has only access to the diagonal elements of the Gram kernel matrix: $\left[\mathbf{K}_L \right]*{i, j} = \frac{1}{p} \mathbf{x}*i^\top \mathbf{x}_j \mathbf{1}*{\left\lvert i - j \right\rvert < L}$. Under a large-dimensional data regime, we derive the limiting spectral distribution of the banded kernel matrix $\mathbf{K}_L$ and study its isolated eigenvalues and eigenvectors, which behave in an unfamiliar way. We detail how these results can be used to perform efficient online kernel spectral clustering and provide theoretical performance guarantees. Our findings are empirically confirmed on image clustering tasks. Leveraging on optimality results of spectral methods for clustering, this work offers insights on efficient online clustering techniques for high-dimensional data.

Neural Tangent Kernel Analysis of Deep Narrow Neural Networks

- Jongmin Lee, Joo Young Choi, Ernest K Ryu, Albert No
- abstract: The tremendous recent progress in analyzing the training dynamics of overparameterized neural networks has primarily focused on wide networks and therefore does not sufficiently address the role of depth in deep learning. In this work, we present the first trainability guarantee of infinitely deep but narrow neural networks. We study the infinite-depth limit of a multilayer perceptron (MLP) with a specific initialization and establish a trainability guarantee using the NTK theory. We then extend the analysis to an infinitely deep convolutional neural network (CNN) and perform brief experiments.

Dataset Condensation with Contrastive Signals

- Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, Sungroh Yoon
- abstract: Recent studies have demonstrated that gradient matching-based dataset synthesis, or dataset condensation (DC), methods can achieve state-of-theart performance when applied to data-efficient learning tasks. However, in this study, we prove that the existing DC methods can perform worse than the random selection method when taskirrelevant information forms a significant part of the training dataset. We attribute this to the lack of participation of the contrastive signals between the classes resulting from the class-wise gradient matching strategy. To address this problem, we propose Dataset Condensation with Contrastive signals (DCC) by modifying the loss function to enable the DC methods to effectively capture the differences between classes. In addition, we analyze the new loss function in terms of training dynamics by tracking the kernel velocity. Furthermore, we introduce a bi-level warm-up strategy to stabilize the optimization. Our experimental results indicate that while the existing methods are ineffective for fine-grained image classification tasks, the proposed method can successfully generate informative synthetic datasets for the same tasks. Moreover, we demonstrate that the proposed method outperforms the baselines even on benchmark datasets such as SVHN, CIFAR-10, and CIFAR-100. Finally, we demonstrate the high applicability of the proposed method by applying it to continual learning tasks.

Confidence Score for Source-Free Unsupervised Domain Adaptation

- Jonghyun Lee, Dahuin Jung, Junho Yim, Sungroh Yoon
- abstract: Source-free unsupervised domain adaptation (SFUDA) aims to obtain high performance in the unlabeled target domain using the pre-trained source model, not the source data. Existing SFUDA methods assign the same importance to all target samples, which is vulnerable to incorrect pseudo-labels. To differentiate between sample importance, in this study, we propose a novel sample-wise confidence score, the Joint Model-Data Structure (JMDS) score for SFUDA. Unlike existing confidence scores that use only one of the source or target domain knowledge, the JMDS score uses both knowledge. We then propose a Confidence score Weighting Adaptation using the JMDS (CoWA-JMDS) framework for SFUDA. CoWA-JMDS consists of the JMDS scores as sample weights and weight Mixup that is our proposed variant of Mixup. Weight Mixup promotes the model make more use of the target domain knowledge. The experimental results show that the JMDS score outperforms the existing confidence scores. Moreover, CoWA-JMDS achieves state-of-the-art performance on various SFUDA scenarios: closed, open, and partial-set scenarios.

A Statistical Manifold Framework for Point Cloud Data

- Yonghyeon Lee, Seungyeon Kim, Jinwon Choi, Frank Park
- abstract: Many problems in machine learning involve data sets in which each data point is a point cloud in $\mathbb{R}^D$. A growing number of applications require a means of measuring not only distances between point clouds, but also angles, volumes, derivatives, and other more advanced concepts. To formulate and quantify these concepts in a coordinate-invariant way, we develop a Riemannian geometric framework for point cloud data. By interpreting each point in a point cloud as a sample drawn from some given underlying probability density, the space of point cloud data can be given the structure of a statistical manifold – each point on this manifold represents a point cloud – with the Fisher information metric acting as a natural Riemannian metric. Two autoencoder applications of our framework are presented: (i) smoothly deforming one 3D object into another via interpolation between the two corresponding point clouds; (ii) learning an optimal set of latent space coordinates for point cloud data that best preserves angles and distances, and thus produces a more discriminative representation space. Experiments with large-scale standard benchmark point cloud data show greatly improved classification accuracy vis-á-vis existing methods. Code is available at https://github.com/seungyeon-k/SMF-public.

Low-Complexity Deep Convolutional Neural Networks on Fully Homomorphic Encryption Using Multiplexed Parallel Convolutions

- Eunsang Lee, Joon-Woo Lee, Junghyun Lee, Young-Sik Kim, Yongjune Kim, Jong-Seon No, Woosuk Choi
- abstract: Recently, the standard ResNet-20 network was successfully implemented on the fully homomorphic encryption scheme, residue number system variant Cheon-Kim-Kim-Song (RNS-CKKS) scheme using bootstrapping, but the implementation lacks practicality due to high latency and low security level. To improve the performance, we first minimize total bootstrapping runtime using multiplexed parallel convolution that collects sparse output data for multiple channels compactly. We also propose the imaginary-removing bootstrapping to prevent the deep neural networks from catastrophic divergence during approximate ReLU operations. In addition, we optimize level consumptions and use lighter and tighter parameters. Simulation results show that we have 4.67x lower inference latency and 134x less amortized runtime (runtime per image) for ResNet-20 compared to the state-of-the-art previous work, and we achieve standard 128-bit security. Furthermore, we successfully implement ResNet-110 with high accuracy on the RNS-CKKS scheme for the first time.

Statistical inference with implicit SGD: proximal Robbins-Monro vs. Polyak-Ruppert

- Yoonhyung Lee, Sungdong Lee, Joong-Ho Won

- abstract: The implicit stochastic gradient descent (ISGD), a proximal version of SGD, is gaining interest in the literature due to its stability over (explicit) SGD. In this paper, we conduct an in-depth analysis of the two modes of ISGD for smooth convex functions, namely proximal Robbins-Monro (proxRM) and proximal Poylak-Ruppert (proxPR) procedures, for their use in statistical inference on model parameters. Specifically, we derive non-asymptotic point estimation error bounds of both proxRM and proxPR iterates and their limiting distributions, and propose on-line estimators of their asymptotic covariance matrices that require only a single run of ISGD. The latter estimators are used to construct valid confidence intervals for the model parameters. Our analysis is free of the generalized linear model assumption that has limited the preceding analyses, and employs feasible procedures. Our on-line covariance matrix estimators appear to be the first of this kind in the ISGD literature.

Maslow's Hammer in Catastrophic Forgetting: Node Re-Use vs. Node Activation

- Sebastian Lee, Stefano Sarao Mannelli, Claudia Clopath, Sebastian Goldt, Andrew Saxe
- abstract: Continual learning—learning new tasks in sequence while maintaining performance on old tasks—remains particularly challenging for artificial neural networks. Surprisingly, the amount of forgetting does not increase with the dissimilarity between the learned tasks, but appears to be worst in an intermediate similarity regime. In this paper we theoretically analyse both a synthetic teacher-student framework and a real data setup to provide an explanation of this phenomenon that we name Maslow's Hammer hypothesis. Our analysis reveals the presence of a trade-off between node activation and node re-use that results in worst forgetting in the intermediate regime. Using this understanding we reinterpret popular algorithmic interventions for catastrophic interference in terms of this trade-off, and identify the regimes in which they are most effective.

Query-Efficient and Scalable Black-Box Adversarial Attacks on Discrete Sequential Data via Bayesian Optimization

- Deokjae Lee, Seungyong Moon, Junhyeok Lee, Hyun Oh Song
- abstract: We focus on the problem of adversarial attacks against models on discrete sequential data in the black-box setting where the attacker aims to craft adversarial examples with limited query access to the victim model. Existing black-box attacks, mostly based on greedy algorithms, find adversarial examples using pre-computed key positions to perturb, which severely limits the search space and might result in suboptimal solutions. To this end, we propose a query-efficient black-box attack using Bayesian optimization, which dynamically computes important positions using an automatic relevance determination (ARD) categorical kernel. We introduce block decomposition and history subsampling techniques to improve the scalability of Bayesian optimization when an input sequence becomes long. Moreover, we develop a post-optimization algorithm that finds adversarial examples with smaller perturbation size. Experiments on natural language and protein classification tasks demonstrate that our method consistently achieves higher attack success rate with significant reduction in query count and modification rate compared to the previous state-of-the-art methods.

Least Squares Estimation using Sketched Data with Heteroskedastic Errors

- Sokbae Lee, Serena Ng
- abstract: Researchers may perform regressions using a sketch of data of size m instead of the full sample of size n for a variety of reasons. This paper considers the case when the regression errors do not have constant variance and heteroskedasticity robust standard errors would normally be needed for test statistics to provide accurate inference. We show that estimates using data sketched by random projections will behave 'as if' the errors were homoskedastic. Estimation by random sampling would not have this property. The result arises because the sketched estimates in the case of random projections can be expressed as degenerate U-statistics, and under certain conditions, these statistics are asymptotically normal with homoskedastic variance. We verify that the conditions hold not only in the case of least squares regression when the covariates are exogenous, but also in instrumental variables estimation when the covariates are endogenous. The result implies that inference can be simpler than the full sample case if the sketching scheme is appropriately chosen.

Why the Rich Get Richer? On the Balancedness of Random Partition Models

- Changwoo J Lee, Huiyan Sang
- abstract: Random partition models are widely used in Bayesian methods for various clustering tasks, such as mixture models, topic models, and community detection problems. While the number of clusters induced by random partition models has been studied extensively, another important model property regarding the balancedness of partition has been largely neglected. We formulate a framework to define and theoretically study the balancedness of exchangeable random partition models, by analyzing how a model assigns probabilities to partitions with different levels of balancedness. We demonstrate that the "rich-get-richer" characteristic of many existing popular random partition models is an inevitable consequence of two common assumptions: product-form exchangeability and projectivity. We propose a principled way to compare the balancedness of random partition models, which gives a better understanding of what model works better and what doesn't for different applications. We also introduce the "rich-get-poorer" random partition models and illustrate their application to entity resolution tasks.

Model Selection in Batch Policy Optimization

- Jonathan Lee, George Tucker, Ofir Nachum, Bo Dai
- abstract: We study the problem of model selection in batch policy optimization: given a fixed, partial-feedback dataset and M model classes, learn a policy with performance that is competitive with the policy derived from the best model class. We formalize the problem in the contextual bandit setting with linear model classes by identifying three sources of error that any model selection algorithm should optimally trade-off in order to be competitive: (1) approximation error, (2) statistical complexity, and (3) coverage. The first two sources are common in model selection for supervised learning, where optimally trading off these two is well-studied. In contrast, the third source is unique to batch policy optimization and is due to dataset shift inherent to the setting. We first show that no batch policy optimization algorithm can achieve a guarantee addressing all three simultaneously, revealing a stark contrast between difficulties in batch policy optimization and the positive results available in supervised learning. Despite this negative result, we show that relaxing any one of the three error sources enables the design of algorithms achieving near-oracle inequalities for the remaining two. We conclude with experiments demonstrating the efficacy of these algorithms.

Supervised Learning with General Risk Functionals

- Liu Leqi, Audrey Huang, Zachary Lipton, Kamyar Azizzadenesheli
- abstract: Standard uniform convergence results bound the generalization gap of the expected loss over a hypothesis class. The emergence of risk-sensitive learning requires generalization guarantees for functionals of the loss distribution beyond the expectation. While prior works specialize in uniform convergence of particular functionals, our work provides uniform convergence for a general class of Hölder risk functionals for which the closeness in the Cumulative Distribution Function (CDF) entails closeness in risk. We establish the first uniform convergence results for estimating the CDF of the loss distribution, which yield uniform convergence guarantees that hold simultaneously both over a class of Hölder risk functionals and over a hypothesis class. Thus licensed to perform empirical risk minimization, we develop practical gradient-based methods for minimizing distortion risks (widely studied subset of Hölder risks that subsumes the spectral risks, including the mean, conditional value at risk, cumulative prospect theory risks, and others) and provide convergence guarantees. In experiments, we demonstrate the efficacy of our learning procedure, both in settings where uniform convergence results hold and in high-dimensional settings with deep networks.

Generalized Strategic Classification and the Case of Aligned Incentives

- Sagi Levanon, Nir Rosenfeld
- abstract: Strategic classification studies learning in settings where self-interested users can strategically modify their features to obtain favorable predictive outcomes. A key working assumption, however, is that “favorable” always means “positive”; this may be appropriate in some applications (e.g., loan approval), but reduces to a fairly narrow view of what user interests can be. In this work we argue for a broader perspective on what accounts for strategic user behavior, and propose and study a flexible model of generalized strategic classification. Our generalized model subsumes most current models but includes other novel settings; among these, we identify and target one intriguing sub-class of problems in which the interests of users and the system are aligned. This setting reveals a surprising fact: that standard max-margin losses are ill-suited for strategic inputs. Returning to our fully generalized model, we propose a novel max-margin framework for strategic learning that is practical and effective, and which we analyze theoretically. We conclude with a set of experiments that empirically demonstrate the utility of our approach.

A Simple Unified Framework for High Dimensional Bandit Problems

- Wenjie Li, Adarsh Barik, Jean Honorio
- abstract: Stochastic high dimensional bandit problems with low dimensional structures are useful in different applications such as online advertising and drug discovery. In this work, we propose a simple unified algorithm for such problems and present a general analysis framework for the regret upper bound of our algorithm. We show that under some mild unified assumptions, our algorithm can be applied to different high-dimensional bandit problems. Our framework utilizes the low dimensional structure to guide the parameter estimation in the problem, therefore our algorithm achieves the comparable regret bounds in the LASSO bandit as a sanity check, as well as novel bounds that depend logarithmically on dimensions in the low-rank matrix bandit, the group sparse matrix bandit, and in a new problem: the multi-agent LASSO bandit.

Robust Training of Neural Networks Using Scale Invariant Architectures

- Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, Sanjiv Kumar
- abstract: In contrast to SGD, adaptive gradient methods like Adam allow robust training of modern deep networks, especially large language models. However, the use of adaptivity not only comes at the cost of extra memory but also raises the fundamental question: can non-adaptive methods like SGD enjoy similar benefits? In this paper, we provide an affirmative answer to this question by proposing to achieve both robust and memory-efficient training via the following general recipe: (1) modify the architecture and make it scale invariant, (2) train with SGD and weight decay, and optionally (3) clip the global gradient norm proportional to weight norm multiplied by $\sqrt{\frac{2\lambda}{\eta}}$, where η is learning rate and λ is weight decay. We show that this general approach is robust to rescaling of parameter and loss by proving that its convergence only depends logarithmically on the scale of initialization and loss, whereas the standard SGD might not even converge for many initializations. Following our recipe, we design a scale invariant version of BERT, called SIBERT, which when trained simply by vanilla SGD achieves performance comparable to BERT trained by adaptive methods like Adam on downstream tasks.

Spatial-Channel Token Distillation for Vision MLPs

- Yanxi Li, Xinghao Chen, Minjing Dong, Yehui Tang, Yunhe Wang, Chang Xu
- abstract: Recently, neural architectures with all Multi-layer Perceptrons (MLPs) have attracted great research interest from the computer vision community. However, the inefficient mixing of spatial-channel information causes MLP-like vision models to demand tremendous pre-training on large-scale datasets. This work solves the problem from a novel knowledge distillation perspective. We propose a novel Spatial-channel Token Distillation (STD) method, which improves the information mixing in the two dimensions by introducing distillation tokens to each of them. A mutual information regularization is further introduced to let distillation tokens focus on their specific dimensions and maximize the performance gain. Extensive experiments on ImageNet for several MLP-like architectures demonstrate that the proposed token distillation mechanism can efficiently improve the accuracy. For example, the proposed STD boosts the top-1 accuracy of Mixer-S16 on ImageNet from 73.8% to 75.7% without any costly pre-training on JFT-300M. When applied to stronger architectures, e.g. CycleMLP-B1 and CycleMLP-B2, STD can still harvest about 1.1% and 0.5% accuracy gains, respectively.

An Analytical Update Rule for General Policy Optimization

- Hepeng Li, Nicholas Clavette, Haibo He
- abstract: We present an analytical policy update rule that is independent of parametric function approximators. The policy update rule is suitable for optimizing general stochastic policies and has a monotonic improvement guarantee. It is derived from a closed-form solution to trust-region optimization using calculus of variation, following a new theoretical result that tightens existing bounds for policy improvement using trust-region methods. The update rule builds a connection between policy search methods and value function methods. Moreover, off-policy reinforcement learning algorithms can be derived from the update rule since it does not need to compute integration over on-policy states. In addition, the update rule extends immediately to cooperative multi-agent systems when policy updates are performed by one agent at a time.

On Convergence of Gradient Descent Ascent: A Tight Local Analysis

- Haochuan Li, Farzan Farnia, Subhro Das, Ali Jadbabaie
- abstract: Gradient Descent Ascent (GDA) methods are the mainstream algorithms for minimax optimization in generative adversarial networks (GANs). Convergence properties of GDA have drawn significant interest in the recent literature. Specifically, for $\min_{x} \max_{y} f(x;y)$ where f is strongly-concave in y and possibly nonconvex in x, (Lin et al., 2020) proved the convergence of GDA with a stepsize ratio $\eta_y/\eta_x=\Theta(\kappa^2)$ where η_x and η_y are the stepsizes for x and y and κ is the condition number for y. While this stepsize ratio suggests a slow training of the min player, practical GAN algorithms typically adopt similar stepsizes for both variables, indicating a wide gap between theoretical and empirical results. In this paper, we aim to bridge this gap by analyzing the local convergence of general nonconvex-nonconcave minimax problems. We demonstrate that a stepsize ratio of $\Theta(\kappa)$ is necessary and sufficient for local convergence of GDA to a Stackelberg Equilibrium, where κ is the local condition number for y. We prove a nearly tight convergence rate with a matching lower bound. We further extend the convergence guarantees to stochastic GDA and extra-gradient methods (EG). Finally, we conduct several numerical experiments to support our theoretical findings.

On the Finite-Time Performance of the Knowledge Gradient Algorithm

- Yanwen Li, Siyang Gao
- abstract: The knowledge gradient (KG) algorithm is a popular and effective algorithm for the best arm identification (BAI) problem. Due to the complex calculation of KG, theoretical analysis of this algorithm is difficult, and existing results are mostly about the asymptotic performance of it, e.g., consistency, asymptotic sample allocation, etc. In this research, we present new theoretical results about the finite-time performance of the KG algorithm. Under independent and normally distributed rewards, we derive lower bounds and upper bounds for the probability of error and simple regret of the algorithm. With these bounds, existing asymptotic results become simple corollaries. We also show the performance of the algorithm for the multi-armed bandit (MAB) problem. These developments not only extend the existing analysis of the KG algorithm, but can also be used to analyze other improvement-based algorithms. Last, we use numerical experiments to further demonstrate the finite-time behavior of the KG algorithm.

Phasic Self-Imitative Reduction for Sparse-Reward Goal-Conditioned Reinforcement Learning

- Yunfei Li, Tian Gao, Jiaqi Yang, Huazhe Xu, Yi Wu
- abstract: It has been a recent trend to leverage the power of supervised learning (SL) towards more effective reinforcement learning (RL) methods. We propose a novel phasic solution by alternating online RL and offline SL for tackling sparse-reward goal-conditioned problems. In the online phase, we perform RL training and collect rollout data while in the offline phase, we perform SL on those successful trajectories from the dataset. To further improve sample efficiency, we adopt additional techniques in the online phase including task reduction to generate more feasible trajectories and a value-difference-based intrinsic reward to alleviate the sparse-reward issue. We call this overall framework, PhAsic self-Imitative Reduction (PAIR). PAIR is compatible with various online and offline RL methods and substantially outperforms both non-phasic RL and phasic SL baselines on sparse-reward robotic control problems, including a particularly challenging stacking task. PAIR is the first RL method that learns to stack 6 cubes with only 0/1 success rewards from scratch.

G^{2}CN: Graph Gaussian Convolution Networks with Concentrated Graph Filters

- Mingjie Li, Xiaojun Guo, Yifei Wang, Yisen Wang, Zhouchen Lin
- abstract: Recently, linear GCNs have shown competitive performance against non-linear ones with less computation cost, and the key lies in their propagation layers. Spectral analysis has been widely adopted in designing and analyzing existing graph propagations. Nevertheless, we notice that existing spectral analysis fails to explain why existing graph propagations with the same global tendency, such as low-pass or high-pass, still yield very different results. Motivated by this situation, we develop a new framework for spectral analysis in this paper called concentration analysis. In particular, we propose three attributes: concentration centre, maximum response, and bandwidth for our analysis. Through a dissection of the limitations of existing graph propagations via the above analysis, we propose a new kind of propagation layer, Graph Gaussian Convolution Networks (G^2CN), in which the three properties are decoupled and the whole structure becomes more flexible and applicable to different kinds of graphs. Extensive experiments show that we can obtain state-of-the-art performance on heterophily and homophily datasets with our proposed G^2CN.

Decomposing Temporal High-Order Interactions via Latent ODEs

- Shibo Li, Robert Kirby, Shandian Zhe
- abstract: High-order interactions between multiple objects are common in real-world applications. Although tensor decomposition is a popular framework for high-order interaction analysis and prediction, most methods cannot well exploit the valuable timestamp information in data. The existent methods either discard the timestamps or convert them into discrete steps or use over-simplistic decomposition models. As a result, these methods might not be capable enough of capturing complex, fine-grained temporal dynamics or making accurate predictions for long-term interaction results. To overcome these limitations, we propose a novel Temporal High-order Interaction decompoSition model based on Ordinary Differential Equations (THIS-ODE). We model the time-varying interaction result with a latent ODE. To capture the complex temporal dynamics, we use a neural network (NN) to learn the time derivative of the ODE state. We use the representation of the interaction objects to model the initial value of the ODE and to constitute a part of the NN input to compute the state. In this way, the temporal relationships of the participant objects can be estimated and encoded into their representations. For tractable and scalable inference, we use forward sensitivity analysis to efficiently compute the gradient of ODE state, based on which we use integral transform to develop a stochastic mini-batch learning algorithm. We demonstrate the advantage of our approach in simulation and four real-world applications.

Neural Inverse Transform Sampler

- Henry Li, Yuval Kluger
- abstract: Any explicit functional representation f of a density is hampered by two main obstacles when we wish to use it as a generative model: designing f so that sampling is fast, and estimating $Z = \int f$ so that $Z^{-1}f$ integrates to 1. This becomes increasingly complicated as f itself becomes complicated. In this paper, we show that when modeling one-dimensional conditional densities with a neural network, Z can be exactly and efficiently computed by letting the network represent the cumulative distribution function of a target density, and applying a generalized fundamental theorem of calculus. We also derive a fast algorithm for sampling from the resulting representation by the inverse transform method. By extending these principles to higher dimensions, we introduce the \textbf{Neural Inverse Transform Sampler (NITS)}, a novel deep learning framework for modeling and sampling from general, multidimensional, compactly-supported probability densities. NITS is a highly expressive density estimator that boasts end-to-end differentiability, fast sampling, and exact and cheap likelihood evaluation. We demonstrate the applicability of NITS by applying it to realistic, high-dimensional density estimation tasks: likelihood-based generative modeling on the CIFAR-10 dataset, and density estimation on the UCI suite of benchmark datasets, where NITS produces compelling results rivaling or surpassing the state of the art.

PLATINUM: Semi-Supervised Model Agnostic Meta-Learning using Submodular Mutual Information

- Changbin Li, Suraj Kothawade, Feng Chen, Rishabh Iyer
- abstract: Few-shot classification (FSC) requires training models using a few (typically one to five) data points per class. Meta-learning has proven to be able to learn a parametrized model for FSC by training on various other classification tasks. In this work, we propose PLATINUM (semi-suPervised modeL Agnostic meTa learnIng usiNg sUbmodular Mutual information), a novel semi-supervised model agnostic meta learning framework that uses the submodular mutual in- formation (SMI) functions to boost the perfor- mance of FSC. PLATINUM leverages unlabeled data in the inner and outer loop using SMI func- tions during meta-training and obtains richer meta- learned parameterizations. We study the per- formance of PLATINUM in two scenarios - 1) where the unlabeled data points belong to the same set of classes as the labeled set of a cer- tain episode, and 2) where there exist out-of-distribution classes that do not belong to the la- beled set. We evaluate our method on various settings on the miniImageNet, tieredImageNet and CIFAR-FS datasets. Our experiments show that PLATINUM outperforms MAML and semi- supervised approaches like pseduo-labeling for semi-supervised FSC, especially for small ratio of labeled to unlabeled samples.

Deconfounded Value Decomposition for Multi-Agent Reinforcement Learning

- Jiahui Li, Kun Kuang, Baoxiang Wang, Furui Liu, Long Chen, Changjie Fan, Fei Wu, Jun Xiao
- abstract: Value decomposition (VD) methods have been widely used in cooperative multi-agent reinforcement learning (MARL), where credit assignment plays an important role in guiding the agents' decentralized execution. In this paper, we investigate VD from a novel perspective of causal inference. We first show that the environment in existing VD methods is an unobserved confounder as the common cause factor of the global state and the joint value function, which leads to the confounding bias on learning credit assignment. We then present our approach, deconfounded value decomposition (DVD), which cuts off the backdoor confounding path from the global state to the joint value function. The cut is implemented by introducing the trajectory graph, which depends only on the local trajectories, as a proxy confounder. DVD is general enough to be applied to various VD methods, and extensive experiments show that DVD can consistently achieve significant performance gains over different state-of-the-art VD methods on StarCraft II and MACO benchmarks.

C-MinHash: Improving Minwise Hashing with Circulant Permutation

- Xiaoyun Li, Ping Li
- abstract: Minwise hashing (MinHash) is an important and practical algorithm for generating random hashes to approximate the Jaccard (resemblance) similarity in massive binary (0/1) data. The basic theory of MinHash requires applying hundreds or even thousands of independent random permutations to each data vector in the dataset, in order to obtain reliable results for (e.g.,) building large-scale learning models or approximate near neighbor search. In

this paper, we propose Circulant MinHash (C-MinHash) and provide the surprising theoretical results that using only two independent random permutations in a circulant manner leads to uniformly smaller Jaccard estimation variance than that of the classical MinHash with K independent permutations. Experiments are conducted to show the effectiveness of the proposed method. We also propose a more convenient C-MinHash variant which reduces two permutations to just one, with extensive numerical results to validate that it achieves essentially the same estimation accuracy as using two permutations.

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

- Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi
- abstract: Vision-Language Pre-training (VLP) has advanced the performance for many vision-language tasks. However, most existing pre-trained models only excel in either understanding-based tasks or generation-based tasks. Furthermore, performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web, which is a suboptimal source of supervision. In this paper, we propose BLIP, a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones. We achieve state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval (+2.7% in average recall@1), image captioning (+2.8% in CIDEr), and VQA (+1.6% in VQA score). BLIP also demonstrates strong generalization ability when directly transferred to video-language tasks in a zero-shot manner. Code and models are available at https://github.com/salesforce/BLIP.

Restarted Nonconvex Accelerated Gradient Descent: No More Polylogarithmic Factor in the $O(ε^-7/4)$ Complexity

- Huan Li, Zhouchen Lin
- abstract: This paper studies the accelerated gradient descent for general nonconvex problems under the gradient Lipschitz and Hessian Lipschitz assumptions. We establish that a simple restarted accelerated gradient descent (AGD) finds an ϵ-approximate first-order stationary point in $O(\epsilon^{-7/4})$ gradient computations with simple proofs. Our complexity does not hide any polylogarithmic factors, and thus it improves over the state-of-the-art one by the $O(\log\frac{1}{\epsilon})$ factor. Our simple algorithm only consists of Nesterov's classical AGD and a restart mechanism, and it does not need the negative curvature exploitation or the optimization of regularized surrogate functions. Technically, our simple proof does not invoke the analysis for the strongly convex AGD, which is crucial to remove the $O(\log\frac{1}{\epsilon})$ factor.

Achieving Fairness at No Utility Cost via Data Reweighing with Influence

- Peizhao Li, Hongfu Liu
- abstract: With the fast development of algorithmic governance, fairness has become a compulsory property for machine learning models to suppress unintentional discrimination. In this paper, we focus on the pre-processing aspect for achieving fairness, and propose a data reweighing approach that only adjusts the weight for samples in the training phase. Different from most previous reweighing methods which usually assign a uniform weight for each (sub)group, we granularly model the influence of each training sample with regard to fairness-related quantity and predictive utility, and compute individual weights based on influence under the constraints from both fairness and utility. Experimental results reveal that previous methods achieve fairness at a non-negligible cost of utility, while as a significant advantage, our approach can empirically release the tradeoff and obtain cost-free fairness for equal opportunity. We demonstrate the cost-free fairness through vanilla classifiers and standard training processes, compared to baseline methods on multiple real-world tabular datasets. Code available at https://github.com/brandeis-machine-learning/influence-fairness.

High Probability Guarantees for Nonconvex Stochastic Gradient Descent with Heavy Tails

- Shaojie Li, Yong Liu
- abstract: Stochastic gradient descent (SGD) is the workhorse in modern machine learning and data-driven optimization. Despite its popularity, existing theoretical guarantees for SGD are mainly derived in expectation and for convex learning problems. High probability guarantees of nonconvex SGD are scarce, and typically rely on "light-tail" noise assumptions and study the optimization and generalization performance separately. In this paper, we develop high probability bounds for nonconvex SGD with a joint perspective of optimization and generalization performance. Instead of the light tail assumption, we consider the gradient noise following a heavy-tailed sub-Weibull distribution, a novel class generalizing the sub-Gaussian and sub-Exponential families to potentially heavier-tailed distributions. Under these complicated settings, we first present high probability bounds with best-known rates in general nonconvex learning, then move to nonconvex learning with a gradient dominance curvature condition, for which we improve the learning guarantees to fast rates. We further obtain sharper learning guarantees by considering a mild Bernstein-type noise condition. Our analysis also reveals the effect of trade-offs between the optimization and generalization performance under different conditions. In the last, we show that gradient clipping can be employed to remove the bounded gradient-type assumptions. Additionally, in this case, the stepsize of SGD is completely oblivious to the knowledge of smoothness.

MetAug: Contrastive Learning via Meta Feature Augmentation

- Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, Hui Xiong
- abstract: What matters for contrastive learning? We argue that contrastive learning heavily relies on informative features, or "hard" (positive or negative) features. Early works include more informative features by applying complex data augmentations and large batch size or memory bank, and recent works design elaborate sampling approaches to explore informative features. The key challenge toward exploring such features is that the source multi-view data is generated by applying random data augmentations, making it infeasible to always add useful information in the augmented data. Consequently, the informativeness of features learned from such augmented data is limited. In response, we propose to directly augment the features in latent space, thereby learning discriminative representations without a large amount of input data. We perform a meta learning technique to build the augmentation generator that updates its network parameters by considering the performance of the encoder. However, insufficient input data may lead the encoder to learn collapsed features and therefore malfunction the augmentation generator. A new margin-injected regularization is further added in the objective function to avoid the encoder learning a degenerate mapping. To contrast all features in one gradient back-propagation step, we adopt the proposed optimization-driven unified contrastive loss instead of the conventional contrastive loss. Empirically, our method achieves state-of-the-art results on several benchmark datasets.

PMIC: Improving Multi-Agent Reinforcement Learning with Progressive Mutual Information Collaboration

- Pengyi Li, Hongyao Tang, Tianpei Yang, Xiaotian Hao, Tong Sang, Yan Zheng, Jianye Hao, Matthew E. Taylor, Wenyuan Tao, Zhen Wang
- abstract: Learning to collaborate is critical in Multi-Agent Reinforcement Learning (MARL). Previous works promote collaboration by maximizing the correlation of agents' behaviors, which is typically characterized by Mutual Information (MI) in different forms. However, we reveal sub-optimal collaborative behaviors also emerge with strong correlations, and simply maximizing the MI can, surprisingly, hinder the learning towards better collaboration. To address this issue, we propose a novel MARL framework, called Progressive Mutual Information Collaboration (PMIC), for more effective MI-driven collaboration. PMIC uses a new collaboration criterion measured by the MI between global states and joint actions. Based on this criterion, the key idea of PMIC is maximizing the MI associated with superior collaborative behaviors and minimizing the MI associated with inferior ones. The two MI objectives play complementary roles by facilitating better collaborations while avoiding falling into sub-optimal ones. Experiments on a wide range of MARL benchmarks show the superior performance of PMIC compared with other algorithms.

CerDEQ: Certifiable Deep Equilibrium Model

- Mingjie Li, Yisen Wang, Zhouchen Lin
- abstract: Recently, certifiable robust training methods via bound propagation have been proposed for training neural networks with certifiable robustness guarantees. However, no neural architectures with regular convolution and linear layers perform better in the certifiable training than the plain CNNs, since the output bounds for the deep explicit models increase quickly as their depth increases. And such a phenomenon significantly hinders certifiable training. Meanwhile, the Deep Equilibrium Model (DEQ) is more representative and robust due to their equivalent infinite depth and controllable global Lipschitz. But no work has been proposed to explore whether DEQ can show advantages in certified training. In this work, we aim to tackle the problem of DEQ's certified training. To obtain the output bound based on the bound propagation scheme in the implicit model, we first involve the adjoint DEQ for bound approximation. Furthermore, we also use the weight orthogonalization method and other tricks specified for DEQ to stabilize the certifiable training. With our approach, we can obtain the certifiable DEQ called CerDEQ. Our CerDEQ can achieve state-of-the-art performance compared with models using regular convolution and linear layers on ℓ_∞ tasks with $\epsilon=8/255$: $64.72%$ certified error for CIFAR-10 and $94.45%$ certified error for Tiny ImageNet.

Generalization Guarantee of Training Graph Convolutional Networks with Graph Topology Sampling

- Hongkang Li, Meng Wang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong
- abstract: Graph convolutional networks (GCNs) have recently achieved great empirical success in learning graph-structured data. To address its scalability issue due to the recursive embedding of neighboring features, graph topology sampling has been proposed to reduce the memory and computational cost of training GCNs, and it has achieved comparable test performance to those without topology sampling in many empirical studies. To the best of our knowledge, this paper provides the first theoretical justification of graph topology sampling in training (up to) three-layer GCNs for semi-supervised node classification. We formally characterize some sufficient conditions on graph topology sampling such that GCN training leads to diminishing generalization error. Moreover, our method tackles the non-convex interaction of weights across layers, which is under-explored in the existing theoretical analyses of GCNs. This paper characterizes the impact of graph structures and topology sampling on the generalization performance and sample complexity explicitly, and the theoretical findings are also justified through numerical experiments.

Let Invariant Rationale Discovery Inspire Graph Contrastive Learning

- Sihang Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, Tat-Seng Chua
- abstract: Leading graph contrastive learning (GCL) methods perform graph augmentations in two fashions: (1) randomly corrupting the anchor graph, which could cause the loss of semantic information, or (2) using domain knowledge to maintain salient features, which undermines the generalization to other domains. Taking an invariance look at GCL, we argue that a high-performing augmentation should preserve the salient semantics of anchor graphs regarding instance-discrimination. To this end, we relate GCL with invariant rationale discovery, and propose a new framework, Rationale-aware Graph Contrastive Learning (RGCL). Specifically, without supervision signals, RGCL uses a rationale generator to reveal salient features about graph instance-discrimination as the rationale, and then creates rationale-aware views for contrastive learning. This rationale-aware pre-training scheme endows the backbone model with the powerful representation ability, further facilitating the fine-tuning on downstream tasks. On MNIST-Superpixel and MUTAG datasets, visual inspections on the discovered rationales showcase that the rationale generator successfully captures the salient features (\ie distinguishing semantic nodes in graphs). On biochemical molecule and social network benchmark datasets, the state-of-the-art performance of RGCL demonstrates the effectiveness of rationale-aware views for contrastive learning. Our codes are available at https://github.com/lsh0520/RGCL.

Difference Advantage Estimation for Multi-Agent Policy Gradients

- Yueheng Li, Guangming Xie, Zongqing Lu
- abstract: Multi-agent policy gradient methods in centralized training with decentralized execution recently witnessed many progresses. During centralized training, multi-agent credit assignment is crucial, which can substantially promote learning performance. However, explicit multi-agent credit assignment in multi-agent policy gradient methods still receives less attention. In this paper, we investigate multi-agent credit assignment induced by reward shaping and provide a theoretical understanding in terms of its credit assignment and policy bias. Based on this, we propose an exponentially weighted advantage estimator, which is analogous to GAE, to enable multi-agent credit assignment while allowing the tradeoff with policy bias. Empirical results show that our approach can successfully perform effective multi-agent credit assignment, and thus substantially outperforms other advantage estimators.

Private Adaptive Optimization with Side information

- Tian Li, Manzil Zaheer, Sashank Reddi, Virginia Smith
- abstract: Adaptive optimization methods have become the default solvers for many machine learning tasks. Unfortunately, the benefits of adaptivity may degrade when training with differential privacy, as the noise added to ensure privacy reduces the effectiveness of the adaptive preconditioner. To this end, we propose AdaDPS, a general framework that uses non-sensitive side information to precondition the gradients, allowing the effective use of adaptive methods in private settings. We formally show AdaDPS reduces the amount of noise needed to achieve similar privacy guarantees, thereby improving optimization performance. Empirically, we leverage simple and readily available side information to explore the performance of AdaDPS in practice, comparing to strong baselines in both centralized and federated settings. Our results show that AdaDPS improves accuracy by 7.7% (absolute) on average —yielding state-of-the-art privacy-utility trade-offs on large-scale text and image benchmarks.

Permutation Search of Tensor Network Structures via Local Sampling

- Chao Li, Junhua Zeng, Zerui Tao, Qibin Zhao
- abstract: Recent works put much effort into tensor network structure search (TN-SS), aiming to select suitable tensor network (TN) structures, involving the TN-ranks, formats, and so on, for the decomposition or learning tasks. In this paper, we consider a practical variant of TN-SS, dubbed TN permutation search (TN-PS), in which we search for good mappings from tensor modes onto TN vertices (core tensors) for compact TN representations. We conduct a theoretical investigation of TN-PS and propose a practically-efficient algorithm to resolve the problem. Theoretically, we prove the counting and metric properties of search spaces of TN-PS, analyzing for the first time the impact of TN structures on these unique properties. Numerically, we propose a novel meta-heuristic algorithm, in which the searching is done by randomly sampling in a neighborhood established in our theory, and then recurrently updating the neighborhood until convergence. Numerical results demonstrate that the new algorithm can reduce the required model size of TNs in extensive benchmarks, implying the improvement in the expressive power of TNs. Furthermore, the computational cost for the new algorithm is significantly less than that in (Li and Sun, 2020).

Hessian-Free High-Resolution Nesterov Acceleration For Sampling

- Ruilin Li, Hongyuan Zha, Molei Tao
- abstract: Nesterov's Accelerated Gradient (NAG) for optimization has better performance than its continuous time limit (noiseless kinetic Langevin) when a finite step-size is employed (Shi et al., 2021). This work explores the sampling counterpart of this phenonemon and proposes a diffusion process, whose discretizations can yield accelerated gradient-based MCMC methods. More precisely, we reformulate the optimizer of NAG for strongly convex functions (NAG-SC) as a Hessian-Free High-Resolution ODE, change its high-resolution coefficient to a hyperparameter, inject appropriate noise, and discretize the

resulting diffusion process. The acceleration effect of the new hyperparameter is quantified and it is not an artificial one created by time-rescaling. Instead, acceleration beyond underdamped Langevin in W_2 distance is quantitatively established for log-strongly-concave-and-smooth targets, at both the continuous dynamics level and the discrete algorithm level. Empirical experiments in both log-strongly-concave and multi-modal cases also numerically demonstrate this acceleration.

Double Sampling Randomized Smoothing

- Linyi Li, Jiawei Zhang, Tao Xie, Bo Li
- abstract: Neural networks (NNs) are known to be vulnerable against adversarial perturbations, and thus there is a line of work aiming to provide robustness certification for NNs, such as randomized smoothing, which samples smoothing noises from a certain distribution to certify the robustness for a smoothed classifier. However, as previous work shows, the certified robust radius in randomized smoothing suffers from scaling to large datasets ("curse of dimensionality"). To overcome this hurdle, we propose a Double Sampling Randomized Smoothing (DSRS) framework, which exploits the sampled probability from an additional smoothing distribution to tighten the robustness certification of the previous smoothed classifier. Theoretically, under mild assumptions, we prove that DSRS can certify $\Theta(\sqrt d)$ robust radius under ℓ_2 norm where d is the input dimension, which implies that DSRS may be able to break the curse of dimensionality of randomized smoothing. We instantiate DSRS for a generalized family of Gaussian smoothing and propose an efficient and sound computing method based on customized dual optimization considering sampling error. Extensive experiments on MNIST, CIFAR-10, and ImageNet verify our theory and show that DSRS certifies larger robust radii than existing baselines consistently under different settings. Code is available at https://github.com/llylly/DSRS.

HousE: Knowledge Graph Embedding with Householder Parameterization

- Rui Li, Jianan Zhao, Chaozhuo Li, Di He, Yiqi Wang, Yuming Liu, Hao Sun, Senzhang Wang, Weiwei Deng, Yanming Shen, Xing Xie, Qi Zhang
- abstract: The effectiveness of knowledge graph embedding (KGE) largely depends on the ability to model intrinsic relation patterns and mapping properties. However, existing approaches can only capture some of them with insufficient modeling capacity. In this work, we propose a more powerful KGE framework named HousE, which involves a novel parameterization based on two kinds of Householder transformations: (1) Householder rotations to achieve superior capacity of modeling relation patterns; (2) Householder projections to handle sophisticated relation mapping properties. Theoretically, HousE is capable of modeling crucial relation patterns and mapping properties simultaneously. Besides, HousE is a generalization of existing rotation-based models while extending the rotations to high-dimensional spaces. Empirically, HousE achieves new state-of-the-art performance on five benchmark datasets. Our code is available at https://github.com/anrep/HousE.

Learning Multiscale Transformer Models for Sequence Generation

- Bei Li, Tong Zheng, Yi Jing, Chengbo Jiao, Tong Xiao, Jingbo Zhu
- abstract: Multiscale feature hierarchies have been witnessed the success in the computer vision area. This further motivates researchers to design multiscale Transformer for natural language processing, mostly based on the self-attention mechanism. For example, restricting the receptive field across heads or extracting local fine-grained features via convolutions. However, most of existing works directly modeled local features but ignored the word-boundary information. This results in redundant and ambiguous attention distributions, which lacks of interpretability. In this work, we define those scales in different linguistic units, including sub-words, words and phrases. We built a multiscale Transformer model by establishing relationships among scales based on word-boundary information and phrase-level prior knowledge. The proposed \textbf{U}niversal \textbf{M}ulti\textbf{S}cale \textbf{T}ransformer, namely \textsc{Umst}, was evaluated on two sequence generation tasks. Notably, it yielded consistent performance gains over the strong baseline on several test sets without sacrificing the efficiency.

Finding Global Homophily in Graph Neural Networks When Meeting Heterophily

- Xiang Li, Renyu Zhu, Yao Cheng, Caihua Shan, Siqiang Luo, Dongsheng Li, Weining Qian
- abstract: We investigate graph neural networks on graphs with heterophily. Some existing methods amplify a node's neighborhood with multi-hop neighbors to include more nodes with homophily. However, it is a significant challenge to set personalized neighborhood sizes for different nodes. Further, for other homophilous nodes excluded in the neighborhood, they are ignored for information aggregation. To address these problems, we propose two models GloGNN and GloGNN++, which generate a node's embedding by aggregating information from global nodes in the graph. In each layer, both models learn a coefficient matrix to capture the correlations between nodes, based on which neighborhood aggregation is performed. The coefficient matrix allows signed values and is derived from an optimization problem that has a closed-form solution. We further accelerate neighborhood aggregation and derive a linear time complexity. We theoretically explain the models' effectiveness by proving that both the coefficient matrix and the generated node embedding matrix have the desired grouping effect. We conduct extensive experiments to compare our models against 11 other competitors on 15 benchmark datasets in a wide range of domains, scales and graph heterophilies. Experimental results show that our methods achieve superior performance and are also very efficient.

Fat–Tailed Variational Inference with Anisotropic Tail Adaptive Flows

- Feynman Liang, Michael Mahoney, Liam Hodgkinson
- abstract: While fat-tailed densities commonly arise as posterior and marginal distributions in robust models and scale mixtures, they present a problematic scenario when Gaussian-based variational inference fails to accurately capture tail decay. We first improve previous theory on tails of Lipschitz flows by quantifying how they affect the rate of tail decay and expanding the theory to non-Lipschitz polynomial flows. Next, we develop an alternative theory for multivariate tail parameters which is sensitive to tail-anisotropy. In doing so, we unveil a fundamental problem which plagues many existing flow-based methods: they can only model tail-isotropic distributions (i.e., distributions having the same tail parameter in every direction). To mitigate this and enable modeling of tail-anisotropic targets, we propose anisotropic tail-adaptive flows (ATAF). Experimental results confirm ATAF on both synthetic and real-world targets is competitive with prior work while also exhibiting appropriate tail-anisotropy.

Exploring and Exploiting Hubness Priors for High-Quality GAN Latent Sampling

- Yuanbang Liang, Jing Wu, Yu-Kun Lai, Yipeng Qin
- abstract: Despite the extensive studies on Generative Adversarial Networks (GANs), how to reliably sample high-quality images from their latent spaces remains an under-explored topic. In this paper, we propose a novel GAN latent sampling method by exploring and exploiting the hubness priors of GAN latent distributions. Our key insight is that the high dimensionality of the GAN latent space will inevitably lead to the emergence of hub latents that usually have much larger sampling densities than other latents in the latent space. As a result, these hub latents are better trained and thus contribute more to the synthesis of high-quality images. Unlike the a posterior "cherry-picking", our method is highly efficient as it is an a priori method that identifies high-quality latents before the synthesis of images. Furthermore, we show that the well-known but purely empirical truncation trick is a naive approximation to the central clustering effect of hub latents, which not only uncovers the rationale of the truncation trick, but also indicates the superiority and fundamentality of our method. Extensive experimental results demonstrate the effectiveness of the proposed method. Our code is available at: https://github.com/Byronliang8/HubnessGANSampling.

Reducing Variance in Temporal-Difference Value Estimation via Ensemble of Deep Networks

- Litian Liang, Yaosheng Xu, Stephen Mcaleer, Dailin Hu, Alexander Ihler, Pieter Abbeel, Roy Fox
- abstract: In temporal-difference reinforcement learning algorithms, variance in value estimation can cause instability and overestimation of the maximal target value. Many algorithms have been proposed to reduce overestimation, including several recent ensemble methods, however none have shown success in sample-efficient learning through addressing estimation variance as the root cause of overestimation. In this paper, we propose MeanQ, a simple ensemble method that estimates target values as ensemble means. Despite its simplicity, MeanQ shows remarkable sample efficiency in experiments on the Atari Learning Environment benchmark. Importantly, we find that an ensemble of size 5 sufficiently reduces estimation variance to obviate the lagging target network, eliminating it as a source of bias and further gaining sample efficiency. We justify intuitively and empirically the design choices in MeanQ, including the necessity of independent experience sampling. On a set of 26 benchmark Atari environments, MeanQ outperforms all tested baselines, including the best available baseline, SUNRISE, at 100K interaction steps in 16/26 environments, and by 68% on average. MeanQ also outperforms Rainbow DQN at 500K steps in 21/26 environments, and by 49% on average, and achieves average human-level performance using 200K ($\pm$100K) interaction steps. Our implementation is available at https://github.com/indylab/MeanQ.

TSPipe: Learn from Teacher Faster with Pipelines

- Hwijoon Lim, Yechan Kim, Sukmin Yun, Jinwoo Shin, Dongsu Han
- abstract: The teacher-student (TS) framework, training a (student) network by utilizing an auxiliary superior (teacher) network, has been adopted as a popular training paradigm in many machine learning schemes, since the seminal work—Knowledge distillation (KD) for model compression and transfer learning. Many recent self-supervised learning (SSL) schemes also adopt the TS framework, where teacher networks are maintained as the moving average of student networks, called the momentum networks. This paper presents TSPipe, a pipelined approach to accelerate the training process of any TS frameworks including KD and SSL. Under the observation that the teacher network does not need a backward pass, our main idea is to schedule the computation of the teacher and student network separately, and fully utilize the GPU during training by interleaving the computations of the two networks and relaxing their dependencies. In case the teacher network requires a momentum update, we use delayed parameter updates only on the teacher network to attain high model accuracy. Compared to existing pipeline parallelism schemes, which sacrifice either training throughput or model accuracy, TSPipe provides better performance trade-offs, achieving up to 12.15x higher throughput.

Order Constraints in Optimal Transport

- Yu Chin Fabian Lim, Laura Wynter, Shiau Hong Lim
- abstract: Optimal transport is a framework for comparing measures whereby a cost is incurred for transporting one measure to another. Recent works have aimed to improve optimal transport plans through the introduction of various forms of structure. We introduce novel order constraints into the optimal transport formulation to allow for the incorporation of structure. We define an efficient method for obtaining explainable solutions to the new formulation that scales far better than standard approaches. The theoretical properties of the method are provided. We demonstrate experimentally that order constraints improve explainability using the e-SNLI (Stanford Natural Language Inference) dataset that includes human-annotated rationales as well as on several image color transfer examples.

Flow-Guided Sparse Transformer for Video Deblurring

- Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, Luc Van Gool
- abstract: Exploiting similar and sharper scene patches in spatio-temporal neighborhoods is critical for video deblurring. However, CNN-based methods show limitations in capturing long-range dependencies and modeling non-local self-similarity. In this paper, we propose a novel framework, Flow-Guided Sparse Transformer (FGST), for video deblurring. In FGST, we customize a self-attention module, Flow-Guided Sparse Window-based Multi-head Self-Attention (FGSW-MSA). For each $query$ element on the blurry reference frame, FGSW-MSA enjoys the guidance of the estimated optical flow to globally sample spatially sparse yet highly related key elements corresponding to the same scene patch in neighboring frames. Besides, we present a Recurrent Embedding (RE) mechanism to transfer information from past frames and strengthen long-range temporal dependencies. Comprehensive experiments demonstrate that our proposed FGST outperforms state-of-the-art (SOTA) methods on both DVD and GOPRO datasets and yields visually pleasant results in real video deblurring. https://github.com/linjing7/VR-Baseline

Federated Learning with Positive and Unlabeled Data

- Xinyang Lin, Hanting Chen, Yixing Xu, Chao Xu, Xiaolin Gui, Yiping Deng, Yunhe Wang
- abstract: We study the problem of learning from positive and unlabeled (PU) data in the federated setting, where each client only labels a little part of their dataset due to the limitation of resources and time. Different from the settings in traditional PU learning where the negative class consists of a single class, the negative samples which cannot be identified by a client in the federated setting may come from multiple classes which are unknown to the client. Therefore, existing PU learning methods can be hardly applied in this situation. To address this problem, we propose a novel framework, namely Federated learning with Positive and Unlabeled data (FedPU), to minimize the expected risk of multiple negative classes by leveraging the labeled data in other clients. We theoretically analyze the generalization bound of the proposed FedPU. Empirical experiments show that the FedPU can achieve much better performance than conventional supervised and semi-supervised federated learning methods.

Decentralized Online Convex Optimization in Networked Systems

- Yiheng Lin, Judy Gan, Guannan Qu, Yash Kanoria, Adam Wierman
- abstract: We study the problem of networked online convex optimization, where each agent individually decides on an action at every time step and agents cooperatively seek to minimize the total global cost over a finite horizon. The global cost is made up of three types of local costs: convex node costs, temporal interaction costs, and spatial interaction costs. In deciding their individual action at each time, an agent has access to predictions of local cost functions for the next k time steps in an r-hop neighborhood. Our work proposes a novel online algorithm, Localized Predictive Control (LPC), which generalizes predictive control to multi-agent systems. We show that LPC achieves a competitive ratio of $1 + \tilde{O}(\rho_T^k) + \tilde{O}(\rho_S^r)$ in an adversarial setting, where ρ_T and ρ_S are constants in $(0, 1)$ that increase with the relative strength of temporal and spatial interaction costs, respectively. This is the first competitive ratio bound on decentralized predictive control for networked online convex optimization. Further, we show that the dependence on k and r in our results is near optimal by lower bounding the competitive ratio of any decentralized online algorithm.

Unsupervised Flow-Aligned Sequence-to-Sequence Learning for Video Restoration

- Jing Lin, Xiaowan Hu, Yuanhao Cai, Haoqian Wang, Youliang Yan, Xueyi Zou, Yulun Zhang, Luc Van Gool
- abstract: How to properly model the inter-frame relation within the video sequence is an important but unsolved challenge for video restoration (VR). In this work, we propose an unsupervised flow-aligned sequence-to-sequence model (S2SVR) to address this problem. On the one hand, the sequence-to-sequence model, which has proven capable of sequence modeling in the field of natural language processing, is explored for the first time in VR. Optimized serialization modeling shows potential in capturing long-range dependencies among frames. On the other hand, we equip the sequence-to-sequence model with an unsupervised optical flow estimator to maximize its potential. The flow estimator is trained with our proposed unsupervised distillation loss, which can alleviate the data discrepancy and inaccurate degraded optical flow issues of previous flow-based methods. With reliable optical flow, we can establish accurate correspondence among multiple frames, narrowing the domain difference between 1D language and 2D misaligned

frames and improving the potential of the sequence-to-sequence model. S2SVR shows superior performance in multiple VR tasks, including video deblurring, video super-resolution, and compressed video quality enhancement. https://github.com/linjing7/VR-Baseline

Constrained Gradient Descent: A Powerful and Principled Evasion Attack Against Neural Networks

- Weiran Lin, Keane Lucas, Lujo Bauer, Michael K. Reiter, Mahmood Sharif
- abstract: We propose new, more efficient targeted white-box attacks against deep neural networks. Our attacks better align with the attacker's goal: (1) tricking a model to assign higher probability to the target class than to any other class, while (2) staying within an ϵ-distance of the attacked input. First, we demonstrate a loss function that explicitly encodes (1) and show that Auto-PGD finds more attacks with it. Second, we propose a new attack method, Constrained Gradient Descent (CGD), using a refinement of our loss function that captures both (1) and (2). CGD seeks to satisfy both attacker objectives—misclassification and bounded ℓ_{p}-norm—in a principled manner, as part of the optimization, instead of via ad hoc post-processing techniques (e.g., projection or clipping). We show that CGD is more successful on CIFAR10 (0.9–4.2%) and ImageNet (8.6–13.6%) than state-of-the-art attacks while consuming less time (11.4–18.8%). Statistical tests confirm that our attack outperforms others against leading defenses on different datasets and values of ϵ.

Learning Augmented Binary Search Trees

- Honghao Lin, Tian Luo, David Woodruff
- abstract: A treap is a classic randomized binary search tree data structure that is easy to implement and supports O(log n) expected time access. However, classic treaps do not take advantage of the input distribution or patterns in the input. Given recent advances in algorithms with predictions, we propose pairing treaps with machine advice to form a learning-augmented treap. We are the first to propose a learning-augmented data structure that supports binary search tree operations such as range-query and successor functionalities. With the assumption that we have access to advice from a frequency estimation oracle, we assign learned priorities to the nodes to better improve the treap's structure. We theoretically analyze the learning-augmented treap's performance under various input distributions and show that under those circumstances, our learning-augmented treap has stronger guarantees than classic treaps and other classic tree-based data structures. Further, we experimentally evaluate our learned treap on synthetic datasets and demonstrate a performance advantage over other search tree data structures. We also present experiments on real world datasets with known frequency estimation oracles and show improvements as well.

Online Nonsubmodular Minimization with Delayed Costs: From Full Information to Bandit Feedback

- Tianyi Lin, Aldo Pacchiano, Yaodong Yu, Michael Jordan
- abstract: Motivated by applications to online learning in sparse estimation and Bayesian optimization, we consider the problem of online unconstrained nonsubmodular minimization with delayed costs in both full information and bandit feedback settings. In contrast to previous works on online unconstrained submodular minimization, we focus on a class of nonsubmodular functions with special structure, and prove regret guarantees for several variants of the online and approximate online bandit gradient descent algorithms in static and delayed scenarios. We derive bounds for the agent's regret in the full information and bandit feedback setting, even if the delay between choosing a decision and receiving the incurred cost is unbounded. Key to our approach is the notion of (α, β)-regret and the extension of the generic convex relaxation model from \citet{El-2020-Optimal}, the analysis of which is of independent interest. We conduct and showcase several simulation studies to demonstrate the efficacy of our algorithms.

Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments

- Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, Siddhartha Sen
- abstract: We develop a new, principled algorithm for estimating the contribution of training data points to the behavior of a deep learning model, such as a specific prediction it makes. Our algorithm estimates the AME, a quantity that measures the expected (average) marginal effect of adding a data point to a subset of the training data, sampled from a given distribution. When subsets are sampled from the uniform distribution, the AME reduces to the well-known Shapley value. Our approach is inspired by causal inference and randomized experiments: we sample different subsets of the training data to train multiple submodels, and evaluate each submodel's behavior. We then use a LASSO regression to jointly estimate the AME of each data point, based on the subset compositions. Under sparsity assumptions ($k \ll N$ datapoints have large AME), our estimator requires only $O(k\log N)$ randomized submodel trainings, improving upon the best prior Shapley value estimators.

Interactively Learning Preference Constraints in Linear Bandits

- David Lindner, Sebastian Tschiatschek, Katja Hofmann, Andreas Krause
- abstract: We study sequential decision-making with known rewards and unknown constraints, motivated by situations where the constraints represent expensive-to-evaluate human preferences, such as safe and comfortable driving behavior. We formalize the challenge of interactively learning about these constraints as a novel linear bandit problem which we call constrained linear best-arm identification. To solve this problem, we propose the Adaptive Constraint Learning (ACOL) algorithm. We provide an instance-dependent lower bound for constrained linear best-arm identification and show that ACOL's sample complexity matches the lower bound in the worst-case. In the average case, ACOL's sample complexity bound is still significantly tighter than bounds of simpler approaches. In synthetic experiments, ACOL performs on par with an oracle solution and outperforms a range of baselines. As an application, we consider learning constraints to represent human preferences in a driving simulation. ACOL is significantly more sample efficient than alternatives for this application. Further, we find that learning preferences as constraints is more robust to changes in the driving scenario than encoding the preferences directly in the reward function.

Delayed Reinforcement Learning by Imitation

- Pierre Liotet, Davide Maran, Lorenzo Bisi, Marcello Restelli
- abstract: When the agent's observations or interactions are delayed, classic reinforcement learning tools usually fail. In this paper, we propose a simple yet new and efficient solution to this problem. We assume that, in the undelayed environment, an efficient policy is known or can be easily learnt, but the task may suffer from delays in practice and we thus want to take them into account. We present a novel algorithm, Delayed Imitation with Dataset Aggregation (DIDA), which builds upon imitation learning methods to learn how to act in a delayed environment from undelayed demonstrations. We provide a theoretical analysis of the approach that will guide the practical design of DIDA. These results are also of general interest in the delayed reinforcement learning literature by providing bounds on the performance between delayed and undelayed tasks, under smoothness conditions. We show empirically that DIDA obtains high performances with a remarkable sample efficiency on a variety of tasks, including robotic locomotion, classic control, and trading.

CITRIS: Causal Identifiability from Temporal Intervened Sequences

- Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, Stratis Gavves
- abstract: Understanding the latent causal factors of a dynamical system from visual observations is considered a crucial step towards agents reasoning in complex environments. In this paper, we propose CITRIS, a variational autoencoder framework that learns causal representations from temporal sequences of images in which underlying causal factors have possibly been intervened upon. In contrast to the recent literature, CITRIS exploits temporality and observing intervention targets to identify scalar and multidimensional causal factors, such as 3D rotation angles. Furthermore, by

introducing a normalizing flow, CITRIS can be easily extended to leverage and disentangle representations obtained by already pretrained autoencoders. Extending previous results on scalar causal factors, we prove identifiability in a more general setting, in which only some components of a causal factor are affected by interventions. In experiments on 3D rendered image sequences, CITRIS outperforms previous methods on recovering the underlying causal variables. Moreover, using pretrained autoencoders, CITRIS can even generalize to unseen instantiations of causal factors, opening future research areas in sim-to-real generalization for causal representation learning.

StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models

- Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, Angeliki Lazaridou
- abstract: Knowledge and language understanding of models evaluated through question answering (QA) has been usually studied on static snapshots of knowledge, like Wikipedia. However, our world is dynamic, evolves over time, and our models' knowledge becomes outdated. To study how semi-parametric QA models and their underlying parametric language models (LMs) adapt to evolving knowledge, we construct a new large-scale dataset, StreamingQA, with human written and generated questions asked on a given date, to be answered from 14 years of time-stamped news articles. We evaluate our models quarterly as they read new articles not seen in pre-training. We show that parametric models can be updated without full retraining, while avoiding catastrophic forgetting. For semi-parametric models, adding new articles into the search space allows for rapid adaptation, however, models with an outdated underlying LM under-perform those with a retrained LM. For questions about higher-frequency named entities, parametric updates are particularly beneficial. In our dynamic world, the StreamingQA dataset enables a more realistic evaluation of QA models, and our experiments highlight several promising directions for future research.

Distributionally Robust Q-Learning

- Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, Zhengyuan Zhou
- abstract: Reinforcement learning (RL) has demonstrated remarkable achievements in simulated environments. However, carrying this success to real environments requires the important attribute of robustness, which the existing RL algorithms often lack as they assume that the future deployment environment is the same as the training environment (i.e. simulator) in which the policy is learned. This assumption often does not hold due to the discrepancy between the simulator and the real environment and, as a result, and hence renders the learned policy fragile when deployed. In this paper, we propose a novel distributionally robust Q-learning algorithm that learns the best policy in the worst distributional perturbation of the environment. Our algorithm first transforms the infinite-dimensional learning problem (since the environment MDP perturbation lies in an infinite-dimensional space) into a finite-dimensional dual problem and subsequently uses a multi-level Monte-Carlo scheme to approximate the dual value using samples from the simulator. Despite the complexity, we show that the resulting distributionally robust Q-learning algorithm asymptotically converges to optimal worst-case policy, thus making it robust to future environment changes. Simulation results further demonstrate its strong empirical robustness.

Constrained Variational Policy Optimization for Safe Reinforcement Learning

- Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, Ding Zhao
- abstract: Safe reinforcement learning (RL) aims to learn policies that satisfy certain constraints before deploying them to safety-critical applications. Previous primal-dual style approaches suffer from instability issues and lack optimality guarantees. This paper overcomes the issues from the perspective of probabilistic inference. We introduce a novel Expectation-Maximization approach to naturally incorporate constraints during the policy learning: 1) a provable optimal non-parametric variational distribution could be computed in closed form after a convex optimization (E-step); 2) the policy parameter is improved within the trust region based on the optimal variational distribution (M-step). The proposed algorithm decomposes the safe RL problem into a convex optimization phase and a supervised learning phase, which yields a more stable training performance. A wide range of experiments on continuous robotic tasks shows that the proposed method achieves significantly better constraint satisfaction performance and better sample efficiency than baselines. The code is available at https://github.com/liuzuxin/cvpo-safe-rl.

Benefits of Overparameterized Convolutional Residual Networks: Function Approximation under Smoothness Constraint

- Hao Liu, Minshuo Chen, Siawpeng Er, Wenjing Liao, Tong Zhang, Tuo Zhao
- abstract: Overparameterized neural networks enjoy great representation power on complex data, and more importantly yield sufficiently smooth output, which is crucial to their generalization and robustness. Most existing function approximation theories suggest that with sufficiently many parameters, neural networks can well approximate certain classes of functions in terms of the function value. The neural network themselves, however, can be highly nonsmooth. To bridge this gap, we take convolutional residual networks (ConvResNets) as an example, and prove that large ConvResNets can not only approximate a target function in terms of function value, but also exhibit sufficient first-order smoothness. Moreover, we extend our theory to approximating functions supported on a low-dimensional manifold. Our theory partially justifies the benefits of using deep and wide networks in practice. Numerical experiments on adversarial robust image classification are provided to support our theory.

Boosting Graph Structure Learning with Dummy Nodes

- Xin Liu, Jiayang Cheng, Yangqiu Song, Xin Jiang
- abstract: With the development of graph kernels and graph representation learning, many superior methods have been proposed to handle scalability and oversmoothing issues on graph structure learning. However, most of those strategies are designed based on practical experience rather than theoretical analysis. In this paper, we use a particular dummy node connecting to all existing vertices without affecting original vertex and edge properties. We further prove that such the dummy node can help build an efficient monomorphic edge-to-vertex transform and an epimorphic inverse to recover the original graph back. It also indicates that adding dummy nodes can preserve local and global structures for better graph representation learning. We extend graph kernels and graph neural networks with dummy nodes and conduct experiments on graph classification and subgraph isomorphism matching tasks. Empirical results demonstrate that taking graphs with dummy nodes as input significantly boosts graph structure learning, and using their edge-to-vertex graphs can also achieve similar results. We also discuss the gain of expressive power from the dummy in neural networks.

Equivalence Analysis between Counterfactual Regret Minimization and Online Mirror Descent

- Weiming Liu, Huacong Jiang, Bin Li, Houqiang Li
- abstract: Follow-the-Regularized-Leader (FTRL) and Online Mirror Descent (OMD) are regret minimization algorithms for Online Convex Optimization (OCO), they are mathematically elegant but less practical in solving Extensive-Form Games (EFGs). Counterfactual Regret Minimization (CFR) is a technique for approximating Nash equilibria in EFGs. CFR and its variants have a fast convergence rate in practice, but their theoretical results are not satisfactory. In recent years, researchers have been trying to link CFRs with OCO algorithms, which may provide new theoretical results and inspire new algorithms. However, existing analysis is restricted to local decision points. In this paper, we show that CFRs with Regret Matching and Regret Matching+ are equivalent to special cases of FTRL and OMD, respectively. According to these equivalences, a new FTRL and a new OMD algorithm, which can be considered as extensions of vanilla CFR and CFR+, are derived. The experimental results show that the two variants converge faster than conventional FTRL and OMD, even faster than vanilla CFR and CFR+ in some EFGs.

Deep Probability Estimation

- Sheng Liu, Aakash Kaku, Weicheng Zhu, Matan Leibovich, Sreyas Mohan, Boyang Yu, Haoxiang Huang, Laure Zanna, Narges Razavian, Jonathan Niles-Weed, Carlos Fernandez-Granda
- abstract: Reliable probability estimation is of crucial importance in many real-world applications where there is inherent (aleatoric) uncertainty. Probability-estimation models are trained on observed outcomes (e.g. whether it has rained or not, or whether a patient has died or not), because the ground-truth probabilities of the events of interest are typically unknown. The problem is therefore analogous to binary classification, with the difference that the objective is to estimate probabilities rather than predicting the specific outcome. This work investigates probability estimation from high-dimensional data using deep neural networks. There exist several methods to improve the probabilities generated by these models but they mostly focus on model (epistemic) uncertainty. For problems with inherent uncertainty, it is challenging to evaluate performance without access to ground-truth probabilities. To address this, we build a synthetic dataset to study and compare different computable metrics. We evaluate existing methods on the synthetic data as well as on three real-world probability estimation tasks, all of which involve inherent uncertainty: precipitation forecasting from radar images, predicting cancer patient survival from histopathology images, and predicting car crashes from dashcam videos. We also give a theoretical analysis of a model for high-dimensional probability estimation which reproduces several of the phenomena evinced in our experiments. Finally, we propose a new method for probability estimation using neural networks, which modifies the training process to promote output probabilities that are consistent with empirical probabilities computed from the data. The method outperforms existing approaches on most metrics on the simulated as well as real-world data.

Gating Dropout: Communication-efficient Regularization for Sparsely Activated Transformers

- Rui Liu, Young Jin Kim, Alexandre Muzio, Hany Hassan
- abstract: Sparsely activated transformers, such as Mixture of Experts (MoE), have received great interest due to their outrageous scaling capability which enables dramatical increases in model size without significant increases in computational cost. To achieve this, MoE models replace the feedforward sub-layer with Mixture-of-Experts sub-layer in transformers and use a gating network to route each token to its assigned experts. Since the common practice for efficient training of such models requires distributing experts and tokens across different machines, this routing strategy often incurs huge cross-machine communication cost because tokens and their assigned experts likely reside in different machines. In this paper, we propose Gating Dropout, which allows tokens to ignore the gating network and stay at their local machines, thus reducing the cross-machine communication. Similar to traditional dropout, we also show that Gating Dropout has a regularization effect during training, resulting in improved generalization performance. We validate the effectiveness of Gating Dropout on multilingual machine translation tasks. Our results demonstrate that Gating Dropout improves a state-of-the-art MoE model with faster wall-clock time convergence rates and better BLEU scores for a variety of model sizes and datasets.

Simplex Neural Population Learning: Any-Mixture Bayes-Optimality in Symmetric Zero-sum Games

- Siqi Liu, Marc Lanctot, Luke Marris, Nicolas Heess
- abstract: Learning to play optimally against any mixture over a diverse set of strategies is of important practical interests in competitive games. In this paper, we propose simplex-NeuPL that satisfies two desiderata simultaneously: i) learning a population of strategically diverse basis policies, represented by a single conditional network; ii) using the same network, learn best-responses to any mixture over the simplex of basis policies. We show that the resulting conditional policies incorporate prior information about their opponents effectively, enabling near optimal returns against arbitrary mixture policies in a game with tractable best-responses. We verify that such policies behave Bayes-optimally under uncertainty and offer insights in using this flexibility at test time. Finally, we offer evidence that learning best-responses to any mixture policies is an effective auxiliary task for strategic exploration, which, by itself, can lead to more performant populations.

Rethinking Attention-Model Explainability through Faithfulness Violation Test

- Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, Shiqi Wang
- abstract: Attention mechanisms are dominating the explainability of deep models. They produce probability distributions over the input, which are widely deemed as feature-importance indicators. However, in this paper, we find one critical limitation in attention explanations: weakness in identifying the polarity of feature impact. This would be somehow misleading – features with higher attention weights may not faithfully contribute to model predictions; instead, they can impose suppression effects. With this finding, we reflect on the explainability of current attention-based techniques, such as Attention $\bigodot$ Gradient and LRP-based attention explanations. We first propose an actionable diagnostic methodology (henceforth faithfulness violation test) to measure the consistency between explanation weights and the impact polarity. Through the extensive experiments, we then show that most tested explanation methods are unexpectedly hindered by the faithfulness violation issue, especially the raw attention. Empirical analyses on the factors affecting violation issues further provide useful observations for adopting explanation methods in attention models.

Optimization-Derived Learning with Essential Convergence Analysis of Training and Hyper-training

- Risheng Liu, Xuan Liu, Shangzhi Zeng, Jin Zhang, Yixuan Zhang
- abstract: Recently, Optimization-Derived Learning (ODL) has attracted attention from learning and vision areas, which designs learning models from the perspective of optimization. However, previous ODL approaches regard the training and hyper-training procedures as two separated stages, meaning that the hyper-training variables have to be fixed during the training process, and thus it is also impossible to simultaneously obtain the convergence of training and hyper-training variables. In this work, we design a Generalized Krasnoselskii-Mann (GKM) scheme based on fixed-point iterations as our fundamental ODL module, which unifies existing ODL methods as special cases. Under the GKM scheme, a Bilevel Meta Optimization (BMO) algorithmic framework is constructed to solve the optimal training and hyper-training variables together. We rigorously prove the essential joint convergence of the fixed-point iteration for training and the process of optimizing hyper-parameters for hyper-training, both on the approximation quality, and on the stationary analysis. Experiments demonstrate the efficiency of BMO with competitive performance on sparse coding and real-world applications such as image deconvolution and rain streak removal.

Deep Neural Network Fusion via Graph Matching with Applications to Model Ensemble and Federated Learning

- Chang Liu, Chenfei Lou, Runzhong Wang, Alan Yuhan Xi, Li Shen, Junchi Yan
- abstract: Model fusion without accessing training data in machine learning has attracted increasing interest due to the practical resource-saving and data privacy issues. During the training process, the neural weights of each model can be randomly permuted, and we have to align the channels of each layer before fusing them. Regrading the channels as nodes and weights as edges, aligning the channels to maximize weight similarity is a challenging NP-hard assignment problem. Due to its quadratic assignment nature, we formulate the model fusion problem as a graph matching task, considering the second-order similarity of model weights instead of previous work merely formulating model fusion as a linear assignment problem. For the rising problem scale and multi-model consistency issues, we propose an efficient graduated assignment-based model fusion method, dubbed GAMF, which iteratively updates the matchings in a consistency-maintaining manner. We apply GAMF to tackle the compact model ensemble task and federated learning task on MNIST, CIFAR-10, CIFAR-100, and Tiny-Imagenet. The performance shows the efficacy of our GAMF compared to state-of-the-art baselines.

Welfare Maximization in Competitive Equilibrium: Reinforcement Learning for Markov Exchange Economy

- Zhihan Liu, Miao Lu, Zhaoran Wang, Michael Jordan, Zhuoran Yang
- abstract: We study a bilevel economic system, which we refer to as a Markov exchange economy (MEE), from the point of view of multi-agent reinforcement learning (MARL). An MEE involves a central planner and a group of self-interested agents. The goal of the agents is to form a Competitive

Equilibrium (CE), where each agent myopically maximizes her own utility at each step. The goal of the central planner is to steer the system so as to maximize social welfare, which is defined as the sum of the utilities of all agents. Working in a setting in which the utility function and the system dynamics are both unknown, we propose to find the socially optimal policy and the CE from data via both online and offline variants of MARL. Concretely, we first devise a novel suboptimality metric specifically tailored to MEE, such that minimizing such a metric certifies globally optimal policies for both the planner and the agents. Second, in the online setting, we propose an algorithm, dubbed as \texttt{MOLM}, which combines the optimism principle for exploration with subgame CE seeking. Our algorithm can readily incorporate general function approximation tools for handling large state spaces and achieves a sublinear regret. Finally, we adapt the algorithm to an offline setting based on the pessimism principle and establish an upper bound on the suboptimality.

Generating 3D Molecules for Target Protein Binding

- Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, Shuiwang Ji
- abstract: A fundamental problem in drug discovery is to design molecules that bind to specific proteins. To tackle this problem using machine learning methods, here we propose a novel and effective framework, known as GraphBP, to generate 3D molecules that bind to given proteins by placing atoms of specific types and locations to the given binding site one by one. In particular, at each step, we first employ a 3D graph neural network to obtain geometry-aware and chemically informative representations from the intermediate contextual information. Such context includes the given binding site and atoms placed in the previous steps. Second, to preserve the desirable equivariance property, we select a local reference atom according to the designed auxiliary classifiers and then construct a local spherical coordinate system. Finally, to place a new atom, we generate its atom type and relative location w.r.t. the constructed local coordinate system via a flow model. We also consider generating the variables of interest sequentially to capture the underlying dependencies among them. Experiments demonstrate that our GraphBP is effective to generate 3D molecules with binding ability to target protein binding sites. Our implementation is available at https://github.com/divelab/GraphBP.

Communication-efficient Distributed Learning for Large Batch Optimization

- Rui Liu, Barzan Mozafari
- abstract: Many communication-efficient methods have been proposed for distributed learning, whereby gradient compression is used to reduce the communication cost. However, given recent advances in large batch optimization (e.g., large batch SGD and its variant LARS with layerwise adaptive learning rates), the compute power of each machine is being fully utilized. This means, in modern distributed learning, the per-machine computation cost is no longer negligible compared to the communication cost. In this paper, we propose new gradient compression methods for large batch optimization, JointSpar and its variant JointSpar-LARS with layerwise adaptive learning rates, that jointly reduce both the computation and the communication cost. To achieve this, we take advantage of the redundancy in the gradient computation, unlike the existing methods compute all coordinates of the gradient vector, even if some coordinates are later dropped for communication efficiency. JointSpar and its variant further reduce the training time by avoiding the wasted computation on dropped coordinates. While computationally more efficient, we prove that JointSpar and its variant also maintain the same convergence rates as their respective baseline methods. Extensive experiments show that, by reducing the time per iteration, our methods converge faster than state-of-the-art compression methods in terms of wall-clock time.

Adaptive Accelerated (Extra-)Gradient Methods with Variance Reduction

- Zijian Liu, Ta Duy Nguyen, Alina Ene, Huy Nguyen
- abstract: In this paper, we study the finite-sum convex optimization problem focusing on the general convex case. Recently, the study of variance reduced (VR) methods and their accelerated variants has made exciting progress. However, the step size used in the existing VR algorithms typically depends on the smoothness parameter, which is often unknown and requires tuning in practice. To address this problem, we propose two novel adaptive VR algorithms: Adaptive Variance Reduced Accelerated Extra-Gradient (AdaVRAE) and Adaptive Variance Reduced Accelerated Gradient (AdaVRAG). Our algorithms do not require knowledge of the smoothness parameter. AdaVRAE uses $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta}{\epsilon}}\right)$ and AdaVRAG uses $\mathcal{O}\left(n\log\log n+\sqrt{\frac{n\beta\log\beta}{\epsilon}}\right)$ gradient evaluations to attain an $\mathcal{O}(\epsilon)$-suboptimal solution, where n is the number of functions in the finite sum and β is the smoothness parameter. This result matches the best-known convergence rate of non-adaptive VR methods and it improves upon the convergence of the state of the art adaptive VR method, AdaSVRG. We demonstrate the superior performance of our algorithms compared with previous methods in experiments on real-world datasets.

REvolveR: Continuous Evolutionary Models for Robot-to-robot Policy Transfer

- Xingyu Liu, Deepak Pathak, Kris Kitani
- abstract: A popular paradigm in robotic learning is to train a policy from scratch for every new robot. This is not only inefficient but also often impractical for complex robots. In this work, we consider the problem of transferring a policy across two different robots with significantly different parameters such as kinematics and morphology. Existing approaches that train a new policy by matching the action or state transition distribution, including imitation learning methods, fail due to optimal action and/or state distribution being mismatched in different robots. In this paper, we propose a novel method named REvolveR of using continuous evolutionary models for robotic policy transfer implemented in a physics simulator. We interpolate between the source robot and the target robot by finding a continuous evolutionary change of robot parameters. An expert policy on the source robot is transferred through training on a sequence of intermediate robots that gradually evolve into the target robot. Experiments on a physics simulator show that the proposed continuous evolutionary model can effectively transfer the policy across robots and achieve superior sample efficiency on new robots. The proposed method is especially advantageous in sparse reward settings where exploration can be significantly reduced.

Kill a Bird with Two Stones: Closing the Convergence Gaps in Non-Strongly Convex Optimization by Directly Accelerated SVRG with Double Compensation and Snapshots

- Yuanyuan Liu, Fanhua Shang, Weixin An, Hongying Liu, Zhouchen Lin
- abstract: Recently, some accelerated stochastic variance reduction algorithms such as Katyusha and ASVRG-ADMM achieve faster convergence than non-accelerated methods such as SVRG and SVRG-ADMM. However, there are still some gaps between the oracle complexities and their lower bounds. To fill in these gaps, this paper proposes a novel Directly Accelerated stochastic Variance reductIon (DAVIS) algorithm with two Snapshots for non-strongly convex (non-SC) unconstrained problems. Our theoretical results show that DAVIS achieves the optimal convergence rate O(1/(nS^2)) and optimal gradient complexity O(n+\sqrt{nL/\epsilon}), which is identical to its lower bound. To the best of our knowledge, this is the first directly accelerated algorithm that attains the optimal lower bound and improves the convergence rate from O(1/S^2) to O(1/(nS^2)). Moreover, we extend DAVIS and theoretical results to non-SC problems with a structured regularizer, and prove that the proposed algorithm with double-snapshots also attains the optimal convergence rate O(1/(nS)) and optimal oracle complexity O(n+L/\epsilon) for such problems, and it is at least a factor n/S faster than existing accelerated stochastic algorithms, where n\gg S in general.

Learning Markov Games with Adversarial Opponents: Efficient Algorithms and Fundamental Limits

- Qinghua Liu, Yuanhao Wang, Chi Jin
- abstract: An ideal strategy in zero-sum games should not only grant the player an average reward no less than the value of Nash equilibrium, but also exploit the (adaptive) opponents when they are suboptimal. While most existing works in Markov games focus exclusively on the former objective, it remains open whether we can achieve both objectives simultaneously. To address this problem, this work studies no-regret learning in Markov games with

adversarial opponents when competing against the best fixed policy in hindsight. Along this direction, we present a new complete set of positive and negative results: When the policies of the opponents are revealed at the end of each episode, we propose new efficient algorithms achieving $\sqrt{K}$ regret bounds when either (1) the baseline policy class is small or (2) the opponent's policy class is small. This is complemented with an exponential lower bound when neither conditions are true. When the policies of the opponents are not revealed, we prove a statistical hardness result even in the most favorable scenario when both above conditions are true. Our hardness result is much stronger than the existing hardness results which either only involve computational hardness, or require further restrictions on the algorithms.

Local Augmentation for Graph Neural Networks

- Songtao Liu, Rex Ying, Hanze Dong, Lanqing Li, Tingyang Xu, Yu Rong, Peilin Zhao, Junzhou Huang, Dinghao Wu
- abstract: Graph Neural Networks (GNNs) have achieved remarkable performance on graph-based tasks. The key idea for GNNs is to obtain informative representation through aggregating information from local neighborhoods. However, it remains an open question whether the neighborhood information is adequately aggregated for learning representations of nodes with few neighbors. To address this, we propose a simple and efficient data augmentation strategy, local augmentation, to learn the distribution of the node representations of the neighbors conditioned on the central node's representation and enhance GNN's expressive power with generated features. Local augmentation is a general framework that can be applied to any GNN model in a plug-and-play manner. It samples feature vectors associated with each node from the learned conditional distribution as additional input for the backbone model at each training iteration. Extensive experiments and analyses show that local augmentation consistently yields performance improvement when applied to various GNN architectures across a diverse set of benchmarks. For example, experiments show that plugging in local augmentation to GCN and GAT improves by an average of 3.4% and 1.6% in terms of test accuracy on Cora, Citeseer, and Pubmed. Besides, our experimental results on large graphs (OGB) show that our model consistently improves performance over backbones. Code is available at https://github.com/SongtaoLiu0823/LAGNN.

Asking for Knowledge (AFK): Training RL Agents to Query External Knowledge Using Language

- Iou-Jen Liu, Xingdi Yuan, Marc-Alexandre Côté, Pierre-Yves Oudeyer, Alexander Schwing
- abstract: To solve difficult tasks, humans ask questions to acquire knowledge from external sources. In contrast, classical reinforcement learning agents lack such an ability and often resort to exploratory behavior. This is exacerbated as few present-day environments support querying for knowledge. In order to study how agents can be taught to query external knowledge via language, we first introduce two new environments: the grid-world-based Q-BabyAI and the text-based Q-TextWorld. In addition to physical interactions, an agent can query an external knowledge source specialized for these environments to gather information. Second, we propose the 'Asking for Knowledge' (AFK) agent, which learns to generate language commands to query for meaningful knowledge that helps solve the tasks. AFK leverages a non-parametric memory, a pointer mechanism and an episodic exploration bonus to tackle (1) irrelevant information, (2) a large query language space, (3) delayed reward for making meaningful queries. Extensive experiments demonstrate that the AFK agent outperforms recent baselines on the challenging Q-BabyAI and Q-TextWorld environments.

Learning from Demonstration: Provably Efficient Adversarial Policy Imitation with Linear Function Approximation

- Zhihan Liu, Yufeng Zhang, Zuyue Fu, Zhuoran Yang, Zhaoran Wang
- abstract: In generative adversarial imitation learning (GAIL), the agent aims to learn a policy from an expert demonstration so that its performance cannot be discriminated from the expert policy on a certain predefined reward set. In this paper, we study GAIL in both online and offline settings with linear function approximation, where both the transition and reward function are linear in the feature maps. Besides the expert demonstration, in the online setting the agent can interact with the environment, while in the offline setting the agent only accesses an additional dataset collected by a prior. For online GAIL, we propose an optimistic generative adversarial policy imitation algorithm (OGAPI) and prove that OGAPI achieves $\widetilde{\mathcal{O}}(\sqrt{H^4d^3K}+\sqrt{H^3d^2K^2/N_1})$ regret. Here N_1 represents the number of trajectories of the expert demonstration, d is the feature dimension, and K is the number of episodes. For offline GAIL, we propose a pessimistic generative adversarial policy imitation algorithm (PGAPI). We also obtain the optimality gap of PGAPI, achieving the minimax lower bound in the utilization of the additional dataset. Assuming sufficient coverage on the additional dataset, we show that PGAPI achieves $\widetilde{\mathcal{O}}(\sqrt{H^4d^2/K}+\sqrt{H^4d^3/N_2}+\sqrt{H^3d^2/N_1})$ optimality gap. Here N_2 represents the number of trajectories of the additional dataset with sufficient coverage.

GACT: Activation Compressed Training for Generic Network Architectures

- Xiaoxuan Liu, Lianmin Zheng, Dequan Wang, Yukuo Cen, Weize Chen, Xu Han, Jianfei Chen, Zhiyuan Liu, Jie Tang, Joey Gonzalez, Michael Mahoney, Alvin Cheung
- abstract: Training large neural network (NN) models requires extensive memory resources, and Activation Compression Training (ACT) is a promising approach to reduce training memory footprint. This paper presents GACT, an ACT framework to support a broad range of machine learning tasks for generic NN architectures with limited domain knowledge. By analyzing a linearized version of ACT's approximate gradient, we prove the convergence of GACT without prior knowledge on operator type or model architecture. To make training stable, we propose an algorithm that decides the compression ratio for each tensor by estimating its impact on the gradient at run time. We implement GACT as a PyTorch library that readily applies to any NN architecture. GACT reduces the activation memory for convolutional NNs, transformers, and graph NNs by up to 8.1x, enabling training with a 4.2x to 24.7x larger batch size, with negligible accuracy loss.

Robust Training under Label Noise by Over-parameterization

- Sheng Liu, Zhihui Zhu, Qing Qu, Chong You
- abstract: Recently, over-parameterized deep networks, with increasingly more network parameters than training samples, have dominated the performances of modern machine learning. However, when the training data is corrupted, it has been well-known that over-parameterized networks tend to overfit and do not generalize. In this work, we propose a principled approach for robust training of over-parameterized deep networks in classification tasks where a proportion of training labels are corrupted. The main idea is yet very simple: label noise is sparse and incoherent with the network learned from clean data, so we model the noise and learn to separate it from the data. Specifically, we model the label noise via another sparse over-parameterization term, and exploit implicit algorithmic regularizations to recover and separate the underlying corruptions. Remarkably, when trained using such a simple method in practice, we demonstrate state-of-the-art test accuracy against label noise on a variety of real datasets. Furthermore, our experimental results are corroborated by theory on simplified linear models, showing that exact separation between sparse noise and low-rank data can be achieved under incoherent conditions. The work opens many interesting directions for improving over-parameterized models by using sparse over-parameterization and implicit regularization. Code is available at https://github.com/shengliu66/SOP.

Plan Your Target and Learn Your Skills: Transferable State-Only Imitation Learning via Decoupled Policy Optimization

- Minghuan Liu, Zhengbang Zhu, Yuzheng Zhuang, Weinan Zhang, Jianye Hao, Yong Yu, Jun Wang
- abstract: Recent progress in state-only imitation learning extends the scope of applicability of imitation learning to real-world settings by relieving the need for observing expert actions. However, existing solutions only learn to extract a state-to-action mapping policy from the data, without considering how the expert plans to the target. This hinders the ability to leverage demonstrations and limits the flexibility of the policy. In this paper, we introduce Decoupled Policy Optimization (DePO), which explicitly decouples the policy as a high-level state planner and an inverse dynamics model. With embedded decoupled policy gradient and generative adversarial training, DePO enables knowledge transfer to different action spaces or state transition dynamics, and can generalize the planner to out-of-demonstration state regions. Our in-depth experimental analysis shows the effectiveness of DePO on

learning a generalized target state planner while achieving the best imitation performance. We demonstrate the appealing usage of DePO for transferring across different tasks by pre-training, and the potential for co-training agents with various skills.

On the Impossibility of Learning to Cooperate with Adaptive Partner Strategies in Repeated Games

- Robert Loftin, Frans A Oliehoek
- abstract: Learning to cooperate with other agents is challenging when those agents also possess the ability to adapt to our own behavior. Practical and theoretical approaches to learning in cooperative settings typically assume that other agents' behaviors are stationary, or else make very specific assumptions about other agents' learning processes. The goal of this work is to understand whether we can reliably learn to cooperate with other agents without such restrictive assumptions, which are unlikely to hold in real-world applications. Our main contribution is a set of impossibility results, which show that no learning algorithm can reliably learn to cooperate with all possible adaptive partners in a repeated matrix game, even if that partner is guaranteed to cooperate with some stationary strategy. Motivated by these results, we then discuss potential alternative assumptions which capture the idea that an adaptive partner will only adapt rationally to our behavior.

AutoIP: A United Framework to Integrate Physics into Gaussian Processes

- Da Long, Zheng Wang, Aditi Krishnapriyan, Robert Kirby, Shandian Zhe, Michael Mahoney
- abstract: Physical modeling is critical for many modern science and engineering applications. From a data science or machine learning perspective, where more domain-agnostic, data-driven models are pervasive, physical knowledge {—} often expressed as differential equations {—} is valuable in that it is complementary to data, and it can potentially help overcome issues such as data sparsity, noise, and inaccuracy. In this work, we propose a simple, yet powerful and general framework {—} AutoIP, for Automatically Incorporating Physics {—} that can integrate all kinds of differential equations into Gaussian Processes (GPs) to enhance prediction accuracy and uncertainty quantification. These equations can be linear or nonlinear, spatial, temporal, or spatio-temporal, complete or incomplete with unknown source terms, and so on. Based on kernel differentiation, we construct a GP prior to sample the values of the target function, equation related derivatives, and latent source functions, which are all jointly from a multivariate Gaussian distribution. The sampled values are fed to two likelihoods: one to fit the observations, and the other to conform to the equation. We use the whitening method to evade the strong dependency between the sampled function values and kernel parameters, and we develop a stochastic variational learning algorithm. AutoIP shows improvement upon vanilla GPs in both simulation and several real-world applications, even using rough, incomplete equations.

Bayesian Model Selection, the Marginal Likelihood, and Generalization

- Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, Andrew Gordon Wilson
- abstract: How do we compare between hypotheses that are entirely consistent with observations? The marginal likelihood (aka Bayesian evidence), which represents the probability of generating our observations from a prior, provides a distinctive approach to this foundational question, automatically encoding Occam's razor. Although it has been observed that the marginal likelihood can overfit and is sensitive to prior assumptions, its limitations for hyperparameter learning and discrete model comparison have not been thoroughly investigated. We first revisit the appealing properties of the marginal likelihood for learning constraints and hypothesis testing. We then highlight the conceptual and practical issues in using the marginal likelihood as a proxy for generalization. Namely, we show how marginal likelihood can be negatively correlated with generalization, with implications for neural architecture search, and can lead to both underfitting and overfitting in hyperparameter learning. We provide a partial remedy through a conditional marginal likelihood, which we show is more aligned with generalization, and practically valuable for large-scale hyperparameter learning, such as in deep kernel learning.

Feature Learning and Signal Propagation in Deep Neural Networks

- Yizhang Lou, Chris E Mingard, Soufiane Hayou
- abstract: Recent work by Baratin et al. (2021) sheds light on an intriguing pattern that occurs during the training of deep neural networks: some layers align much more with data compared to other layers (where the alignment is defined as the normalize euclidean product of the tangent features matrix and the data labels matrix). The curve of the alignment as a function of layer index (generally) exhibits a ascent-descent pattern where the maximum is reached for some hidden layer. In this work, we provide the first explanation for this phenomenon. We introduce the Equilibrium Hypothesis which connects this alignment pattern to signal propagation in deep neural networks. Our experiments demonstrate an excellent match with the theoretical predictions.

Fluctuations, Bias, Variance & Ensemble of Learners: Exact Asymptotics for Convex Losses in High-Dimension

- Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, Florent Krzakala
- abstract: From the sampling of data to the initialisation of parameters, randomness is ubiquitous in modern Machine Learning practice. Understanding the statistical fluctuations engendered by the different sources of randomness in prediction is therefore key to understanding robust generalisation. In this manuscript we develop a quantitative and rigorous theory for the study of fluctuations in an ensemble of generalised linear models trained on different, but correlated, features in high-dimensions. In particular, we provide a complete description of the asymptotic joint distribution of the empirical risk minimiser for generic convex loss and regularisation in the high-dimensional limit. Our result encompasses a rich set of classification and regression tasks, such as the lazy regime of overparametrised neural networks, or equivalently the random features approximation of kernels. While allowing to study directly the mitigating effect of ensembling (or bagging) on the bias-variance decomposition of the test error, our analysis also helps disentangle the contribution of statistical fluctuations, and the singular role played by the interpolation threshold that are at the roots of the "double-descent" phenomenon.

A Single-Loop Gradient Descent and Perturbed Ascent Algorithm for Nonconvex Functional Constrained Optimization

- Songtao Lu
- abstract: Nonconvex constrained optimization problems can be used to model a number of machine learning problems, such as multi-class Neyman-Pearson classification and constrained Markov decision processes. However, such kinds of problems are challenging because both the objective and constraints are possibly nonconvex, so it is difficult to balance the reduction of the loss value and reduction of constraint violation. Although there are a few methods that solve this class of problems, all of them are double-loop or triple-loop algorithms, and they require oracles to solve some subproblems up to certain accuracy by tuning multiple hyperparameters at each iteration. In this paper, we propose a novel gradient descent and perturbed ascent (GDPA) algorithm to solve a class of smooth nonconvex inequality constrained problems. The GDPA is a primal-dual algorithm, which only exploits the first-order information of both the objective and constraint functions to update the primal and dual variables in an alternating way. The key feature of the proposed algorithm is that it is a single-loop algorithm, where only two step-sizes need to be tuned. We show that under a mild regularity condition GDPA is able to find Karush-Kuhn-Tucker (KKT) points of nonconvex functional constrained problems with convergence rate guarantees. To the best of our knowledge, it is the first single-loop algorithm that can solve the general nonconvex smooth problems with nonconvex inequality constraints. Numerical results also showcase the superiority of GDPA compared with the best-known algorithms (in terms of both stationarity measure and feasibility of the obtained solutions).

Additive Gaussian Processes Revisited

- Xiaoyu Lu, Alexis Boukouvalas, James Hensman
- abstract: Gaussian Process (GP) models are a class of flexible non-parametric models that have rich representational power. By using a Gaussian process with additive structure, complex responses can be modelled whilst retaining interpretability. Previous work showed that additive Gaussian process models require high-dimensional interaction terms. We propose the orthogonal additive kernel (OAK), which imposes an orthogonality constraint on the additive functions, enabling an identifiable, low-dimensional representation of the functional relationship. We connect the OAK kernel to functional ANOVA decomposition, and show improved convergence rates for sparse computation methods. With only a small number of additive low-dimensional terms, we demonstrate the OAK model achieves similar or better predictive performance compared to black-box models, while retaining interpretability.

ModLaNets: Learning Generalisable Dynamics via Modularity and Physical Inductive Bias

- Yupu Lu, Shijie Lin, Guanqi Chen, Jia Pan
- abstract: Deep learning models are able to approximate one specific dynamical system but struggle at learning generalisable dynamics, where dynamical systems obey the same laws of physics but contain different numbers of elements (e.g., double- and triple-pendulum systems). To relieve this issue, we proposed the Modular Lagrangian Network (ModLaNet), a structural neural network framework with modularity and physical inductive bias. This framework models the energy of each element using modularity and then construct the target dynamical system via Lagrangian mechanics. Modularity is beneficial for reusing trained networks and reducing the scale of networks and datasets. As a result, our framework can learn from the dynamics of simpler systems and extend to more complex ones, which is not feasible using other relevant physics-informed neural networks. We examine our framework for modelling double-pendulum or three-body systems with small training datasets, where our models achieve the best data efficiency and accuracy performance compared with counterparts. We also reorganise our models as extensions to model multi-pendulum and multi-body systems, demonstrating the intriguing reusable feature of our framework.

Model-Free Opponent Shaping

- Christopher Lu, Timon Willi, Christian A Schroeder De Witt, Jakob Foerster
- abstract: In general-sum games the interaction of self-interested learning agents commonly leads to collectively worst-case outcomes, such as defect-defect in the iterated prisoner's dilemma (IPD). To overcome this, some methods, such as Learning with Opponent-Learning Awareness (LOLA), directly shape the learning process of their opponents. However, these methods are myopic since only a small number of steps can be anticipated, are asymmetric since they treat other agents as naive learners, and require the use of higher-order derivatives, which are calculated through white-box access to an opponent's differentiable learning algorithm. To address these issues, we propose Model-Free Opponent Shaping (M-FOS). M-FOS learns in a meta-game in which each meta-step is an episode of the underlying game. The meta-state consists of the policies in the underlying game and the meta-policy produces a new policy to be used in the next episode. M-FOS then uses generic model-free optimisation methods to learn meta-policies that accomplish long-horizon opponent shaping. Empirically, M-FOS near-optimally exploits naive learners and other, more sophisticated algorithms from the literature. For example, to the best of our knowledge, it is the first method to learn the well-known ZD extortion strategy in the IPD. In the same settings, M-FOS leads to socially optimal outcomes under meta-self-play. Finally, we show that M-FOS can be scaled to high-dimensional settings.

Multi-slots Online Matching with High Entropy

- Xingyu Lu, Qintong Wu, Wenliang Zhong
- abstract: Online matching with diversity and fairness pursuit, a common building block in the recommendation and advertising, can be modeled as constrained convex programming with high entropy. While most existing approaches are based on the "single slot" assumption (i.e., assigning one item per iteration), they cannot be directly applied to cases with multiple slots, e.g., stock-aware top-N recommendation and advertising at multiple places. Particularly, the gradient computation and resource allocation are both challenging under this setting due to the absence of a closed-form solution. To overcome these obstacles, we develop a novel algorithm named Online subGradient descent for Multi-slots Allocation (OG-MA). It uses an efficient pooling algorithm to compute closed-form of the gradient then performs a roulette swapping for allocation, yielding a sub-linear regret with linear cost per iteration. Extensive experiments on synthetic and industrial data sets demonstrate that OG-MA is a fast and promising method for multi-slots online matching.

Maximum Likelihood Training for Score-based Diffusion ODEs by High Order Denoising Score Matching

- Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, Jun Zhu
- abstract: Score-based generative models have excellent performance in terms of generation quality and likelihood. They model the data distribution by matching a parameterized score network with first-order data score functions. The score network can be used to define an ODE ("score-based diffusion ODE") for exact likelihood evaluation. However, the relationship between the likelihood of the ODE and the score matching objective is unclear. In this work, we prove that matching the first-order score is not sufficient to maximize the likelihood of the ODE, by showing a gap between the maximum likelihood and score matching objectives. To fill up this gap, we show that the negative likelihood of the ODE can be bounded by controlling the first, second, and third-order score matching errors; and we further present a novel high-order denoising score matching method to enable maximum likelihood training of score-based diffusion ODEs. Our algorithm guarantees that the higher-order matching error is bounded by the training error and the lower-order errors. We empirically observe that by high-order score matching, score-based diffusion ODEs achieve better likelihood on both synthetic data and CIFAR-10, while retaining the high generation quality.

Orchestra: Unsupervised Federated Learning via Globally Consistent Clustering

- Ekdeep Lubana, Chi Ian Tang, Fahim Kawsar, Robert Dick, Akhil Mathur
- abstract: Federated learning is generally used in tasks where labels are readily available (e.g., next word prediction). Relaxing this constraint requires design of unsupervised learning techniques that can support desirable properties for federated training: robustness to statistical/systems heterogeneity, scalability with number of participants, and communication efficiency. Prior work on this topic has focused on directly extending centralized self-supervised learning techniques, which are not designed to have the properties listed above. To address this situation, we propose Orchestra, a novel unsupervised federated learning technique that exploits the federation's hierarchy to orchestrate a distributed clustering task and enforce a globally consistent partitioning of clients' data into discriminable clusters. We show the algorithmic pipeline in Orchestra guarantees good generalization performance under a linear probe, allowing it to outperform alternative techniques in a broad range of conditions, including variation in heterogeneity, number of clients, participation ratio, and local epochs.

A Rigorous Study of Integrated Gradients Method and Extensions to Internal Neuron Attributions

- Daniel D Lundstrom, Tianjian Huang, Meisam Razaviyayn
- abstract: As deep learning (DL) efficacy grows, concerns for poor model explainability grow also. Attribution methods address the issue of explainability by quantifying the importance of an input feature for a model prediction. Among various methods, Integrated Gradients (IG) sets itself apart by claiming other methods failed to satisfy desirable axioms, while IG and methods like it uniquely satisfy said axioms. This paper comments on fundamental aspects of IG and its applications/extensions: 1) We identify key differences between IG function spaces and the supporting literature's function spaces which problematize previous claims of IG uniqueness. We show that with the introduction of an additional axiom, non-decreasing positivity, the uniqueness claims can be established. 2) We address the question of input sensitivity by identifying function classes where IG is/is not Lipschitz in the attributed input. 3) We show that axioms for single-baseline methods have analogous properties for methods with probability distribution baselines. 4) We introduce

a computationally efficient method of identifying internal neurons that contribute to specified regions of an IG attribution map. Finally, we present experimental results validating this method.

BAMDT: Bayesian Additive Semi-Multivariate Decision Trees for Nonparametric Regression

- Zhao Tang Luo, Huiyan Sang, Bani Mallick
- abstract: Bayesian additive regression trees (BART; Chipman et al., 2010) have gained great popularity as a flexible nonparametric function estimation and modeling tool. Nearly all existing BART models rely on decision tree weak learners with axis-parallel univariate split rules to partition the Euclidean feature space into rectangular regions. In practice, however, many regression problems involve features with multivariate structures (e.g., spatial locations) possibly lying in a manifold, where rectangular partitions may fail to respect irregular intrinsic geometry and boundary constraints of the structured feature space. In this paper, we develop a new class of Bayesian additive multivariate decision tree models that combine univariate split rules for handling possibly high dimensional features without known multivariate structures and novel multivariate split rules for features with multivariate structures in each weak learner. The proposed multivariate split rules are built upon stochastic predictive spanning tree bipartition models on reference knots, which are capable of achieving highly flexible nonlinear decision boundaries on manifold feature spaces while enabling efficient dimension reduction computations. We demonstrate the superior performance of the proposed method using simulation data and a Sacramento housing price data set.

Disentangled Federated Learning for Tackling Attributes Skew via Invariant Aggregation and Diversity Transferring

- Zhengquan Luo, Yunlong Wang, Zilei Wang, Zhenan Sun, Tieniu Tan
- abstract: Attributes skew hinders the current federated learning (FL) frameworks from consistent optimization directions among the clients, which inevitably leads to performance reduction and unstable convergence. The core problems lie in that: 1) Domain-specific attributes, which are non-causal and only locally valid, are indeliberately mixed into global aggregation. 2) The one-stage optimizations of entangled attributes cannot simultaneously satisfy two conflicting objectives, i.e., generalization and personalization. To cope with these, we proposed disentangled federated learning (DFL) to disentangle the domain-specific and cross-invariant attributes into two complementary branches, which are trained by the proposed alternating local-global optimization independently. Importantly, convergence analysis proves that the FL system can be stably converged even if incomplete client models participate in the global aggregation, which greatly expands the application scope of FL. Extensive experiments verify that DFL facilitates FL with higher performance, better interpretability, and faster convergence rate, compared with SOTA FL methods on both manually synthesized and realistic attributes skew datasets.

Channel Importance Matters in Few-Shot Image Classification

- Xu Luo, Jing Xu, Zenglin Xu
- abstract: Few-Shot Learning (FSL) requires vision models to quickly adapt to brand-new classification tasks with a shift in task distribution. Understanding the difficulties posed by this task distribution shift is central to FSL. In this paper, we show that a simple channel-wise feature transformation may be the key to unraveling this secret from a channel perspective. When facing novel few-shot tasks in the test-time datasets, this transformation can greatly improve the generalization ability of learned image representations, while being agnostic to the choice of datasets and training algorithms. Through an in-depth analysis of this transformation, we find that the difficulty of representation transfer in FSL stems from the severe channel bias problem of image representations: channels may have different importance in different tasks, while convolutional neural networks are likely to be insensitive, or respond incorrectly to such a shift. This points out a core problem of the generalization ability of modern vision systems which needs further attention in the future.

Learning Dynamics and Generalization in Deep Reinforcement Learning

- Clare Lyle, Mark Rowland, Will Dabney, Marta Kwiatkowska, Yarin Gal
- abstract: Solving a reinforcement learning (RL) problem poses two competing challenges: fitting a potentially discontinuous value function, and generalizing well to new observations. In this paper, we analyze the learning dynamics of temporal difference algorithms to gain novel insight into the tension between these two objectives. We show theoretically that temporal difference learning encourages agents to fit non-smooth components of the value function early in training, and at the same time induces the second-order effect of discouraging generalization. We corroborate these findings in deep RL agents trained on a range of environments, finding that neural networks trained using temporal difference algorithms on dense reward tasks exhibit weaker generalization between states than randomly initialized networks and networks trained with policy gradient methods. Finally, we investigate how post-training policy distillation may avoid this pitfall, and show that this approach improves generalization to novel environments in the ProcGen suite and improves robustness to input perturbations.

On Finite-Sample Identifiability of Contrastive Learning-Based Nonlinear Independent Component Analysis

- Qi Lyu, Xiao Fu
- abstract: Nonlinear independent component analysis (nICA) aims at recovering statistically independent latent components that are mixed by unknown nonlinear functions. Central to nICA is the identifiability of the latent components, which had been elusive until very recently. Specifically, Hyvärinen et al. have shown that the nonlinearly mixed latent components are identifiable (up to often inconsequential ambiguities) under a generalized contrastive learning (GCL) formulation, given that the latent components are independent conditioned on a certain auxiliary variable. The GCL-based identifiability of nICA is elegant, and establishes interesting connections between nICA and popular unsupervised/self-supervised learning paradigms in representation learning, causal learning, and factor disentanglement. However, existing identifiability analyses of nICA all build upon an unlimited sample assumption and the use of ideal universal function learners—which creates a non-negligible gap between theory and practice. Closing the gap is a nontrivial challenge, as there is a lack of established “textbook” routine for finite sample analysis of such unsupervised problems. This work puts forth a finite-sample identifiability analysis of GCL-based nICA. Our analytical framework judiciously combines the properties of the GCL loss function, statistical generalization analysis, and numerical differentiation. Our framework also takes the learning function’s approximation error into consideration, and reveals an intuitive trade-off between the complexity and expressiveness of the employed function learner. Numerical experiments are used to validate the theorems.

Pessimism meets VCG: Learning Dynamic Mechanism Design via Offline Reinforcement Learning

- Boxiang Lyu, Zhaoran Wang, Mladen Kolar, Zhuoran Yang
- abstract: Dynamic mechanism design has garnered significant attention from both computer scientists and economists in recent years. By allowing agents to interact with the seller over multiple rounds, where agents’ reward functions may change with time and are state-dependent, the framework is able to model a rich class of real-world problems. In these works, the interaction between agents and sellers is often assumed to follow a Markov Decision Process (MDP). We focus on the setting where the reward and transition functions of such an MDP are not known a priori, and we are attempting to recover the optimal mechanism using an a priori collected data set. In the setting where the function approximation is employed to handle large state spaces, with only mild assumptions on the expressiveness of the function class, we are able to design a dynamic mechanism using offline reinforcement learning algorithms. Moreover, learned mechanisms approximately have three key desiderata: efficiency, individual rationality, and truthfulness. Our algorithm is based on the pessimism principle and only requires a mild assumption on the coverage of the offline data set. To the best of our knowledge, our work provides the first offline RL algorithm for dynamic mechanism design without assuming uniform coverage.

Versatile Offline Imitation from Observations and Examples via Regularized State-Occupancy Matching

- Yecheng Ma, Andrew Shen, Dinesh Jayaraman, Osbert Bastani
- abstract: We propose State Matching Offline DIstribution Correction Estimation (SMODICE), a novel and versatile regression-based offline imitation learning algorithm derived via state-occupancy matching. We show that the SMODICE objective admits a simple optimization procedure through an application of Fenchel duality and an analytic solution in tabular MDPs. Without requiring access to expert actions, SMODICE can be effectively applied to three offline IL settings: (i) imitation from observations (IfO), (ii) IfO with dynamics or morphologically mismatched expert, and (iii) example-based reinforcement learning, which we show can be formulated as a state-occupancy matching problem. We extensively evaluate SMODICE on both gridworld environments as well as on high-dimensional offline benchmarks. Our results demonstrate that SMODICE is effective for all three problem settings and significantly outperforms prior state-of-art.

Quantification and Analysis of Layer-wise and Pixel-wise Information Discarding

- Haotian Ma, Hao Zhang, Fan Zhou, Yinqing Zhang, Quanshi Zhang
- abstract: This paper presents a method to explain how the information of each input variable is gradually discarded during the forward propagation in a deep neural network (DNN), which provides new perspectives to explain DNNs. We define two types of entropy-based metrics, i.e. (1) the discarding of pixel-wise information used in the forward propagation, and (2) the uncertainty of the input reconstruction, to measure input information contained by a specific layer from two perspectives. Unlike previous attribution metrics, the proposed metrics ensure the fairness of comparisons between different layers of different DNNs. We can use these metrics to analyze the efficiency of information processing in DNNs, which exhibits strong connections to the performance of DNNs. We analyze information discarding in a pixel-wise manner, which is different from the information bottleneck theory measuring feature information w.r.t. the sample distribution. Experiments have shown the effectiveness of our metrics in analyzing classic DNNs and explaining existing deep-learning techniques. The code is available at https://github.com/haotianSustc/deepinfo.

Interpretable Neural Networks with Frank-Wolfe: Sparse Relevance Maps and Relevance Orderings

- Jan Macdonald, Mathieu E. Besançon, Sebastian Pokutta
- abstract: We study the effects of constrained optimization formulations and Frank-Wolfe algorithms for obtaining interpretable neural network predictions. Reformulating the Rate-Distortion Explanations (RDE) method for relevance attribution as a constrained optimization problem provides precise control over the sparsity of relevance maps. This enables a novel multi-rate as well as a relevance-ordering variant of RDE that both empirically outperform standard RDE and other baseline methods in a well-established comparison test. We showcase several deterministic and stochastic variants of the Frank-Wolfe algorithm and their effectiveness for RDE.

A Tighter Analysis of Spectral Clustering, and Beyond

- Peter Macgregor, He Sun
- abstract: This work studies the classical spectral clustering algorithm which embeds the vertices of some graph G=(V_G, E_G) into R^k using k eigenvectors of some matrix of G, and applies k-means to partition V_G into k clusters. Our first result is a tighter analysis on the performance of spectral clustering, and explains why it works under some much weaker condition than the ones studied in the literature. For the second result, we show that, by applying fewer than k eigenvectors to construct the embedding, spectral clustering is able to produce better output for many practical instances; this result is the first of its kind in spectral clustering. Besides its conceptual and theoretical significance, the practical impact of our work is demonstrated by the empirical analysis on both synthetic and real-world data sets, in which spectral clustering produces comparable or better results with fewer than k eigenvectors.

Zero-Shot Reward Specification via Grounded Natural Language

- Parsa Mahmoudieh, Deepak Pathak, Trevor Darrell
- abstract: Reward signals in reinforcement learning are expensive to design and often require access to the true state which is not available in the real world. Common alternatives are usually demonstrations or goal images which can be labor-intensive to collect. On the other hand, text descriptions provide a general, natural, and low-effort way of communicating the desired task. However, prior works in learning text-conditioned policies still rely on rewards that are defined using either true state or labeled expert demonstrations. We use recent developments in building large-scale visuolanguage models like CLIP to devise a framework that generates the task reward signal just from goal text description and raw pixel observations which is then used to learn the task policy. We evaluate the proposed framework on control and robotic manipulation tasks. Finally, we distill the individual task policies into a single goal text conditioned policy that can generalize in a zero-shot manner to new tasks with unseen objects and unseen goal text descriptions.

Feature selection using e-values

- Subhabrata Majumdar, Snigdhansu Chatterjee
- abstract: In the context of supervised learning, we introduce the concept of e-value. An e-value is a scalar quantity that represents the proximity of the sampling distribution of parameter estimates in a model trained on a subset of features to that of the model trained on all features (i.e. the full model). Under general conditions, a rank ordering of e-values separates models that contain all essential features from those that do not. For a p-dimensional feature space, this requires fitting only the full model and evaluating p+1 models, as opposed to the traditional requirement of fitting and evaluating 2^p models. The above e-values framework is applicable to a wide range of parametric models. We use data depths and a fast resampling-based algorithm to implement a feature selection procedure, providing consistency results. Through experiments across several model settings and synthetic and real datasets, we establish that the e-values can be a promising general alternative to existing model-specific methods of feature selection.

Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations

- Bodhisattwa Prasad Majumder, Oana Camburu, Thomas Lukasiewicz, Julian Mcauley
- abstract: Models that generate extractive rationales (i.e., subsets of features) or natural language explanations (NLEs) for their predictions are important for explainable AI. While an extractive rationale provides a quick view of the features most responsible for a prediction, an NLE allows for a comprehensive description of the decision-making process behind a prediction. However, current models that generate the best extractive rationales or NLEs often fall behind the state-of-the-art (SOTA) in terms of task performance. In this work, we bridge this gap by introducing RExC, a self-rationalizing framework that grounds its predictions and two complementary types of explanations (NLEs and extractive rationales) in background knowledge. Our framework improves over previous methods by: (i) reaching SOTA task performance while also providing explanations, (ii) providing two types of explanations, while existing models usually provide only one type, and (iii) beating by a large margin the previous SOTA in terms of quality of both types of explanations. Furthermore, a perturbation analysis in RExC shows a high degree of association between explanations and predictions, a necessary property of faithful explanations.

Nonparametric Involutive Markov Chain Monte Carlo

- Carol Mak, Fabian Zaiser, Luke Ong

- abstract: A challenging problem in probabilistic programming is to develop inference algorithms that work for arbitrary programs in a universal probabilistic programming language (PPL). We present the nonparametric involutive Markov chain Monte Carlo (NP-iMCMC) algorithm as a method for constructing MCMC inference algorithms for nonparametric models expressible in universal PPLs. Building on the unifying involutive MCMC framework, and by providing a general procedure for driving state movement between dimensions, we show that NP-iMCMC can generalise numerous existing iMCMC algorithms to work on nonparametric models. We prove the correctness of the NP-iMCMC sampler. Our empirical study shows that the existing strengths of several iMCMC algorithms carry over to their nonparametric extensions. Applying our method to the recently proposed Nonparametric HMC, an instance of (Multiple Step) NP-iMCMC, we have constructed several nonparametric extensions (all of which new) that exhibit significant performance improvements.

Architecture Agnostic Federated Learning for Neural Networks

- Disha Makhija, Xing Han, Nhat Ho, Joydeep Ghosh
- abstract: With growing concerns regarding data privacy and rapid increase in data volume, Federated Learning (FL) has become an important learning paradigm. However, jointly learning a deep neural network model in a FL setting proves to be a non-trivial task because of the complexities associated with the neural networks, such as varied architectures across clients, permutation invariance of the neurons, and presence of non-linear transformations in each layer. This work introduces a novel framework, Federated Heterogeneous Neural Networks (FedHeNN), that allows each client to build a personalised model without enforcing a common architecture across clients. This allows each client to optimize with respect to local data and compute constraints, while still benefiting from the learnings of other (potentially more powerful) clients. The key idea of FedHeNN is to use the instance-level representations obtained from peer clients to guide the simultaneous training on each client. The extensive experimental results demonstrate that the FedHeNN framework is capable of learning better performing models on clients in both the settings of homogeneous and heterogeneous architectures across clients.

Robustness in Multi-Objective Submodular Optimization: a Quantile Approach

- Cedric Malherbe, Kevin Scaman
- abstract: The optimization of multi-objective submodular systems appears in a wide variety of applications. However, there are currently very few techniques which are able to provide a robust allocation to such systems. In this work, we propose to design and analyse novel algorithms for the robust allocation of submodular systems through lens of quantile maximization. We start by observing that identifying an exact solution for this problem is computationally intractable. To tackle this issue, we propose a proxy for the quantile function using a softmax formulation, and show that this proxy is well suited to submodular optimization. Based on this relaxation, we propose a novel and simple algorithm called SOFTSAT. Theoretical properties are provided for this algorithm as well as novel approximation guarantees. Finally, we provide numerical experiments showing the efficiency of our algorithm with regards to state-of-the-art methods in a test bed of real-world applications, and show that SOFTSAT is particularly robust and well-suited to online scenarios.

More Efficient Sampling for Tensor Decomposition With Worst-Case Guarantees

- Osman Asif Malik
- abstract: Recent papers have developed alternating least squares (ALS) methods for CP and tensor ring decomposition with a per-iteration cost which is sublinear in the number of input tensor entries for low-rank decomposition. However, the per-iteration cost of these methods still has an exponential dependence on the number of tensor modes when parameters are chosen to achieve certain worst-case guarantees. In this paper, we propose sampling-based ALS methods for the CP and tensor ring decompositions whose cost does not have this exponential dependence, thereby significantly improving on the previous state-of-the-art. We provide a detailed theoretical analysis and also apply the methods in a feature extraction experiment.

Unaligned Supervision for Automatic Music Transcription in The Wild

- Ben Maman, Amit H Bermano
- abstract: Multi-instrument Automatic Music Transcription (AMT), or the decoding of a musical recording into semantic musical content, is one of the holy grails of Music Information Retrieval. Current AMT approaches are restricted to piano and (some) guitar recordings, due to difficult data collection. In order to overcome data collection barriers, previous AMT approaches attempt to employ musical scores in the form of a digitized version of the same song or piece. The scores are typically aligned using audio features and strenuous human intervention to generate training labels. We introduce Note$_{EM}$, a method for simultaneously training a transcriber and aligning the scores to their corresponding performances, in a fully-automated process. Using this unaligned supervision scheme, complemented by pseudo-labels and pitch shift augmentation, our method enables training on in-the-wild recordings with unprecedented accuracy and instrumental variety. Using only synthetic data and unaligned supervision, we report SOTA note-level accuracy of the MAPS dataset, and large favorable margins on cross-dataset evaluations. We also demonstrate robustness and ease of use; we report comparable results when training on a small, easily obtainable, self-collected dataset, and we propose alternative labeling to the MusicNet dataset, which we show to be more accurate. Our project page is available at https://benadar293.github.io.

Decision-Focused Learning: Through the Lens of Learning to Rank

- Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, Tias Guns
- abstract: In the last years decision-focused learning framework, also known as predict-and-optimize, have received increasing attention. In this setting, the predictions of a machine learning model are used as estimated cost coefficients in the objective function of a discrete combinatorial optimization problem for decision making. Decision-focused learning proposes to train the ML models, often neural network models, by directly optimizing the quality of decisions made by the optimization solvers. Based on a recent work that proposed a noise contrastive estimation loss over a subset of the solution space, we observe that decision-focused learning can more generally be seen as a learning-to-rank problem, where the goal is to learn an objective function that ranks the feasible points correctly. This observation is independent of the optimization method used and of the form of the objective function. We develop pointwise, pairwise and listwise ranking loss functions, which can be differentiated in closed form given a subset of solutions. We empirically investigate the quality of our generic methods compared to existing decision-focused learning approaches with competitive results. Furthermore, controlling the subset of solutions allows controlling the runtime considerably, with limited effect on regret.

Differentially Private Coordinate Descent for Composite Empirical Risk Minimization

- Paul Mangold, Aurélien Bellet, Joseph Salmon, Marc Tommasi
- abstract: Machine learning models can leak information about the data used to train them. To mitigate this issue, Differentially Private (DP) variants of optimization algorithms like Stochastic Gradient Descent (DP-SGD) have been designed to trade-off utility for privacy in Empirical Risk Minimization (ERM) problems. In this paper, we propose Differentially Private proximal Coordinate Descent (DP-CD), a new method to solve composite DP-ERM problems. We derive utility guarantees through a novel theoretical analysis of inexact coordinate descent. Our results show that, thanks to larger step sizes, DP-CD can exploit imbalance in gradient coordinates to outperform DP-SGD. We also prove new lower bounds for composite DP-ERM under coordinate-wise regularity assumptions, that are nearly matched by DP-CD. For practical implementations, we propose to clip gradients using coordinate-wise thresholds that emerge from our theory, avoiding costly hyperparameter tuning. Experiments on real and synthetic data support our results, and show that DP-CD compares favorably with DP-SGD.

Refined Convergence Rates for Maximum Likelihood Estimation under Finite Mixture Models

- Tudor Manole, Nhat Ho
- abstract: We revisit the classical problem of deriving convergence rates for the maximum likelihood estimator (MLE) in finite mixture models. The Wasserstein distance has become a standard loss function for the analysis of parameter estimation in these models, due in part to its ability to circumvent label switching and to accurately characterize the behaviour of fitted mixture components with vanishing weights. However, the Wasserstein distance is only able to capture the worst-case convergence rate among the remaining fitted mixture components. We demonstrate that when the log-likelihood function is penalized to discourage vanishing mixing weights, stronger loss functions can be derived to resolve this shortcoming of the Wasserstein distance. These new loss functions accurately capture the heterogeneity in convergence rates of fitted mixture components, and we use them to sharpen existing pointwise and uniform convergence rates in various classes of mixture models. In particular, these results imply that a subset of the components of the penalized MLE typically converge significantly faster than could have been anticipated from past work. We further show that some of these conclusions extend to the traditional MLE. Our theoretical findings are supported by a simulation study to illustrate these improved convergence rates.

On Improving Model-Free Algorithms for Decentralized Multi-Agent Reinforcement Learning

- Weichao Mao, Lin Yang, Kaiqing Zhang, Tamer Basar
- abstract: Multi-agent reinforcement learning (MARL) algorithms often suffer from an exponential sample complexity dependence on the number of agents, a phenomenon known as the curse of multiagents. We address this challenge by investigating sample-efficient model-free algorithms in decentralized MARL, and aim to improve existing algorithms along this line. For learning (coarse) correlated equilibria in general-sum Markov games, we propose stage-based V-learning algorithms that significantly simplify the algorithmic design and analysis of recent works, and circumvent a rather complicated no-weighted-regret bandit subroutine. For learning Nash equilibria in Markov potential games, we propose an independent policy gradient algorithm with a decentralized momentum-based variance reduction technique. All our algorithms are decentralized in that each agent can make decisions based on only its local information. Neither communication nor centralized coordination is required during learning, leading to a natural generalization to a large number of agents. Finally, we provide numerical simulations to corroborate our theoretical findings.

On the Effects of Artificial Data Modification

- Antonia Marcu, Adam Prugel-Bennett
- abstract: Data distortion is commonly applied in vision models during both training (e.g methods like MixUp and CutMix) and evaluation (e.g. shape-texture bias and robustness). This data modification can introduce artificial information. It is often assumed that the resulting artefacts are detrimental to training, whilst being negligible when analysing models. We investigate these assumptions and conclude that in some cases they are unfounded and lead to incorrect results. Specifically, we show current shape bias identification methods and occlusion robustness measures are biased and propose a fairer alternative for the latter. Subsequently, through a series of experiments we seek to correct and strengthen the community's perception of how augmenting affects learning of vision models. Based on our empirical results we argue that the impact of the artefacts must be understood and exploited rather than eliminated.

Personalized Federated Learning through Local Memorization

- Othmane Marfoq, Giovanni Neglia, Richard Vidal, Laetitia Kameni
- abstract: Federated learning allows clients to collaboratively learn statistical models while keeping their data local. Federated learning was originally used to train a unique global model to be served to all clients, but this approach might be sub-optimal when clients' local data distributions are heterogeneous. In order to tackle this limitation, recent personalized federated learning methods train a separate model for each client while still leveraging the knowledge available at other clients. In this work, we exploit the ability of deep neural networks to extract high quality vectorial representations (embeddings) from non-tabular data, e.g., images and text, to propose a personalization mechanism based on local memorization. Personalization is obtained by interpolating a collectively trained global model with a local k-nearest neighbors (kNN) model based on the shared representation provided by the global model. We provide generalization bounds for the proposed approach in the case of binary classification, and we show on a suite of federated datasets that this approach achieves significantly higher accuracy and fairness than state-of-the-art methods.

Nested Bandits

- Matthieu Martin, Panayotis Mertikopoulos, Thibaud Rahier, Houssam Zenati
- abstract: In many online decision processes, the optimizing agent is called to choose between large numbers of alternatives with many inherent similarities; in turn, these similarities imply closely correlated losses that may confound standard discrete choice models and bandit algorithms. We study this question in the context of nested bandits, a class of adversarial multi-armed bandit problems where the learner seeks to minimize their regret in the presence of a large number of distinct alternatives with a hierarchy of embedded (non-combinatorial) similarities. In this setting, optimal algorithms based on the exponential weights blueprint (like Hedge, EXP3, and their variants) may incur significant regret because they tend to spend excessive amounts of time exploring irrelevant alternatives with similar, suboptimal costs. To account for this, we propose a nested exponential weights (NEW) algorithm that performs a layered exploration of the learner's set of alternatives based on a nested, step-by-step selection method. In so doing, we obtain a series of tight bounds for the learner's regret showing that online learning problems with a high degree of similarity between alternatives can be resolved efficiently, without a red bus / blue bus paradox occurring.

Closed-Form Diffeomorphic Transformations for Time Series Alignment

- Iñigo Martinez, Elisabeth Viles, Igor G. Olaizola
- abstract: Time series alignment methods call for highly expressive, differentiable and invertible warping functions which preserve temporal topology, i.e diffeomorphisms. Diffeomorphic warping functions can be generated from the integration of velocity fields governed by an ordinary differential equation (ODE). Gradient-based optimization frameworks containing diffeomorphic transformations require to calculate derivatives to the differential equation's solution with respect to the model parameters, i.e. sensitivity analysis. Unfortunately, deep learning frameworks typically lack automatic-differentiation-compatible sensitivity analysis methods; and implicit functions, such as the solution of ODE, require particular care. Current solutions appeal to adjoint sensitivity methods, ad-hoc numerical solvers or ResNet's Eulerian discretization. In this work, we present a closed-form expression for the ODE solution and its gradient under continuous piecewise-affine (CPA) velocity functions. We present a highly optimized implementation of the results on CPU and GPU. Furthermore, we conduct extensive experiments on several datasets to validate the generalization ability of our model to unseen data for time-series joint alignment. Results show significant improvements both in terms of efficiency and accuracy.

SPECTRE: Spectral Conditioning Helps to Overcome the Expressivity Limits of One-shot Graph Generators

- Karolis Martinkus, Andreas Loukas, Nathanaël Perraudin, Roger Wattenhofer
- abstract: We approach the graph generation problem from a spectral perspective by first generating the dominant parts of the graph Laplacian spectrum and then building a graph matching these eigenvalues and eigenvectors. Spectral conditioning allows for direct modeling of the global and local graph structure and helps to overcome the expressivity and mode collapse issues of one-shot graph generators. Our novel GAN, called SPECTRE, enables the one-shot generation of much larger graphs than previously possible with one-shot models. SPECTRE outperforms state-of-the-art deep autoregressive

generators in terms of modeling fidelity, while also avoiding expensive sequential generation and dependence on node ordering. A case in point, in sizable synthetic and real-world graphs SPECTRE achieves a 4-to-170 fold improvement over the best competitor that does not overfit and is 23-to-30 times faster than autoregressive generators.

Modular Conformal Calibration

- Charles Marx, Shengjia Zhao, Willie Neiswanger, Stefano Ermon
- abstract: Uncertainty estimates must be calibrated (i.e., accurate) and sharp (i.e., informative) in order to be useful. This has motivated a variety of methods for recalibration, which use held-out data to turn an uncalibrated model into a calibrated model. However, the applicability of existing methods is limited due to their assumption that the original model is also a probabilistic model. We introduce a versatile class of algorithms for recalibration in regression that we call modular conformal calibration (MCC). This framework allows one to transform any regression model into a calibrated probabilistic model. The modular design of MCC allows us to make simple adjustments to existing algorithms that enable well-behaved distribution predictions. We also provide finite-sample calibration guarantees for MCC algorithms. Our framework recovers isotonic recalibration, conformal calibration, and conformal interval prediction, implying that our theoretical results apply to those methods as well. Finally, we conduct an empirical study of MCC on 17 regression datasets. Our results show that new algorithms designed in our framework achieve near-perfect calibration and improve sharpness relative to existing methods.

Continual Repeated Annealed Flow Transport Monte Carlo

- Alex Matthews, Michael Arbel, Danilo Jimenez Rezende, Arnaud Doucet
- abstract: We propose Continual Repeated Annealed Flow Transport Monte Carlo (CRAFT), a method that combines a sequential Monte Carlo (SMC) sampler (itself a generalization of Annealed Importance Sampling) with variational inference using normalizing flows. The normalizing flows are directly trained to transport between annealing temperatures using a KL divergence for each transition. This optimization objective is itself estimated using the normalizing flow/SMC approximation. We show conceptually and using multiple empirical examples that CRAFT improves on Annealed Flow Transport Monte Carlo (Arbel et al., 2021), on which it builds and also on Markov chain Monte Carlo (MCMC) based Stochastic Normalizing Flows (Wu et al., 2020). By incorporating CRAFT within particle MCMC, we show that such learnt samplers can achieve impressively accurate results on a challenging lattice field theory example.

How to Stay Curious while avoiding Noisy TVs using Aleatoric Uncertainty Estimation

- Augustine Mavor-Parker, Kimberly Young, Caswell Barry, Lewis Griffin
- abstract: When extrinsic rewards are sparse, artificial agents struggle to explore an environment. Curiosity, implemented as an intrinsic reward for prediction errors, can improve exploration but it is known to fail when faced with action-dependent noise sources ('noisy TVs'). In an attempt to make exploring agents robust to Noisy TVs, we present a simple solution: aleatoric mapping agents (AMAs). AMAs are a novel form of curiosity that explicitly ascertain which state transitions of the environment are unpredictable, even if those dynamics are induced by the actions of the agent. This is achieved by generating separate forward predictions for the mean and aleatoric uncertainty of future states, with the aim of reducing intrinsic rewards for those transitions that are unpredictable. We demonstrate that in a range of environments AMAs are able to circumvent action-dependent stochastic traps that immobilise conventional curiosity driven agents. Furthermore, we demonstrate empirically that other common exploration approaches—previously thought to be immune to agent-induced randomness—can be trapped by stochastic dynamics.

How to Steer Your Adversary: Targeted and Efficient Model Stealing Defenses with Gradient Redirection

- Mantas Mazeika, Bo Li, David Forsyth
- abstract: Model stealing attacks present a dilemma for public machine learning APIs. To protect financial investments, companies may be forced to withhold important information about their models that could facilitate theft, including uncertainty estimates and prediction explanations. This compromise is harmful not only to users but also to external transparency. Model stealing defenses seek to resolve this dilemma by making models harder to steal while preserving utility for benign users. However, existing defenses have poor performance in practice, either requiring enormous computational overheads or severe utility trade-offs. To meet these challenges, we present a new approach to model stealing defenses called gradient redirection. At the core of our approach is a provably optimal, efficient algorithm for steering an adversary's training updates in a targeted manner. Combined with improvements to surrogate networks and a novel coordinated defense strategy, our gradient redirection defense, called GRAD^2, achieves small utility trade-offs and low computational overhead, outperforming the best prior defenses. Moreover, we demonstrate how gradient redirection enables reprogramming the adversary with arbitrary behavior, which we hope will foster work on new avenues of defense.

Quant-BnB: A Scalable Branch-and-Bound Method for Optimal Decision Trees with Continuous Features

- Rahul Mazumder, Xiang Meng, Haoyue Wang
- abstract: Decision trees are one of the most useful and popular methods in the machine learning toolbox. In this paper, we consider the problem of learning optimal decision trees, a combinatorial optimization problem that is challenging to solve at scale. A common approach in the literature is to use greedy heuristics, which may not be optimal. Recently there has been significant interest in learning optimal decision trees using various approaches (e.g., based on integer programming, dynamic programming)—to achieve computational scalability, most of these approaches focus on classification tasks with binary features. In this paper, we present a new discrete optimization method based on branch-and-bound (BnB) to obtain optimal decision trees. Different from existing customized approaches, we consider both regression and classification tasks with continuous features. The basic idea underlying our approach is to split the search space based on the quantiles of the feature distribution—leading to upper and lower bounds for the underlying optimization problem along the BnB iterations. Our proposed algorithm Quant-BnB shows significant speedups compared to existing approaches for shallow optimal trees on various real datasets.

Optimizing Tensor Network Contraction Using Reinforcement Learning

- Eli Meirom, Haggai Maron, Shie Mannor, Gal Chechik
- abstract: Quantum Computing (QC) stands to revolutionize computing, but is currently still limited. To develop and test quantum algorithms today, quantum circuits are often simulated on classical computers. Simulating a complex quantum circuit requires computing the contraction of a large network of tensors. The order (path) of contraction can have a drastic effect on the computing cost, but finding an efficient order is a challenging combinatorial optimization problem. We propose a Reinforcement Learning (RL) approach combined with Graph Neural Networks (GNN) to address the contraction ordering problem. The problem is extremely challenging due to the huge search space, the heavy-tailed reward distribution, and the challenging credit assignment. We show how a carefully implemented RL-agent that uses a GNN as the basic policy construct can address these challenges and obtain significant improvements over state-of-the-art techniques in three varieties of circuits, including the largest scale networks used in contemporary QC.

Causal Transformer for Estimating Counterfactual Outcomes

- Valentyn Melnychuk, Dennis Frauen, Stefan Feuerriegel

- abstract: Estimating counterfactual outcomes over time from observational data is relevant for many applications (e.g., personalized medicine). Yet, state-of-the-art methods build upon simple long short-term memory (LSTM) networks, thus rendering inferences for complex, long-range dependencies challenging. In this paper, we develop a novel Causal Transformer for estimating counterfactual outcomes over time. Our model is specifically designed to capture complex, long-range dependencies among time-varying confounders. For this, we combine three transformer subnetworks with separate inputs for time-varying covariates, previous treatments, and previous outcomes into a joint network with in-between cross-attentions. We further develop a custom, end-to-end training procedure for our Causal Transformer. Specifically, we propose a novel counterfactual domain confusion loss to address confounding bias: it aims to learn adversarial balanced representations, so that they are predictive of the next outcome but non-predictive of the current treatment assignment. We evaluate our Causal Transformer based on synthetic and real-world datasets, where it achieves superior performance over current baselines. To the best of our knowledge, this is the first work proposing transformer-based architecture for estimating counterfactual outcomes from longitudinal data.

Steerable 3D Spherical Neurons

- Pavlo Melnyk, Michael Felsberg, Mårten Wadenbäck
- abstract: Emerging from low-level vision theory, steerable filters found their counterpart in prior work on steerable convolutional neural networks equivariant to rigid transformations. In our work, we propose a steerable feed-forward learning-based approach that consists of neurons with spherical decision surfaces and operates on point clouds. Such spherical neurons are obtained by conformal embedding of Euclidean space and have recently been revisited in the context of learning representations of point sets. Focusing on 3D geometry, we exploit the isometry property of spherical neurons and derive a 3D steerability constraint. After training spherical neurons to classify point clouds in a canonical orientation, we use a tetrahedron basis to quadruplicate the neurons and construct rotation-equivariant spherical filter banks. We then apply the derived constraint to interpolate the filter bank outputs and, thus, obtain a rotation-invariant network. Finally, we use a synthetic point set and real-world 3D skeleton data to verify our theoretical findings. The code is available at https://github.com/pavlo-melnyk/steerable-3d-neurons.

Transformers are Meta-Reinforcement Learners

- Luckeciano C Melo
- abstract: The transformer architecture and variants presented a remarkable success across many machine learning tasks in recent years. This success is intrinsically related to the capability of handling long sequences and the presence of context-dependent weights from the attention mechanism. We argue that these capabilities suit the central role of a Meta-Reinforcement Learning algorithm. Indeed, a meta-RL agent needs to infer the task from a sequence of trajectories. Furthermore, it requires a fast adaptation strategy to adapt its policy for a new task - which can be achieved using the self-attention mechanism. In this work, we present TrMRL (Transformers for Meta-Reinforcement Learning), a meta-RL agent that mimics the memory reinstatement mechanism using the transformer architecture. It associates the recent past of working memories to build an episodic memory recursively through the transformer layers. We show that the self-attention computes a consensus representation that minimizes the Bayes Risk at each layer and provides meaningful features to compute the best actions. We conducted experiments in high-dimensional continuous control environments for locomotion and dexterous manipulation. Results show that TrMRL presents comparable or superior asymptotic performance, sample efficiency, and out-of-distribution generalization compared to the baselines in these environments.

ButterflyFlow: Building Invertible Layers with Butterfly Matrices

- Chenlin Meng, Linqi Zhou, Kristy Choi, Tri Dao, Stefano Ermon
- abstract: Normalizing flows model complex probability distributions using maps obtained by composing invertible layers. Special linear layers such as masked and 1{\texttimes}1 convolutions play a key role in existing architectures because they increase expressive power while having tractable Jacobians and inverses. We propose a new family of invertible linear layers based on butterfly layers, which are known to theoretically capture complex linear structures including permutations and periodicity, yet can be inverted efficiently. This representational power is a key advantage of our approach, as such structures are common in many real-world datasets. Based on our invertible butterfly layers, we construct a new class of normalizing flow mod- els called ButterflyFlow. Empirically, we demonstrate that ButterflyFlows not only achieve strong density estimation results on natural images such as MNIST, CIFAR-10, and ImageNet-32{\texttimes}32, but also obtain significantly better log-likelihoods on structured datasets such as galaxy images and MIMIC-III patient cohorts{—}all while being more efficient in terms of memory and computation than relevant baselines.

In defense of dual-encoders for neural ranking

- Aditya Menon, Sadeep Jayasumana, Ankit Singh Rawat, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar
- abstract: Transformer-based models such as BERT have proven successful in information retrieval problem, which seek to identify relevant documents for a given query. There are two broad flavours of such models: cross-attention (CA) models, which learn a joint embedding for the query and document, and dual-encoder (DE) models, which learn separate embeddings for the query and document. Empirically, CA models are often found to be more accurate, which has motivated a series of works seeking to bridge this gap. However, a more fundamental question remains less explored: does this performance gap reflect an inherent limitation in the capacity of DE models, or a limitation in the training of such models? And does such an understanding suggest a principled means of improving DE models? In this paper, we study these questions, with three contributions. First, we establish theoretically that with a sufficiently large embedding dimension, DE models have the capacity to model a broad class of score distributions. Second, we show empirically that on real-world problems, DE models may overfit to spurious correlations in the training set, and thus under-perform on test samples. To mitigate this behaviour, we propose a suitable distillation strategy, and confirm its practical efficacy on the MSMARCO-Passage and Natural Questions benchmarks.

Equivariant Quantum Graph Circuits

- Peter Mernyei, Konstantinos Meichanetzidis, Ismail Ilkan Ceylan
- abstract: We investigate quantum circuits for graph representation learning, and propose equivariant quantum graph circuits (EQGCs), as a class of parameterized quantum circuits with strong relational inductive bias for learning over graph-structured data. Conceptually, EQGCs serve as a unifying framework for quantum graph representation learning, allowing us to define several interesting subclasses which subsume existing proposals. In terms of the representation power, we prove that the studied subclasses of EQGCs are universal approximators for functions over the bounded graph domain. This theoretical perspective on quantum graph machine learning methods opens many directions for further work, and could lead to models with capabilities beyond those of classical approaches. We empirically verify the expressive power of EQGCs through a dedicated experiment on synthetic data, and additionally observe that the performance of EQGCs scales well with the depth of the model and does not suffer from barren plateu issues.

Stochastic Rising Bandits

- Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, Marcello Restelli
- abstract: This paper is in the field of stochastic Multi-Armed Bandits (MABs), i.e., those sequential selection techniques able to learn online using only the feedback given by the chosen option (a.k.a. arm). We study a particular case of the rested and restless bandits in which the arms' expected payoff is monotonically non-decreasing. This characteristic allows designing specifically crafted algorithms that exploit the regularity of the payoffs to provide tight regret bounds. We design an algorithm for the rested case (R-ed-UCB) and one for the restless case (R-less-UCB), providing a regret bound depending on the properties of the instance and, under certain circumstances, of $\widetilde{\mathcal{O}}(T^{\frac{2}{3}})$. We empirically compare our algorithms with state-of-the-art methods for non-stationary MABs over several synthetically generated tasks and an online model selection problem for a real-world

dataset. Finally, using synthetic and real-world data, we illustrate the effectiveness of the proposed approaches compared with state-of-the-art algorithms for the non-stationary bandits.

Minimizing Control for Credit Assignment with Strong Feedback

- Alexander Meulemans, Matilde Tristany Farinha, Maria R. Cervera, João Sacramento, Benjamin F. Grewe
- abstract: The success of deep learning ignited interest in whether the brain learns hierarchical representations using gradient-based learning. However, current biologically plausible methods for gradient-based credit assignment in deep neural networks need infinitesimally small feedback signals, which is problematic in biologically realistic noisy environments and at odds with experimental evidence in neuroscience showing that top-down feedback can significantly influence neural activity. Building upon deep feedback control (DFC), a recently proposed credit assignment method, we combine strong feedback influences on neural activity with gradient-based learning and show that this naturally leads to a novel view on neural network optimization. Instead of gradually changing the network weights towards configurations with low output loss, weight updates gradually minimize the amount of feedback required from a controller that drives the network to the supervised output label. Moreover, we show that the use of strong feedback in DFC allows learning forward and feedback connections simultaneously, using learning rules fully local in space and time. We complement our theoretical results with experiments on standard computer-vision benchmarks, showing competitive performance to backpropagation as well as robustness to noise. Overall, our work presents a fundamentally novel view of learning as control minimization, while sidestepping biologically unrealistic assumptions.

A Dynamical System Perspective for Lipschitz Neural Networks

- Laurent Meunier, Blaise J Delattre, Alexandre Araujo, Alexandre Allauzen
- abstract: The Lipschitz constant of neural networks has been established as a key quantity to enforce the robustness to adversarial examples. In this paper, we tackle the problem of building 1-Lipschitz Neural Networks. By studying Residual Networks from a continuous time dynamical system perspective, we provide a generic method to build 1-Lipschitz Neural Networks and show that some previous approaches are special cases of this framework. Then, we extend this reasoning and show that ResNet flows derived from convex potentials define 1-Lipschitz transformations, that lead us to define the Convex Potential Layer (CPL). A comprehensive set of experiments on several datasets demonstrates the scalability of our architecture and the benefits as an ℓ_2-provable defense against adversarial examples. Our code is available at \url{https://github.com/MILES-PSL/Convex-Potential-Layer}

Distribution Regression with Sliced Wasserstein Kernels

- Dimitri Meunier, Massimiliano Pontil, Carlo Ciliberto
- abstract: The problem of learning functions over spaces of probabilities - or distribution regression - is gaining significant interest in the machine learning community. The main challenge in these settings is to identify a suitable representation capturing all relevant properties of a distribution. The well-established approach in this sense is to use kernel mean embeddings, which lift kernel-induced similarity on the input domain at the probability level. This strategy effectively tackles the two-stage sampling nature of the problem, enabling one to derive estimators with strong statistical guarantees, such as universal consistency and excess risk bounds. However, kernel mean embeddings implicitly hinge on the maximum mean discrepancy (MMD), a metric on probabilities, which is not the most suited to capture geometrical relations between distributions. In contrast, optimal transport (OT) metrics, are potentially more appealing. In this work, we propose an OT-based estimator for distribution regression. We build on the Sliced Wasserstein distance to obtain an OT-based representation. We study the theoretical properties of a kernel ridge regression estimator based on such representation, for which we prove universal consistency and excess risk bounds. Preliminary experiments complement our theoretical findings by showing the effectiveness of the proposed approach and compare it with MMD-based estimators.

Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism

- Siqi Miao, Mia Liu, Pan Li
- abstract: Interpretable graph learning is in need as many scientific applications depend on learning models to collect insights from graph-structured data. Previous works mostly focused on using post-hoc approaches to interpret pre-trained models (graph neural networks in particular). They argue against inherently interpretable models because the good interpretability of these models is often at the cost of their prediction accuracy. However, those post-hoc methods often fail to provide stable interpretation and may extract features that are spuriously correlated with the task. In this work, we address these issues by proposing Graph Stochastic Attention (GSAT). Derived from the information bottleneck principle, GSAT injects stochasticity to the attention weights to block the information from task-irrelevant graph components while learning stochasticity-reduced attention to select task-relevant subgraphs for interpretation. The selected subgraphs provably do not contain patterns that are spuriously correlated with the task under some assumptions. Extensive experiments on eight datasets show that GSAT outperforms the state-of-the-art methods by up to 20% in interpretation AUC and 5% in prediction accuracy. Our code is available at https://github.com/Graph-COM/GSAT.

Modeling Structure with Undirected Neural Networks

- Tsvetomila Mihaylova, Vlad Niculae, Andre Martins
- abstract: Neural networks are powerful function estimators, leading to their status as a paradigm of choice for modeling structured data. However, unlike other structured representations that emphasize the modularity of the problem {–} e.g., factor graphs {–} neural networks are usually monolithic mappings from inputs to outputs, with a fixed computation order. This limitation prevents them from capturing different directions of computation and interaction between the modeled variables. In this paper, we combine the representational strengths of factor graphs and of neural networks, proposing undirected neural networks (UNNs): a flexible framework for specifying computations that can be performed in any order. For particular choices, our proposed models subsume and extend many existing architectures: feed-forward, recurrent, self-attention networks, auto-encoders, and networks with implicit layers. We demonstrate the effectiveness of undirected neural architectures, both unstructured and structured, on a range of tasks: tree-constrained dependency parsing, convolutional image classification, and sequence completion with attention. By varying the computation order, we show how a single UNN can be used both as a classifier and a prototype generator, and how it can fill in missing parts of an input sequence, making them a promising field for further research.

Universal Hopfield Networks: A General Framework for Single-Shot Associative Memory Models

- Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, Rafal Bogacz
- abstract: A large number of neural network models of associative memory have been proposed in the literature. These include the classical Hopfield networks (HNs), sparse distributed memories (SDMs), and more recently the modern continuous Hopfield networks (MCHNs), which possess close links with self-attention in machine learning. In this paper, we propose a general framework for understanding the operation of such memory networks as a sequence of three operations: similarity, separation, and projection. We derive all these memory models as instances of our general framework with differing similarity and separation functions. We extend the mathematical framework of Krotov et al (2020) to express general associative memory models using neural network dynamics with local computation, and derive a general energy function that is a Lyapunov function of the dynamics. Finally, using our framework, we empirically investigate the capacity of using different similarity functions for these associative memory models, beyond the dot product similarity measure, and demonstrate empirically that Euclidean or Manhattan distance similarity metrics perform substantially better in practice on many tasks, enabling a more robust retrieval and higher memory capacity than existing models.

Learning Stochastic Shortest Path with Linear Function Approximation

- Yifei Min, Jiafan He, Tianhao Wang, Quanquan Gu
- abstract: We study the stochastic shortest path (SSP) problem in reinforcement learning with linear function approximation, where the transition kernel is represented as a linear mixture of unknown models. We call this class of SSP problems as linear mixture SSPs. We propose a novel algorithm with Hoeffding-type confidence sets for learning the linear mixture SSP, which can attain an $\tilde{\mathcal{O}}(d B_{\star}^{1.5}\sqrt{K/c_{\min}})$ regret. Here K is the number of episodes, d is the dimension of the feature mapping in the mixture model, $B_{\star}$ bounds the expected cumulative cost of the optimal policy, and $c_{\min}>0$ is the lower bound of the cost function. Our algorithm also applies to the case when $c_{\min} = 0$, and an $\tilde{\mathcal{O}}(K^{2/3})$ regret is guaranteed. To the best of our knowledge, this is the first algorithm with a sublinear regret guarantee for learning linear mixture SSP. Moreover, we design a refined Bernstein-type confidence set and propose an improved algorithm, which provably achieves an $\tilde{\mathcal{O}}(d B_{\star}\sqrt{K/c_{\min}})$ regret. In complement to the regret upper bounds, we also prove a lower bound of $\Omega(dB_{\star} \sqrt{K})$. Hence, our improved algorithm matches the lower bound up to a $1/\sqrt{c_{\min}}$ factor and poly-logarithmic factors, achieving a near-optimal regret guarantee.

Prioritized Training on Points that are Learnable, Worth Learning, and not yet Learnt

- Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, Yarin Gal
- abstract: Training on web-scale data can take months. But much computation and time is wasted on redundant and noisy points that are already learnt or not learnable. To accelerate training, we introduce Reducible Holdout Loss Selection (RHO-LOSS), a simple but principled technique which selects approximately those points for training that most reduce the model's generalization loss. As a result, RHO-LOSS mitigates the weaknesses of existing data selection methods: techniques from the optimization literature typically select "hard" (e.g. high loss) points, but such points are often noisy (not learnable) or less task-relevant. Conversely, curriculum learning prioritizes "easy" points, but such points need not be trained on once learned. In contrast, RHO-LOSS selects points that are learnable, worth learning, and not yet learnt. RHO-LOSS trains in far fewer steps than prior art, improves accuracy, and speeds up training on a wide range of datasets, hyperparameters, and architectures (MLPs, CNNs, and BERT). On the large web-scraped image dataset Clothing-1M, RHO-LOSS trains in 18x fewer steps and reaches 2% higher final accuracy than uniform data shuffling.

POEM: Out-of-Distribution Detection with Posterior Sampling

- Yifei Ming, Ying Fan, Yixuan Li
- abstract: Out-of-distribution (OOD) detection is indispensable for machine learning models deployed in the open world. Recently, the use of an auxiliary outlier dataset during training (also known as outlier exposure) has shown promising performance. As the sample space for potential OOD data can be prohibitively large, sampling informative outliers is essential. In this work, we propose a novel posterior sampling based outlier mining framework, POEM, which facilitates efficient use of outlier data and promotes learning a compact decision boundary between ID and OOD data for improved detection. We show that POEM establishes state-of-the-art performance on common benchmarks. Compared to the current best method that uses a greedy sampling strategy, POEM improves the relative performance by 42.0% and 24.2% (FPR95) on CIFAR-10 and CIFAR-100, respectively. We further provide theoretical insights on the effectiveness of POEM for OOD detection.

A Simple Reward-free Approach to Constrained Reinforcement Learning

- Sobhan Miryoosefi, Chi Jin
- abstract: In constrained reinforcement learning (RL), a learning agent seeks to not only optimize the overall reward but also satisfy the additional safety, diversity, or budget constraints. Consequently, existing constrained RL solutions require several new algorithmic ingredients that are notably different from standard RL. On the other hand, reward-free RL is independently developed in the unconstrained literature, which learns the transition dynamics without using the reward information, and thus naturally capable of addressing RL with multiple objectives under the common dynamics. This paper bridges reward-free RL and constrained RL. Particularly, we propose a simple meta-algorithm such that given any reward-free RL oracle, the approachability and constrained RL problems can be directly solved with negligible overheads in sample complexity. Utilizing the existing reward-free RL solvers, our framework provides sharp sample complexity results for constrained RL in the tabular MDP setting, matching the best existing results up to a factor of horizon dependence; our framework directly extends to a setting of tabular two-player Markov games, and gives a new result for constrained RL with linear function approximation.

Wide Neural Networks Forget Less Catastrophically

- Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, Mehrdad Farajtabar
- abstract: A primary focus area in continual learning research is alleviating the "catastrophic forgetting" problem in neural networks by designing new algorithms that are more robust to the distribution shifts. While the recent progress in continual learning literature is encouraging, our understanding of what properties of neural networks contribute to catastrophic forgetting is still limited. To address this, instead of focusing on continual learning algorithms, in this work, we focus on the model itself and study the impact of "width" of the neural network architecture on catastrophic forgetting, and show that width has a surprisingly significant effect on forgetting. To explain this effect, we study the learning dynamics of the network from various perspectives such as gradient orthogonality, sparsity, and lazy training regime. We provide potential explanations that are consistent with the empirical results across different architectures and continual learning benchmarks.

Proximal and Federated Random Reshuffling

- Konstantin Mishchenko, Ahmed Khaled, Peter Richtarik
- abstract: Random Reshuffling (RR), also known as Stochastic Gradient Descent (SGD) without replacement, is a popular and theoretically grounded method for finite-sum minimization. We propose two new algorithms: Proximal and Federated Random Reshuffling (ProxRR and FedRR). The first algorithm, ProxRR, solves composite finite-sum minimization problems in which the objective is the sum of a (potentially non-smooth) convex regularizer and an average of n smooth objectives. ProxRR evaluates the proximal operator once per epoch only. When the proximal operator is expensive to compute, this small difference makes ProxRR up to n times faster than algorithms that evaluate the proximal operator in every iteration, such as proximal (stochastic) gradient descent. We give examples of practical optimization tasks where the proximal operator is difficult to compute and ProxRR has a clear advantage. One such task is federated or distributed optimization, where the evaluation of the proximal operator corresponds to communication across the network. We obtain our second algorithm, FedRR, as a special case of ProxRR applied to federated optimization, and prove it has a smaller communication footprint than either distributed gradient descent or Local SGD. Our theory covers both constant and decreasing stepsizes, and allows for importance resampling schemes that can improve conditioning, which may be of independent interest. Our theory covers both convex and nonconvex regimes. Finally, we corroborate our results with experiments on real data sets.

ProxSkip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration! Finally!

- Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, Peter Richtarik

- abstract: We introduce ProxSkip—a surprisingly simple and provably efficient method for minimizing the sum of a smooth (f) and an expensive nonsmooth proximable (ψ) function. The canonical approach to solving such problems is via the proximal gradient descent (ProxGD) algorithm, which is based on the evaluation of the gradient of f and the prox operator of ψ in each iteration. In this work we are specifically interested in the regime in which the evaluation of prox is costly relative to the evaluation of the gradient, which is the case in many applications. ProxSkip allows for the expensive prox operator to be skipped in most iterations: while its iteration complexity is $\mathcal{O}(\kappa \log \nicefrac{1}{\varepsilon})$, where κ is the condition number of f, the number of prox evaluations is $\mathcal{O}(\sqrt{\kappa} \log \nicefrac{1}{\varepsilon})$ only. Our main motivation comes from federated learning, where evaluation of the gradient operator corresponds to taking a local GD step independently on all devices, and evaluation of prox corresponds to (expensive) communication in the form of gradient averaging. In this context, ProxSkip offers an effective acceleration of communication complexity. Unlike other local gradient-type methods, such as FedAvg, SCAFFOLD, S-Local-GD and FedLin, whose theoretical communication complexity is worse than, or at best matching, that of vanilla GD in the heterogeneous data regime, we obtain a provable and large improvement without any heterogeneity-bounding assumptions.

Fast Convex Optimization for Two-Layer ReLU Networks: Equivalent Model Classes and Cone Decompositions

- Aaron Mishkin, Arda Sahiner, Mert Pilanci
- abstract: We develop fast algorithms and robust software for convex optimization of two-layer neural networks with ReLU activation functions. Our work leverages a convex re-formulation of the standard weight-decay penalized training problem as a set of group-l1-regularized data-local models, where locality is enforced by polyhedral cone constraints. In the special case of zero-regularization, we show that this problem is exactly equivalent to unconstrained optimization of a convex "gated ReLU" network. For problems with non-zero regularization, we show that convex gated ReLU models obtain data-dependent approximation bounds for the ReLU training problem. To optimize the convex re-formulations, we develop an accelerated proximal gradient method and a practical augmented Lagrangian solver. We show that these approaches are faster than standard training heuristics for the non-convex problem, such as SGD, and outperform commercial interior-point solvers. Experimentally, we verify our theoretical results, explore the group-l1 regularization path, and scale convex optimization for neural networks to image classification on MNIST and CIFAR-10.

Memory-Based Model Editing at Scale

- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, Chelsea Finn
- abstract: Even the largest neural networks make errors, and once-correct predictions can become invalid as the world changes. Model editors make local updates to the behavior of base (pre-trained) models to inject updated knowledge or correct undesirable behaviors. Existing model editors have shown promise, but also suffer from insufficient expressiveness: they struggle to accurately model an edit's intended scope (examples affected by the edit), leading to inaccurate predictions for test inputs loosely related to the edit, and they often fail altogether after many edits. As a higher-capacity alternative, we propose Semi-Parametric Editing with a Retrieval-Augmented Counterfactual Model (SERAC), which stores edits in an explicit memory and learns to reason over them to modulate the base model's predictions as needed. To enable more rigorous evaluation of model editors, we introduce three challenging language model editing problems based on question answering, fact-checking, and dialogue generation. We find that only SERAC achieves high performance on all three problems, consistently outperforming existing approaches to model editing by a significant margin. Code, data, and additional project information will be made available at https://sites.google.com/view/serac-editing.

Invariant Ancestry Search

- Phillip B Mogensen, Nikolaj Thams, Jonas Peters
- abstract: Recently, methods have been proposed that exploit the invariance of prediction models with respect to changing environments to infer subsets of the causal parents of a response variable. If the environments influence only few of the underlying mechanisms, the subset identified by invariant causal prediction (ICP), for example, may be small, or even empty. We introduce the concept of minimal invariance and propose invariant ancestry search (IAS). In its population version, IAS outputs a set which contains only ancestors of the response and is a superset of the output of ICP. When applied to data, corresponding guarantees hold asymptotically if the underlying test for invariance has asymptotic level and power. We develop scalable algorithms and perform experiments on simulated and real data.

Differentially Private Community Detection for Stochastic Block Models

- Mohamed S Mohamed, Dung Nguyen, Anil Vullikanti, Ravi Tandon
- abstract: The goal of community detection over graphs is to recover underlying labels/attributes of users (e.g., political affiliation) given the connectivity between users. There has been significant recent progress on understanding the fundamental limits of community detection when the graph is generated from a stochastic block model (SBM). Specifically, sharp information theoretic limits and efficient algorithms have been obtained for SBMs as a function of p and q, which represent the intra-community and inter-community connection probabilities. In this paper, we study the community detection problem while preserving the privacy of the individual connections between the vertices. Focusing on the notion of (ϵ, δ)-edge differential privacy (DP), we seek to understand the fundamental tradeoffs between (p, q), DP budget (ϵ, δ), and computational efficiency for exact recovery of community labels. To this end, we present and analyze the associated information-theoretic tradeoffs for three differentially private community recovery mechanisms: a) stability based mechanism; b) sampling based mechanisms; and c) graph perturbation mechanisms. Our main findings are that stability and sampling based mechanisms lead to a superior tradeoff between (p,q) and the privacy budget (ϵ, δ); however this comes at the expense of higher computational complexity. On the other hand, albeit low complexity, graph perturbation mechanisms require the privacy budget ϵ to scale as $\Omega(\log(n))$ for exact recovery.

A Multi-objective / Multi-task Learning Framework Induced by Pareto Stationarity

- Michinari Momma, Chaosheng Dong, Jia Liu
- abstract: Multi-objective optimization (MOO) and multi-task learning (MTL) have gained much popularity with prevalent use cases such as production model development of regression / classification / ranking models with MOO, and training deep learning models with MTL. Despite the long history of research in MOO, its application to machine learning requires development of solution strategy, and algorithms have recently been developed to solve specific problems such as discovery of any Pareto optimal (PO) solution, and that with a particular form of preference. In this paper, we develop a novel and generic framework to discover a PO solution with multiple forms of preferences. It allows us to formulate a generic MOO / MTL problem to express a preference, which is solved to achieve both alignment with the preference and PO, at the same time. Specifically, we apply the framework to solve the weighted Chebyshev problem and an extension of that. The former is known as a method to discover the Pareto front, the latter helps to find a model that outperforms an existing model with only one run. Experimental results demonstrate not only the method achieves competitive performance with existing methods, but also it allows us to achieve the performance from different forms of preferences.

EqR: Equivariant Representations for Data-Efficient Reinforcement Learning

- Arnab Kumar Mondal, Vineet Jain, Kaleem Siddiqi, Siamak Ravanbakhsh
- abstract: We study a variety of notions of equivariance as an inductive bias in Reinforcement Learning (RL). In particular, we propose new mechanisms for learning representations that are equivariant to both the agent's action, as well as symmetry transformations of the state-action pairs. Whereas prior work on exploiting symmetries in deep RL can only incorporate predefined linear transformations, our approach allows non-linear symmetry transformations of state-action pairs to be learned from the data. This is achieved through 1) equivariant Lie algebraic parameterization of state and action

encodings, 2) equivariant latent transition models, and 3) the incorporation of symmetry-based losses. We demonstrate the advantages of our method, which we call Equivariant representations for RL (EqR), for Atari games in a data-efficient setting limited to 100K steps of interactions with the environment.

Feature and Parameter Selection in Stochastic Linear Bandits

- Ahmadreza Moradipari, Berkay Turan, Yasin Abbasi-Yadkori, Mahnoosh Alizadeh, Mohammad Ghavamzadeh
- abstract: We study two model selection settings in stochastic linear bandits (LB). In the first setting, which we refer to as feature selection, the expected reward of the LB problem is in the linear span of at least one of M feature maps (models). In the second setting, the reward parameter of the LB problem is arbitrarily selected from M models represented as (possibly) overlapping balls in $\mathbb R^d$. However, the agent only has access to misspecified models, i.e., estimates of the centers and radii of the balls. We refer to this setting as parameter selection. For each setting, we develop and analyze a computationally efficient algorithm that is based on a reduction from bandits to full-information problems. This allows us to obtain regret bounds that are not worse (up to a $\sqrt{\log M}$ factor) than the case where the true model is known. This is the best reported dependence on the number of models M in these settings. Finally, we empirically show the effectiveness of our algorithms using synthetic and real-world experiments.

Power-Law Escape Rate of SGD

- Takashi Mori, Liu Ziyin, Kangqiao Liu, Masahito Ueda
- abstract: Stochastic gradient descent (SGD) undergoes complicated multiplicative noise for the mean-square loss. We use this property of SGD noise to derive a stochastic differential equation (SDE) with simpler additive noise by performing a random time change. Using this formalism, we show that the log loss barrier $\Delta\log L=\log[L(\theta^s)/L(\theta^*)]$ *between a local minimum θ^** and a saddle θ^s determines the escape rate of SGD from the local minimum, contrary to the previous results borrowing from physics that the linear loss barrier $\Delta L=L(\theta^s)-L(\theta^*)$ *decides the escape rate. Our escape-rate formula strongly depends on the typical magnitude h^** and the number n of the outlier eigenvalues of the Hessian. This result explains an empirical fact that SGD prefers flat minima with low effective dimensions, giving an insight into implicit biases of SGD.

Rethinking Fano's Inequality in Ensemble Learning

- Terufumi Morishita, Gaku Morio, Shota Horiguchi, Hiroaki Ozaki, Nobuo Nukaga
- abstract: We propose a fundamental theory on ensemble learning that evaluates a given ensemble system by a well-grounded set of metrics. Previous studies used a variant of Fano's inequality of information theory and derived a lower bound of the classification error rate on the basis of the accuracy and diversity of models. We revisit the original Fano's inequality and argue that the studies did not take into account the information lost when multiple model predictions are combined into a final prediction. To address this issue, we generalize the previous theory to incorporate the information loss. Further, we empirically validate and demonstrate the proposed theory through extensive experiments on actual systems. The theory reveals the strengths and weaknesses of systems on each metric, which will push the theoretical understanding of ensemble learning and give us insights into designing systems.

SpeqNets: Sparsity-aware permutation-equivariant graph networks

- Christopher Morris, Gaurav Rattan, Sandra Kiefer, Siamak Ravanbakhsh
- abstract: While message-passing graph neural networks have clear limitations in approximating permutation-equivariant functions over graphs or general relational data, more expressive, higher-order graph neural networks do not scale to large graphs. They either operate on k-order tensors or consider all k-node subgraphs, implying an exponential dependence on k in memory requirements, and do not adapt to the sparsity of the graph. By introducing new heuristics for the graph isomorphism problem, we devise a class of universal, permutation-equivariant graph networks, which, unlike previous architectures, offer a fine-grained control between expressivity and scalability and adapt to the sparsity of the graph. These architectures lead to vastly reduced computation times compared to standard higher-order graph networks in the supervised node- and graph-level classification and regression regime while significantly improving standard graph neural network and graph kernel architectures in terms of predictive performance.

CtrlFormer: Learning Transferable State Representation for Visual Control via Transformer

- Yao Mark Mu, Shoufa Chen, Mingyu Ding, Jianyu Chen, Runjian Chen, Ping Luo
- abstract: Transformer has achieved great successes in learning vision and language representation, which is general across various downstream tasks. In visual control, learning transferable state representation that can transfer between different control tasks is important to reduce the training sample size. However, porting Transformer to sample-efficient visual control remains a challenging and unsolved problem. To this end, we propose a novel Control Transformer (CtrlFormer), possessing many appealing benefits that prior arts do not have. Firstly, CtrlFormer jointly learns self-attention mechanisms between visual tokens and policy tokens among different control tasks, where multitask representation can be learned and transferred without catastrophic forgetting. Secondly, we carefully design a contrastive reinforcement learning paradigm to train CtrlFormer, enabling it to achieve high sample efficiency, which is important in control problems. For example, in the DMControl benchmark, unlike recent advanced methods that failed by producing a zero score in the "Cartpole" task after transfer learning with 100k samples, CtrlFormer can achieve a state-of-the-art score with only 100k samples while maintaining the performance of previous tasks. The code and models are released in our project homepage.

Generalized Beliefs for Cooperative AI

- Darius Muglich, Luisa M Zintgraf, Christian A Schroeder De Witt, Shimon Whiteson, Jakob Foerster
- abstract: Self-play is a common method for constructing solutions in Markov games that can yield optimal policies in collaborative settings. However, these policies often adopt highly-specialized conventions that make playing with a novel partner difficult. To address this, recent approaches rely on encoding symmetry and convention-awareness into policy training, but these require strong environmental assumptions and can complicate policy training. To overcome this, we propose moving the learning of conventions to the belief space. Specifically, we propose a belief learning paradigm that can maintain beliefs over rollouts of policies not seen at training time, and can thus decode and adapt to novel conventions at test time. We show how to leverage this belief model for both search and training of a best response over a pool of policies to greatly improve zero-shot coordination. We also show how our paradigm promotes explainability and interpretability of nuanced agent conventions.

Bounding the Width of Neural Networks via Coupled Initialization A Worst Case Analysis

- Alexander Munteanu, Simon Omlor, Zhao Song, David Woodruff
- abstract: A common method in training neural networks is to initialize all the weights to be independent Gaussian vectors. We observe that by instead initializing the weights into independent pairs, where each pair consists of two identical Gaussian vectors, we can significantly improve the convergence analysis. While a similar technique has been studied for random inputs [Daniely, NeurIPS 2020], it has not been analyzed with arbitrary inputs. Using this technique, we show how to significantly reduce the number of neurons required for two-layer ReLU networks, both in the under-parameterized setting with logistic loss, from roughly γ^{-8} [Ji and Telgarsky, ICLR 2020] to γ^{-2}, where γ denotes the separation margin with a Neural Tangent Kernel, as well as in the over-parameterized setting with squared loss, from roughly n^4 [Song and Yang, 2019] to n^2, implicitly also improving the recent running time bound of [Brand, Peng, Song and Weinstein, ITCS 2021]. For the under-parameterized setting we also prove new lower bounds that improve upon prior work, and that under certain assumptions, are best possible.

Constants Matter: The Performance Gains of Active Learning

- Stephen O Mussmann, Sanjoy Dasgupta
- abstract: Within machine learning, active learning studies the gains in performance made possible by adaptively selecting data points to label. In this work, we show through upper and lower bounds, that for a simple benign setting of well-specified logistic regression on a uniform distribution over a sphere, the expected excess error of both active learning and random sampling have the same inverse proportional dependence on the number of samples. Importantly, due to the nature of lower bounds, any more general setting does not allow a better dependence on the number of samples. Additionally, we show a variant of uncertainty sampling can achieve a faster rate of convergence than random sampling by a factor of the Bayes error, a recent empirical observation made by other work. Qualitatively, this work is pessimistic with respect to the asymptotic dependence on the number of samples, but optimistic with respect to finding performance gains in the constants.

On the Generalization Analysis of Adversarial Learning

- Waleed Mustafa, Yunwen Lei, Marius Kloft
- abstract: Many recent studies have highlighted the susceptibility of virtually all machine-learning models to adversarial attacks. Adversarial attacks are imperceptible changes to an input example of a given prediction model. Such changes are carefully designed to alter the otherwise correct prediction of the model. In this paper, we study the generalization properties of adversarial learning. In particular, we derive high-probability generalization bounds on the adversarial risk in terms of the empirical adversarial risk, the complexity of the function class and the adversarial noise set. Our bounds are generally applicable to many models, losses, and adversaries. We showcase its applicability by deriving adversarial generalization bounds for the multi-class classification setting and various prediction models (including linear models and Deep Neural Networks). We also derive optimistic adversarial generalization bounds for the case of smooth losses. These are the first fast-rate bounds valid for adversarial deep learning to the best of our knowledge.

Universal and data-adaptive algorithms for model selection in linear contextual bandits

- Vidya K Muthukumar, Akshay Krishnamurthy
- abstract: Model selection in contextual bandits is an important complementary problem to regret minimization with respect to a fixed model class. We consider the simplest non-trivial instance of model-selection: distinguishing a simple multi-armed bandit problem from a linear contextual bandit problem. Even in this instance, current state-of-the-art methods explore in a suboptimal manner and require strong "feature-diversity" conditions. In this paper, we introduce new algorithms that a) explore in a data-adaptive manner, and b) provide model selection guarantees of the form O(d^{\alpha} T^{1 - \alpha}) with no feature diversity conditions whatsoever, where d denotes the dimension of the linear model and T denotes the total number of rounds. The first algorithm enjoys a "best-of-both-worlds" property, recovering two prior results that hold under distinct distributional assumptions, simultaneously. The second removes distributional assumptions altogether, expanding the scope for tractable model selection. Our approach extends to model selection among nested linear contextual bandits under some additional assumptions.

The Importance of Non-Markovianity in Maximum State Entropy Exploration

- Mirco Mutti, Riccardo De Santi, Marcello Restelli
- abstract: In the maximum state entropy exploration framework, an agent interacts with a reward-free environment to learn a policy that maximizes the entropy of the expected state visitations it is inducing. Hazan et al. (2019) noted that the class of Markovian stochastic policies is sufficient for the maximum state entropy objective, and exploiting non-Markovianity is generally considered pointless in this setting. In this paper, we argue that non-Markovianity is instead paramount for maximum state entropy exploration in a finite-sample regime. Especially, we recast the objective to target the expected entropy of the induced state visitations in a single trial. Then, we show that the class of non-Markovian deterministic policies is sufficient for the introduced objective, while Markovian policies suffer non-zero regret in general. However, we prove that the problem of finding an optimal non-Markovian policy is NP-hard. Despite this negative result, we discuss avenues to address the problem in a tractable way and how non-Markovian exploration could benefit the sample efficiency of online reinforcement learning in future works.

PAC-Net: A Model Pruning Approach to Inductive Transfer Learning

- Sanghoon Myung, In Huh, Wonik Jang, Jae Myung Choe, Jisu Ryu, Daesin Kim, Kee-Eung Kim, Changwook Jeong
- abstract: Inductive transfer learning aims to learn from a small amount of training data for the target task by utilizing a pre-trained model from the source task. Most strategies that involve large-scale deep learning models adopt initialization with the pre-trained model and fine-tuning for the target task. However, when using over-parameterized models, we can often prune the model without sacrificing the accuracy of the source task. This motivates us to adopt model pruning for transfer learning with deep learning models. In this paper, we propose PAC-Net, a simple yet effective approach for transfer learning based on pruning. PAC-Net consists of three steps: Prune, Allocate, and Calibrate (PAC). The main idea behind these steps is to identify essential weights for the source task, fine-tune on the source task by updating the essential weights, and then calibrate on the target task by updating the remaining redundant weights. Under the various and extensive set of inductive transfer learning experiments, we show that our method achieves state-of-the-art performance by a large margin.

AutoSNN: Towards Energy-Efficient Spiking Neural Networks

- Byunggook Na, Jisoo Mok, Seongsik Park, Dongjin Lee, Hyeokjun Choe, Sungroh Yoon
- abstract: Spiking neural networks (SNNs) that mimic information transmission in the brain can energy-efficiently process spatio-temporal information through discrete and sparse spikes, thereby receiving considerable attention. To improve accuracy and energy efficiency of SNNs, most previous studies have focused solely on training methods, and the effect of architecture has rarely been studied. We investigate the design choices used in the previous studies in terms of the accuracy and number of spikes and figure out that they are not best-suited for SNNs. To further improve the accuracy and reduce the spikes generated by SNNs, we propose a spike-aware neural architecture search framework called AutoSNN. We define a search space consisting of architectures without undesirable design choices. To enable the spike-aware architecture search, we introduce a fitness that considers both the accuracy and number of spikes. AutoSNN successfully searches for SNN architectures that outperform hand-crafted SNNs in accuracy and energy efficiency. We thoroughly demonstrate the effectiveness of AutoSNN on various datasets including neuromorphic datasets.

Implicit Bias of the Step Size in Linear Diagonal Neural Networks

- Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, Daniel Soudry
- abstract: Focusing on diagonal linear networks as a model for understanding the implicit bias in underdetermined models, we show how the gradient descent step size can have a large qualitative effect on the implicit bias, and thus on generalization ability. In particular, we show how using large step size for non-centered data can change the implicit bias from a "kernel" type behavior to a "rich" (sparsity-inducing) regime — even when gradient flow, studied in previous works, would not escape the "kernel" regime. We do so by using dynamic stability, proving that convergence to dynamically stable global minima entails a bound on some weighted ℓ_1-norm of the linear predictor, i.e. a "rich" regime. We prove this leads to good generalization in a sparse regression setting.

DNNR: Differential Nearest Neighbors Regression

- Youssef Nader, Leon Sixt, Tim Landgraf
- abstract: K-nearest neighbors (KNN) is one of the earliest and most established algorithms in machine learning. For regression tasks, KNN averages the targets within a neighborhood which poses a number of challenges: the neighborhood definition is crucial for the predictive performance as neighbors might be selected based on uninformative features, and averaging does not account for how the function changes locally. We propose a novel method called Differential Nearest Neighbors Regression (DNNR) that addresses both issues simultaneously: during training, DNNR estimates local gradients to scale the features; during inference, it performs an n-th order Taylor approximation using estimated gradients. In a large-scale evaluation on over 250 datasets, we find that DNNR performs comparably to state-of-the-art gradient boosting methods and MLPs while maintaining the simplicity and transparency of KNN. This allows us to derive theoretical error bounds and inspect failures. In times that call for transparency of ML models, DNNR provides a good balance between performance and interpretability.

Overcoming Oscillations in Quantization-Aware Training

- Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, Tijmen Blankevoort
- abstract: When training neural networks with simulated quantization, we observe that quantized weights can, rather unexpectedly, oscillate between two grid-points. The importance of this effect and its impact on quantization-aware training (QAT) are not well-understood or investigated in literature. In this paper, we delve deeper into the phenomenon of weight oscillations and show that it can lead to a significant accuracy degradation due to wrongly estimated batch-normalization statistics during inference and increased noise during training. These effects are particularly pronounced in low-bit ($\leq$ 4-bits) quantization of efficient networks with depth-wise separable layers, such as MobileNets and EfficientNets. In our analysis we investigate several previously proposed QAT algorithms and show that most of these are unable to overcome oscillations. Finally, we propose two novel QAT algorithms to overcome oscillations during training: oscillation dampening and iterative weight freezing. We demonstrate that our algorithms achieve state-of-the-art accuracy for low-bit (3 & 4 bits) weight and activation quantization of efficient architectures, such as MobileNetV2, MobileNetV3, and EfficentNet-lite on ImageNet. Our source code is available at https://github.com/qualcomm-ai-research/oscillations-qat.

Strategic Representation

- Vineet Nair, Ganesh Ghalme, Inbal Talgam-Cohen, Nir Rosenfeld
- abstract: Humans have come to rely on machines for reducing excessive information to manageable representations. But this reliance can be abused – strategic machines might craft representations that manipulate their users. How can a user make good choices based on strategic representations? We formalize this as a learning problem, and pursue algorithms for decision-making that are robust to manipulation. In our main setting of interest, the system represents attributes of an item to the user, who then decides whether or not to consume. We model this interaction through the lens of strategic classification (Hardt et al. 2016), reversed: the user, who learns, plays first; and the system, which responds, plays second. The system must respond with representations that reveal ‘nothing but the truth’ but need not reveal the entire truth. Thus, the user faces the problem of learning set functions under strategic subset selection, which presents distinct algorithmic and statistical challenges. Our main result is a learning algorithm that minimizes error despite strategic representations, and our theoretical analysis sheds light on the trade-off between learning effort and susceptibility to manipulation.

Improving Ensemble Distillation With Weight Averaging and Diversifying Perturbation

- Giung Nam, Hyungi Lee, Byeongho Heo, Juho Lee
- abstract: Ensembles of deep neural networks have demonstrated superior performance, but their heavy computational cost hinders applying them for resource-limited environments. It motivates distilling knowledge from the ensemble teacher into a smaller student network, and there are two important design choices for this ensemble distillation: 1) how to construct the student network, and 2) what data should be shown during training. In this paper, we propose a weight averaging technique where a student with multiple subnetworks is trained to absorb the functional diversity of ensemble teachers, but then those subnetworks are properly averaged for inference, giving a single student network with no additional inference cost. We also propose a perturbation strategy that seeks inputs from which the diversities of teachers can be better transferred to the student. Combining these two, our method significantly improves upon previous methods on various image classification tasks.

Measuring Representational Robustness of Neural Networks Through Shared Invariances

- Vedant Nanda, Till Speicher, Camila Kolling, John P Dickerson, Krishna Gummadi, Adrian Weller
- abstract: A major challenge in studying robustness in deep learning is defining the set of “meaningless” perturbations to which a given Neural Network (NN) should be invariant. Most work on robustness implicitly uses a human as the reference model to define such perturbations. Our work offers a new view on robustness by using another reference NN to define the set of perturbations a given NN should be invariant to, thus generalizing the reliance on a reference “human NN” to any NN. This makes measuring robustness equivalent to measuring the extent to which two NNs share invariances. We propose a measure called \stir, which faithfully captures the extent to which two NNs share invariances. \stir re-purposes existing representation similarity measures to make them suitable for measuring shared invariances. Using our measure, we are able to gain insights about how shared invariances vary with changes in weight initialization, architecture, loss functions, and training dataset. Our implementation is available at: \url{https://github.com/nvedant07/STIR}.

Tight and Robust Private Mean Estimation with Few Users

- Shyam Narayanan, Vahab Mirrokni, Hossein Esfandiari
- abstract: In this work, we study high-dimensional mean estimation under user-level differential privacy, and design an (ε,δ)-differentially private mechanism using as few users as possible. In particular, we provide a nearly optimal trade-off between the number of users and the number of samples per user required for private mean estimation, even when the number of users is as low as $O(\frac{1}{\varepsilon}\log\frac{1}{\delta})$. Interestingly, this bound on the number of users is independent of the dimension (though the number of samples per user is allowed to depend polynomially on the dimension), unlike the previous work that requires the number of users to depend polynomially on the dimension. This resolves a problem first proposed by Amin et al. (2019). Moreover, our mechanism is robust against corruptions in up to $49%$ of the users. Finally, our results also apply to optimal algorithms for privately learning discrete distributions with few users, answering a question of Liu et al. (2020), and a broader range of problems such as stochastic convex optimization and a variant of stochastic gradient descent via a reduction to differentially private mean estimation.

Fast Aquatic Swimmer Optimization with Differentiable Projective Dynamics and Neural Network Hydrodynamic Models

- Elvis Nava, John Z Zhang, Mike Yan Michelis, Tao Du, Pingchuan Ma, Benjamin F. Grewe, Wojciech Matusik, Robert Kevin Katzschmann
- abstract: Aquatic locomotion is a classic fluid-structure interaction (FSI) problem of interest to biologists and engineers. Solving the fully coupled FSI equations for incompressible Navier-Stokes and finite elasticity is computationally expensive. Optimizing robotic swimmer design within such a system generally involves cumbersome, gradient-free procedures on top of the already costly simulation. To address this challenge we present a novel, fully differentiable hybrid approach to FSI that combines a 2D direct numerical simulation for the deformable solid structure of the swimmer and a physics-constrained neural network surrogate to capture hydrodynamic effects of the fluid. For the deformable solid simulation of the swimmer’s body, we use state-of-the-art techniques from the field of computer graphics to speed up the finite-element method (FEM). For the fluid simulation, we use a U-Net architecture trained with a physics-based loss function to predict the flow field at each time step. The pressure and velocity field outputs from the neural network are sampled around the boundary of our swimmer using an immersed boundary method (IBM) to compute its swimming motion accurately and

efficiently. We demonstrate the computational efficiency and differentiability of our hybrid simulator on a 2D carangiform swimmer. Due to differentiability, the simulator can be used for computational design of controls for soft bodies immersed in fluids via direct gradient-based optimization.

Multi-Task Learning as a Bargaining Game

- Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, Ethan Fetaya
- abstract: In Multi-task learning (MTL), a joint model is trained to simultaneously make predictions for several tasks. Joint training reduces computation costs and improves data efficiency; however, since the gradients of these different tasks may conflict, training a joint model for MTL often yields lower performance than its corresponding single-task counterparts. A common method for alleviating this issue is to combine per-task gradients into a joint update direction using a particular heuristic. In this paper, we propose viewing the gradients combination step as a bargaining game, where tasks negotiate to reach an agreement on a joint direction of parameter update. Under certain assumptions, the bargaining problem has a unique solution, known as the Nash Bargaining Solution, which we propose to use as a principled approach to multi-task learning. We describe a new MTL optimization procedure, Nash-MTL, and derive theoretical guarantees for its convergence. Empirically, we show that Nash-MTL achieves state-of-the-art results on multiple MTL benchmarks in various domains.

Variational Inference for Infinitely Deep Neural Networks

- Achille Nazaret, David Blei
- abstract: We introduce the unbounded depth neural network (UDN), an infinitely deep probabilistic model that adapts its complexity to the training data. The UDN contains an infinite sequence of hidden layers and places an unbounded prior on a truncation L, the layer from which it produces its data. Given a dataset of observations, the posterior UDN provides a conditional distribution of both the parameters of the infinite neural network and its truncation. We develop a novel variational inference algorithm to approximate this posterior, optimizing a distribution of the neural network weights and of the truncation depth L, and without any upper limit on L. To this end, the variational family has a special structure: it models neural network weights of arbitrary depth, and it dynamically creates or removes free variational parameters as its distribution of the truncation is optimized. (Unlike heuristic approaches to model search, it is solely through gradient-based optimization that this algorithm explores the space of truncations.) We study the UDN on real and synthetic data. We find that the UDN adapts its posterior depth to the dataset complexity; it outperforms standard neural networks of similar computational complexity; and it outperforms other approaches to infinite-depth neural networks.

Stable Conformal Prediction Sets

- Eugene Ndiaye
- abstract: When one observes a sequence of variables $(x_1, y_1), \ldots, (x_n, y_n)$, Conformal Prediction (CP) is a methodology that allows to estimate a confidence set for y_{n+1} given x_{n+1} by merely assuming that the distribution of the data is exchangeable. CP sets have guaranteed coverage for any finite population size n. While appealing, the computation of such a set turns out to be infeasible in general, \eg when the unknown variable y_{n+1} is continuous. The bottleneck is that it is based on a procedure that readjusts a prediction model on data where we replace the unknown target by all its possible values in order to select the most probable one. This requires computing an infinite number of models, which often makes it intractable. In this paper, we combine CP techniques with classical algorithmic stability bounds to derive a prediction set computable with a single model fit. We demonstrate that our proposed confidence set does not lose any coverage guarantees while avoiding the need for data splitting as currently done in the literature. We provide some numerical experiments to illustrate the tightness of our estimation when the sample size is sufficiently large, on both synthetic and real datasets.

Discovering Generalizable Spatial Goal Representations via Graph-based Active Reward Learning

- Aviv Netanyahu, Tianmin Shu, Joshua Tenenbaum, Pulkit Agrawal
- abstract: In this work, we consider one-shot imitation learning for object rearrangement tasks, where an AI agent needs to watch a single expert demonstration and learn to perform the same task in different environments. To achieve a strong generalization, the AI agent must infer the spatial goal specification for the task. However, there can be multiple goal specifications that fit the given demonstration. To address this, we propose a reward learning approach, Graph-based Equivalence Mappings (GEM), that can discover spatial goal representations that are aligned with the intended goal specification, enabling successful generalization in unseen environments. Specifically, GEM represents a spatial goal specification by a reward function conditioned on i) a graph indicating important spatial relationships between objects and ii) state equivalence mappings for each edge in the graph indicating invariant properties of the corresponding relationship. GEM combines inverse reinforcement learning and active reward learning to efficiently improve the reward function by utilizing the graph structure and domain randomization enabled by the equivalence mappings. We conducted experiments with simulated oracles and with human subjects. The results show that GEM can drastically improve the generalizability of the learned goal representations over strong baselines.

Sublinear-Time Clustering Oracle for Signed Graphs

- Stefan Neumann, Pan Peng
- abstract: Social networks are often modeled using signed graphs, where vertices correspond to users and edges have a sign that indicates whether an interaction between users was positive or negative. The arising signed graphs typically contain a clear community structure in the sense that the graph can be partitioned into a small number of polarized communities, each defining a sparse cut and indivisible into smaller polarized sub-communities. We provide a local clustering oracle for signed graphs with such a clear community structure, that can answer membership queries, i.e., “Given a vertex v, which community does v belong to?”, in sublinear time by reading only a small portion of the graph. Formally, when the graph has bounded maximum degree and the number of communities is at most $O(\log n)$, then with $\tilde{O}(\sqrt{n}\operatorname{poly}(1/\varepsilon))$ preprocessing time, our oracle can answer each membership query in $\tilde{O}(\sqrt{n}\operatorname{poly}(1/\varepsilon))$ time, and it correctly classifies a $(1-\varepsilon)$-fraction of vertices w.r.t. a set of hidden planted ground-truth communities. Our oracle is desirable in applications where the clustering information is needed for only a small number of vertices. Previously, such local clustering oracles were only known for unsigned graphs; our generalization to signed graphs requires a number of new ideas and gives a novel spectral analysis of the behavior of random walks with signs. We evaluate our algorithm for constructing such an oracle and answering membership queries on both synthetic and real-world datasets, validating its performance in practice.

Improved Regret for Differentially Private Exploration in Linear MDP

- Dung Daniel T Ngo, Giuseppe Vietri, Steven Wu
- abstract: We study privacy-preserving exploration in sequential decision-making for environments that rely on sensitive data such as medical records. In particular, we focus on solving the problem of reinforcement learning (RL) subject to the constraint of (joint) differential privacy in the linear MDP setting, where both dynamics and rewards are given by linear functions. Prior work on this problem due to (Luyo et al., 2021) achieves a regret rate that has a dependence of O(K^{3/5}) on the number of episodes K. We provide a private algorithm with an improved regret rate with an optimal dependence of O($\sqrt{}$K) on the number of episodes. The key recipe for our stronger regret guarantee is the adaptivity in the policy update schedule, in which an update only occurs when sufficient changes in the data are detected. As a result, our algorithm benefits from low switching cost and only performs O(log(K)) updates, which greatly reduces the amount of privacy noise. Finally, in the most prevalent privacy regimes where the privacy parameter \epsilon is a constant, our algorithm incurs negligible privacy cost{—}in comparison with the existing non-private regret bounds, the additional regret due to privacy appears in lower-order terms.

A Framework for Learning to Request Rich and Contextually Useful Information from Humans

- Khanh X Nguyen, Yonatan Bisk, Hal Daumé Iii
- abstract: When deployed, AI agents will encounter problems that are beyond their autonomous problem-solving capabilities. Leveraging human assistance can help agents overcome their inherent limitations and robustly cope with unfamiliar situations. We present a general interactive framework that enables an agent to request and interpret rich, contextually useful information from an assistant that has knowledge about the task and the environment. We demonstrate the practicality of our framework on a simulated human-assisted navigation problem. Aided with an assistance-requesting policy learned by our method, a navigation agent achieves up to a 7{\texttimes} improvement in success rate on tasks that take place in previously unseen environments, compared to fully autonomous behavior. We show that the agent can take advantage of different types of information depending on the context, and analyze the benefits and challenges of learning the assistance-requesting policy when the assistant can recursively decompose tasks into subtasks.

Transformer Neural Processes: Uncertainty-Aware Meta Learning Via Sequence Modeling

- Tung Nguyen, Aditya Grover
- abstract: Neural Processes (NPs) are a popular class of approaches for meta-learning. Similar to Gaussian Processes (GPs), NPs define distributions over functions and can estimate uncertainty in their predictions. However, unlike GPs, NPs and their variants suffer from underfitting and often have intractable likelihoods, which limit their applications in sequential decision making. We propose Transformer Neural Processes (TNPs), a new member of the NP family that casts uncertainty-aware meta learning as a sequence modeling problem. We learn TNPs via an autoregressive likelihood-based objective and instantiate it with a novel transformer-based architecture that respects the inductive biases inherent to the problem structure, such as invariance to the observed data points and equivariance to the unobserved points. We further design knobs within the TNP architecture to tradeoff the increase in expressivity of the decoding distribution with extra computation. Empirically, we show that TNPs achieve state-of-the-art performance on various benchmark problems, outperforming all previous NP variants on meta regression, image completion, contextual multi-armed bandits, and Bayesian optimization.

Improving Transformers with Probabilistic Attention Keys

- Tam Minh Nguyen, Tan Minh Nguyen, Dung D. D. Le, Duy Khuong Nguyen, Viet-Anh Tran, Richard Baraniuk, Nhat Ho, Stanley Osher
- abstract: Multi-head attention is a driving force behind state-of-the-art transformers, which achieve remarkable performance across a variety of natural language processing (NLP) and computer vision tasks. It has been observed that for many applications, those attention heads learn redundant embedding, and most of them can be removed without degrading the performance of the model. Inspired by this observation, we propose Transformer with a Mixture of Gaussian Keys (Transformer-MGK), a novel transformer architecture that replaces redundant heads in transformers with a mixture of keys at each head. These mixtures of keys follow a Gaussian mixture model and allow each attention head to focus on different parts of the input sequence efficiently. Compared to its conventional transformer counterpart, Transformer-MGK accelerates training and inference, has fewer parameters, and requires fewer FLOPs to compute while achieving comparable or better accuracy across tasks. Transformer-MGK can also be easily extended to use with linear attention. We empirically demonstrate the advantage of Transformer-MGK in a range of practical applications, including language modeling and tasks that involve very long sequences. On the Wikitext-103 and Long Range Arena benchmark, Transformer-MGKs with 4 heads attain comparable or better performance to the baseline transformers with 8 heads.

On Transportation of Mini-batches: A Hierarchical Approach

- Khai Nguyen, Dang Nguyen, Quoc Dinh Nguyen, Tung Pham, Hung Bui, Dinh Phung, Trung Le, Nhat Ho
- abstract: Mini-batch optimal transport (m-OT) has been successfully used in practical applications that involve probability measures with a very high number of supports. The m-OT solves several smaller optimal transport problems and then returns the average of their costs and transportation plans. Despite its scalability advantage, the m-OT does not consider the relationship between mini-batches which leads to undesirable estimation. Moreover, the m-OT does not approximate a proper metric between probability measures since the identity property is not satisfied. To address these problems, we propose a novel mini-batch scheme for optimal transport, named Batch of Mini-batches Optimal Transport (BoMb-OT), that finds the optimal coupling between mini-batches and it can be seen as an approximation to a well-defined distance on the space of probability measures. Furthermore, we show that the m-OT is a limit of the entropic regularized version of the BoMb-OT when the regularized parameter goes to infinity. Finally, we carry out experiments on various applications including deep generative models, deep domain adaptation, approximate Bayesian computation, color transfer, and gradient flow to show that the BoMb-OT can be widely applied and performs well in various applications.

Improving Mini-batch Optimal Transport via Partial Transportation

- Khai Nguyen, Dang Nguyen, The-Anh Vu-Le, Tung Pham, Nhat Ho
- abstract: Mini-batch optimal transport (m-OT) has been widely used recently to deal with the memory issue of OT in large-scale applications. Despite their practicality, m-OT suffers from misspecified mappings, namely, mappings that are optimal on the mini-batch level but are partially wrong in the comparison with the optimal transportation plan between the original measures. Motivated by the misspecified mappings issue, we propose a novel mini-batch method by using partial optimal transport (POT) between mini-batch empirical measures, which we refer to as mini-batch partial optimal transport (m-POT). Leveraging the insight from the partial transportation, we explain the source of misspecified mappings from the m-OT and motivate why limiting the amount of transported masses among mini-batches via POT can alleviate the incorrect mappings. Finally, we carry out extensive experiments on various applications such as deep domain adaptation, partial domain adaptation, deep generative model, color transfer, and gradient flow to demonstrate the favorable performance of m-POT compared to current mini-batch methods.

Recurrent Model-Free RL Can Be a Strong Baseline for Many POMDPs

- Tianwei Ni, Benjamin Eysenbach, Ruslan Salakhutdinov
- abstract: Many problems in RL, such as meta-RL, robust RL, generalization in RL, and temporal credit assignment, can be cast as POMDPs. In theory, simply augmenting model-free RL with memory-based architectures, such as recurrent neural networks, provides a general approach to solving all types of POMDPs. However, prior work has found that such recurrent model-free RL methods tend to perform worse than more specialized algorithms that are designed for specific types of POMDPs. This paper revisits this claim. We find that careful architecture and hyperparameter decisions can often yield a recurrent model-free implementation that performs on par with (and occasionally substantially better than) more sophisticated recent techniques. We compare to 21 environments from 6 prior specialized methods and find that our implementation achieves greater sample efficiency and asymptotic performance than these methods on 18/21 environments. We also release a simple and efficient implementation of recurrent model-free RL for future work to use as a baseline for POMDPs.

Optimal Estimation of Policy Gradient via Double Fitted Iteration

- Chengzhuo Ni, Ruiqi Zhang, Xiang Ji, Xuezhou Zhang, Mengdi Wang
- abstract: Policy gradient (PG) estimation becomes a challenge when we are not allowed to sample with the target policy but only have access to a dataset generated by some unknown behavior policy. Conventional methods for off-policy PG estimation often suffer from either significant bias or exponentially large variance. In this paper, we propose the double Fitted PG estimation (FPG) algorithm. FPG can work with an arbitrary policy parameterization,

assuming access to a Bellman-complete value function class. In the case of linear value function approximation, we provide a tight finite-sample upper bound on policy gradient estimation error, that is governed by the amount of distribution mismatch measured in feature space. We also establish the asymptotic normality of FPG estimation error with a precise covariance characterization, which is further shown to be statistically optimal with a matching Cramer-Rao lower bound. Empirically, we evaluate the performance of FPG on both policy gradient estimation and policy optimization, using either softmax tabular or ReLU policy networks. Under various metrics, our results show that FPG significantly outperforms existing off-policy PG estimation methods based on importance sampling and variance reduction techniques.

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, Mark Chen
- abstract: Diffusion models have recently been shown to generate high-quality synthetic images, especially when paired with a guidance technique to trade off diversity for fidelity. We explore diffusion models for the problem of text-conditional image synthesis and compare two different guidance strategies: CLIP guidance and classifier-free guidance. We find that the latter is preferred by human evaluators for both photorealism and caption similarity, and often produces photorealistic samples. Samples from a 3.5 billion parameter text-conditional diffusion model using classifier-free guidance are favored by human evaluators to those from DALL-E, even when the latter uses expensive CLIP reranking. Additionally, we find that our models can be fine-tuned to perform image inpainting, enabling powerful text-driven image editing. We train a smaller model on a filtered dataset and release the code and weights at https://github.com/openai/glide-text2im.

Diffusion Models for Adversarial Purification

- Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, Animashree Anandkumar
- abstract: Adversarial purification refers to a class of defense methods that remove adversarial perturbations using a generative model. These methods do not make assumptions on the form of attack and the classification model, and thus can defend pre-existing classifiers against unseen threats. However, their performance currently falls behind adversarial training methods. In this work, we propose DiffPure that uses diffusion models for adversarial purification: Given an adversarial example, we first diffuse it with a small amount of noise following a forward diffusion process, and then recover the clean image through a reverse generative process. To evaluate our method against strong adaptive attacks in an efficient and scalable way, we propose to use the adjoint method to compute full gradients of the reverse generative process. Extensive experiments on three image datasets including CIFAR-10, ImageNet and CelebA-HQ with three classifier architectures including ResNet, WideResNet and ViT demonstrate that our method achieves the state-of-the-art results, outperforming current adversarial training and adversarial purification methods, often by a large margin.

The Primacy Bias in Deep Reinforcement Learning

- Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, Aaron Courville
- abstract: This work identifies a common flaw of deep reinforcement learning (RL) algorithms: a tendency to rely on early interactions and ignore useful evidence encountered later. Because of training on progressively growing datasets, deep RL agents incur a risk of overfitting to earlier experiences, negatively affecting the rest of the learning process. Inspired by cognitive science, we refer to this effect as the primacy bias. Through a series of experiments, we dissect the algorithmic aspects of deep RL that exacerbate this bias. We then propose a simple yet generally-applicable mechanism that tackles the primacy bias by periodically resetting a part of the agent. We apply this mechanism to algorithms in both discrete (Atari 100k) and continuous action (DeepMind Control Suite) domains, consistently improving their performance.

Causal Conceptions of Fairness and their Consequences

- Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, Sharad Goel
- abstract: Recent work highlights the role of causality in designing equitable decision-making algorithms. It is not immediately clear, however, how existing causal conceptions of fairness relate to one another, or what the consequences are of using these definitions as design principles. Here, we first assemble and categorize popular causal definitions of algorithmic fairness into two broad families: (1) those that constrain the effects of decisions on counterfactual disparities; and (2) those that constrain the effects of legally protected characteristics, like race and gender, on decisions. We then show, analytically and empirically, that both families of definitions almost always—in a measure theoretic sense—result in strongly Pareto dominated decision policies, meaning there is an alternative, unconstrained policy favored by every stakeholder with preferences drawn from a large, natural class. For example, in the case of college admissions decisions, policies constrained to satisfy causal fairness definitions would be disfavored by every stakeholder with neutral or positive preferences for both academic preparedness and diversity. Indeed, under a prominent definition of causal fairness, we prove the resulting policies require admitting all students with the same probability, regardless of academic qualifications or group membership. Our results highlight formal limitations and potential adverse consequences of common mathematical notions of causal fairness.

Efficient Test-Time Model Adaptation without Forgetting

- Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, Mingkui Tan
- abstract: Test-time adaptation provides an effective means of tackling the potential distribution shift between model training and inference, by dynamically updating the model at test time. This area has seen fast progress recently, at the effectiveness of handling test shifts. Nonetheless, prior methods still suffer two key limitations: 1) these methods rely on performing backward computation for each test sample, which takes a considerable amount of time; and 2) these methods focus on improving the performance on out-of-distribution test samples and ignore that the adaptation on test data may result in a catastrophic forgetting issue, \ie, the performance on in-distribution test samples may degrade. To address these issues, we propose an efficient anti-forgetting test-time adaptation (EATA) method. Specifically, we devise a sample-efficient entropy minimization loss to exclude uninformative samples out of backward computation, which improves the overall efficiency and meanwhile boosts the out-of-distribution accuracy. Afterward, we introduce a regularization loss to ensure that critical model weights tend to be preserved during adaptation, thereby alleviating the forgetting issue. Extensive experiments on CIFAR-10-C, ImageNet-C, and ImageNet-R verify the effectiveness and superiority of our EATA.

Generative Trees: Adversarial and Copycat

- Richard Nock, Mathieu Guillame-Bert
- abstract: While Generative Adversarial Networks (GANs) achieve spectacular results on unstructured data like images, there is still a gap on tabular data, data for which state of the art supervised learning still favours decision tree (DT)-based models. This paper proposes a new path forward for the generation of tabular data, exploiting decades-old understanding of the supervised task's best components for DT induction, from losses (properness), models (tree-based) to algorithms (boosting). The properness condition on the supervised loss – which postulates the optimality of Bayes rule – leads us to a variational GAN-style loss formulation which is tight when discriminators meet a calibration property trivially satisfied by DTs, and, under common assumptions about the supervised loss, yields "one loss to train against them all" for the generator: the χ^2. We then introduce tree-based generative models, generative trees (GTs), meant to mirror on the generative side the good properties of DTs for classifying tabular data, with a boosting-compliant adversarial training algorithm for GTs. We also introduce copycat training, in which the generator copies at run time the underlying tree (graph) of the discriminator DT and completes it for the hardest discriminative task, with boosting compliant convergence. We test our algorithms on tasks including fake/real distinction and missing data imputation.

Path-Aware and Structure-Preserving Generation of Synthetically Accessible Molecules

- Juhwan Noh, Dae-Woong Jeong, Kiyoung Kim, Sehui Han, Moontae Lee, Honglak Lee, Yousung Jung
- abstract: Computational chemistry aims to autonomously design specific molecules with target functionality. Generative frameworks provide useful tools to learn continuous representations of molecules in a latent space. While modelers could optimize chemical properties, many generated molecules are not synthesizable. To design synthetically accessible molecules that preserve main structural motifs of target molecules, we propose a reaction-embedded and structure-conditioned variational autoencoder. As the latent space jointly encodes molecular structures and their reaction routes, our new sampling method that measures the path-informed structural similarity allows us to effectively generate structurally analogous synthesizable molecules. When targeting out-of-domain as well as in-domain seed structures, our model generates structurally and property-wisely similar molecules equipped with well-defined reaction paths. By focusing on the important region in chemical space, we also demonstrate that our model can design new molecules with even higher activity than the seed molecules.

Utilizing Expert Features for Contrastive Learning of Time-Series Representations

- Manuel T Nonnenmacher, Lukas Oldenburg, Ingo Steinwart, David Reeb
- abstract: We present an approach that incorporates expert knowledge for time-series representation learning. Our method employs expert features to replace the commonly used data transformations in previous contrastive learning approaches. We do this since time-series data frequently stems from the industrial or medical field where expert features are often available from domain experts, while transformations are generally elusive for time-series data. We start by proposing two properties that useful time-series representations should fulfill and show that current representation learning approaches do not ensure these properties. We therefore devise ExpCLR, a novel contrastive learning approach built on an objective that utilizes expert features to encourage both properties for the learned representation. Finally, we demonstrate on three real-world time-series datasets that ExpCLR surpasses several state-of-the-art methods for both unsupervised and semi-supervised representation learning.

Tranception: Protein Fitness Prediction with Autoregressive Transformers and Inference-time Retrieval

- Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena Hurtado, Aidan N Gomez, Debora Marks, Yarin Gal
- abstract: The ability to accurately model the fitness landscape of protein sequences is critical to a wide range of applications, from quantifying the effects of human variants on disease likelihood, to predicting immune-escape mutations in viruses and designing novel biotherapeutic proteins. Deep generative models of protein sequences trained on multiple sequence alignments have been the most successful approaches so far to address these tasks. The performance of these methods is however contingent on the availability of sufficiently deep and diverse alignments for reliable training. Their potential scope is thus limited by the fact many protein families are hard, if not impossible, to align. Large language models trained on massive quantities of non-aligned protein sequences from diverse families address these problems and show potential to eventually bridge the performance gap. We introduce Tranception, a novel transformer architecture leveraging autoregressive predictions and retrieval of homologous sequences at inference to achieve state-of-the-art fitness prediction performance. Given its markedly higher performance on multiple mutants, robustness to shallow alignments and ability to score indels, our approach offers significant gain of scope over existing approaches. To enable more rigorous model testing across a broader range of protein families, we develop ProteinGym – an extensive set of multiplexed assays of variant effects, substantially increasing both the number and diversity of assays compared to existing benchmarks.

Fast Finite Width Neural Tangent Kernel

- Roman Novak, Jascha Sohl-Dickstein, Samuel S Schoenholz
- abstract: The Neural Tangent Kernel (NTK), defined as the outer product of the neural network (NN) Jacobians, has emerged as a central object of study in deep learning. In the infinite width limit, the NTK can sometimes be computed analytically and is useful for understanding training and generalization of NN architectures. At finite widths, the NTK is also used to better initialize NNs, compare the conditioning across models, perform architecture search, and do meta-learning. Unfortunately, the finite width NTK is notoriously expensive to compute, which severely limits its practical utility. We perform the first in-depth analysis of the compute and memory requirements for NTK computation in finite width networks. Leveraging the structure of neural networks, we further propose two novel algorithms that change the exponent of the compute and memory requirements of the finite width NTK, dramatically improving efficiency. Our algorithms can be applied in a black box fashion to any differentiable function, including those implementing neural networks. We open-source our implementations within the Neural Tangents package at https://github.com/google/neural-tangents.

Multicoated Supermasks Enhance Hidden Networks

- Yasuyuki Okoshi, Ángel López García-Arias, Kazutoshi Hirose, Kota Ando, Kazushi Kawamura, Thiem Van Chu, Masato Motomura, Jaehoon Yu
- abstract: Hidden Networks (Ramanujan et al., 2020) showed the possibility of finding accurate subnetworks within a randomly weighted neural network by training a connectivity mask, referred to as supermask. We show that the supermask stops improving even though gradients are not zero, thus underutilizing backpropagated information. To address this we propose a method that extends Hidden Networks by training an overlay of multiple hierarchical supermasks{—}a multicoated supermask. This method shows that using multiple supermasks for a single task achieves higher accuracy without additional training cost. Experiments on CIFAR-10 and ImageNet show that Multicoated Supermasks enhance the tradeoff between accuracy and model size. A ResNet-101 using a 7-coated supermask outperforms its Hidden Networks counterpart by 4%, matching the accuracy of a dense ResNet-50 while being an order of magnitude smaller.

Generalized Leverage Scores: Geometric Interpretation and Applications

- Bruno Ordozgoiti, Antonis Matakos, Aristides Gionis
- abstract: In problems involving matrix computations, the concept of leverage has found a large number of applications. In particular, leverage scores, which relate the columns of a matrix to the subspaces spanned by its leading singular vectors, are helpful in revealing column subsets to approximately factorize a matrix with quality guarantees. As such, they provide a solid foundation for a variety of machine-learning methods. In this paper we extend the definition of leverage scores to relate the columns of a matrix to arbitrary subsets of singular vectors. We establish a precise connection between column and singular-vector subsets, by relating the concepts of leverage scores and principal angles between subspaces. We employ this result to design approximation algorithms with provable guarantees for two well-known problems: generalized column subset selection and sparse canonical correlation analysis. We run numerical experiments to provide further insight on the proposed methods. The novel bounds we derive improve our understanding of fundamental concepts in matrix approximations. In addition, our insights may serve as building blocks for further contributions.

Practical Almost-Linear-Time Approximation Algorithms for Hybrid and Overlapping Graph Clustering

- Lorenzo Orecchia, Konstantinos Ameranis, Charalampos Tsourakakis, Kunal Talwar
- abstract: Detecting communities in real-world networks and clustering similarity graphs are major data mining tasks with a wide range of applications in graph mining, collaborative filtering, and bioinformatics. In many such applications, overwhelming empirical evidence suggests that communities and clusters are naturally overlapping, i.e., the boundary of a cluster may contain both edges across clusters and nodes that are shared with other clusters, calling for novel hybrid graph partitioning algorithms (HGP). While almost-linear-time approximation algorithms are known for edge-boundary-based graph partitioning, little progress has been made on fast algorithms for HGP, even in the special case of vertex-boundary-based graph partitioning. In this

work, we introduce a frame-work based on two novel clustering objectives, which naturally extend the well-studied notion of conductance to clusters with hybrid vertex-and edge-boundary structure. Our main algorithmic contributions are almost-linear-time algorithms O(log n)-approximation algorithms for both these objectives. To this end, we show that the cut-matching framework of (Khandekar et al., 2014) can be significantly extended to incorporate hybrid partitions. Crucially, we implement our approximation algorithm to produce both hybrid partitions and optimality certificates for large graphs, easily scaling to tens of millions of edges, and test our implementation on real-world datasets against other competitive baselines.

Anticorrelated Noise Injection for Improved Generalization

- Antonio Orvieto, Hans Kersting, Frank Proske, Francis Bach, Aurelien Lucchi
- abstract: Injecting artificial noise into gradient descent (GD) is commonly employed to improve the performance of machine learning models. Usually, uncorrelated noise is used in such perturbed gradient descent (PGD) methods. It is, however, not known if this is optimal or whether other types of noise could provide better generalization performance. In this paper, we zoom in on the problem of correlating the perturbations of consecutive PGD steps. We consider a variety of objective functions for which we find that GD with anticorrelated perturbations ("Anti-PGD") generalizes significantly better than GD and standard (uncorrelated) PGD. To support these experimental findings, we also derive a theoretical analysis that demonstrates that Anti-PGD moves to wider minima, while GD and PGD remain stuck in suboptimal regions or even diverge. This new connection between anticorrelated noise and generalization opens the field to novel ways to exploit noise for training machine learning models.

Scalable Deep Gaussian Markov Random Fields for General Graphs

- Joel Oskarsson, Per Sidén, Fredrik Lindsten
- abstract: Machine learning methods on graphs have proven useful in many applications due to their ability to handle generally structured data. The framework of Gaussian Markov Random Fields (GMRFs) provides a principled way to define Gaussian models on graphs by utilizing their sparsity structure. We propose a flexible GMRF model for general graphs built on the multi-layer structure of Deep GMRFs, originally proposed for lattice graphs only. By designing a new type of layer we enable the model to scale to large graphs. The layer is constructed to allow for efficient training using variational inference and existing software frameworks for Graph Neural Networks. For a Gaussian likelihood, close to exact Bayesian inference is available for the latent field. This allows for making predictions with accompanying uncertainty estimates. The usefulness of the proposed model is verified by experiments on a number of synthetic and real world datasets, where it compares favorably to other both Bayesian and deep learning methods.

Zero-shot AutoML with Pretrained Models

- Ekrem Öztürk, Fabio Ferreira, Hadi Jomaa, Lars Schmidt-Thieme, Josif Grabocka, Frank Hutter
- abstract: Given a new dataset D and a low compute budget, how should we choose a pre-trained model to fine-tune to D, and set the fine-tuning hyperparameters without risking overfitting, particularly if D is small? Here, we extend automated machine learning (AutoML) to best make these choices. Our domain-independent meta-learning approach learns a zero-shot surrogate model which, at test time, allows to select the right deep learning (DL) pipeline (including the pre-trained model and fine-tuning hyperparameters) for a new dataset D given only trivial meta-features describing D such as image resolution or the number of classes. To train this zero-shot model, we collect performance data for many DL pipelines on a large collection of datasets and meta-train on this data to minimize a pairwise ranking objective. We evaluate our approach under the strict time limit of the vision track of the ChaLearn AutoDL challenge benchmark, clearly outperforming all challenge contenders.

History Compression via Language Models in Reinforcement Learning

- Fabian Paischer, Thomas Adler, Vihang Patil, Angela Bitto-Nemling, Markus Holzleitner, Sebastian Lehner, Hamid Eghbal-Zadeh, Sepp Hochreiter
- abstract: In a partially observable Markov decision process (POMDP), an agent typically uses a representation of the past to approximate the underlying MDP. We propose to utilize a frozen Pretrained Language Transformer (PLT) for history representation and compression to improve sample efficiency. To avoid training of the Transformer, we introduce FrozenHopfield, which automatically associates observations with pretrained token embeddings. To form these associations, a modern Hopfield network stores these token embeddings, which are retrieved by queries that are obtained by a random but fixed projection of observations. Our new method, HELM, enables actor-critic network architectures that contain a pretrained language Transformer for history representation as a memory module. Since a representation of the past need not be learned, HELM is much more sample efficient than competitors. On Minigrid and Procgen environments HELM achieves new state-of-the-art results. Our code is available at https://github.com/ml-jku/helm.

A Study on the Ramanujan Graph Property of Winning Lottery Tickets

- Bithika Pal, Arindam Biswas, Sudeshna Kolay, Pabitra Mitra, Biswajit Basu
- abstract: Winning lottery tickets refer to sparse subgraphs of deep neural networks which have classification accuracy close to the original dense networks. Resilient connectivity properties of such sparse networks play an important role in their performance. The attempt is to identify a sparse and yet well-connected network to guarantee unhindered information flow. Connectivity in a graph is best characterized by its spectral expansion property. Ramanujan graphs are robust expanders which lead to sparse but highly-connected networks, and thus aid in studying the winning tickets. A feedforward neural network consists of a sequence of bipartite graphs representing its layers. We analyze the Ramanujan graph property of such bipartite layers in terms of their spectral characteristics using the Cheeger’s inequality for irregular graphs. It is empirically observed that the winning ticket networks preserve the Ramanujan graph property and achieve a high accuracy even when the layers are sparse. Accuracy and robustness to noise start declining as many of the layers lose the property. Next we find a robust winning lottery ticket by pruning individual layers while retaining their respective Ramanujan graph property. This strategy is observed to improve the performance of existing network pruning algorithms.

On Learning Mixture of Linear Regressions in the Non-Realizable Setting

- Soumyabrata Pal, Arya Mazumdar, Rajat Sen, Avishek Ghosh
- abstract: While mixture of linear regressions (MLR) is a well-studied topic, prior works usually do not analyze such models for prediction error. In fact, prediction and loss are not well-defined in the context of mixtures. In this paper, first we show that MLR can be used for prediction where instead of predicting a label, the model predicts a list of values (also known as list-decoding). The list size is equal to the number of components in the mixture, and the loss function is defined to be minimum among the losses resulted by all the component models. We show that with this definition, a solution of the empirical risk minimization (ERM) achieves small probability of prediction error. This begs for an algorithm to minimize the empirical risk for MLR, which is known to be computationally hard. Prior algorithmic works in MLR focus on the realizable setting, i.e., recovery of parameters when data is probabilistically generated by a mixed linear (noisy) model. In this paper we show that a version of the popular expectation minimization (EM) algorithm finds out the best fit lines in a dataset even when a realizable model is not assumed, under some regularity conditions on the dataset and the initial points, and thereby provides a solution for the ERM. We further provide an algorithm that runs in polynomial time in the number of datapoints, and recovers a good approximation of the best fit lines. The two algorithms are experimentally compared.

Plan Better Amid Conservatism: Offline Multi-Agent Reinforcement Learning with Actor Rectification

- Ling Pan, Longbo Huang, Tengyu Ma, Huazhe Xu

- abstract: Conservatism has led to significant progress in offline reinforcement learning (RL) where an agent learns from pre-collected datasets. However, as many real-world scenarios involve interaction among multiple agents, it is important to resolve offline RL in the multi-agent setting. Given the recent success of transferring online RL algorithms to the multi-agent setting, one may expect that offline RL algorithms will also transfer to multi-agent settings directly. Surprisingly, we empirically observe that conservative offline RL algorithms do not work well in the multi-agent setting—the performance degrades significantly with an increasing number of agents. Towards mitigating the degradation, we identify a key issue that non-concavity of the value function makes the policy gradient improvements prone to local optima. Multiple agents exacerbate the problem severely, since the suboptimal policy by any agent can lead to uncoordinated global failure. Following this intuition, we propose a simple yet effective method, Offline Multi-Agent RL with Actor Rectification (OMAR), which combines the first-order policy gradients and zeroth-order optimization methods to better optimize the conservative value functions over the actor parameters. Despite the simplicity, OMAR achieves state-of-the-art results in a variety of multi-agent control tasks.

A Unified Weight Initialization Paradigm for Tensorial Convolutional Neural Networks

- Yu Pan, Zeyong Su, Ao Liu, Wang Jingquan, Nannan Li, Zenglin Xu
- abstract: Tensorial Convolutional Neural Networks (TCNNs) have attracted much research attention for their power in reducing model parameters or enhancing the generalization ability. However, exploration of TCNNs is hindered even from weight initialization methods. To be specific, general initialization methods, such as Xavier or Kaiming initialization, usually fail to generate appropriate weights for TCNNs. Meanwhile, although there are ad-hoc approaches for specific architectures (e.g., Tensor Ring Nets), they are not applicable to TCNNs with other tensor decomposition methods (e.g., CP or Tucker decomposition). To address this problem, we propose a universal weight initialization paradigm, which generalizes Xavier and Kaiming methods and can be widely applicable to arbitrary TCNNs. Specifically, we first present the Reproducing Transformation to convert the backward process in TCNNs to an equivalent convolution process. Then, based on the convolution operators in the forward and backward processes, we build a unified paradigm to control the variance of features and gradients in TCNNs. Thus, we can derive fan-in and fan-out initialization for various TCNNs. We demonstrate that our paradigm can stabilize the training of TCNNs, leading to faster convergence and better results.

Robustness and Accuracy Could Be Reconcilable by (Proper) Definition

- Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, Shuicheng Yan
- abstract: The trade-off between robustness and accuracy has been widely studied in the adversarial literature. Although still controversial, the prevailing view is that this trade-off is inherent, either empirically or theoretically. Thus, we dig for the origin of this trade-off in adversarial training and find that it may stem from the improperly defined robust error, which imposes an inductive bias of local invariance — an overcorrection towards smoothness. Given this, we advocate employing local equivariance to describe the ideal behavior of a robust model, leading to a self-consistent robust error named SCORE. By definition, SCORE facilitates the reconciliation between robustness and accuracy, while still handling the worst-case uncertainty via robust optimization. By simply substituting KL divergence with variants of distance metrics, SCORE can be efficiently minimized. Empirically, our models achieve top-rank performance on RobustBench under AutoAttack. Besides, SCORE provides instructive insights for explaining the overfitting phenomenon and semantic input gradients observed on robust models.

Towards Coherent and Consistent Use of Entities in Narrative Generation

- Pinelopi Papalampidi, Kris Cao, Tomas Kocisky
- abstract: Large pre-trained language models (LMs) have demonstrated impressive capabilities in generating long, fluent text; however, there is little to no analysis on their ability to maintain entity coherence and consistency. In this work, we focus on the end task of narrative generation and systematically analyse the long-range entity coherence and consistency in generated stories. First, we propose a set of automatic metrics for measuring model performance in terms of entity usage. Given these metrics, we quantify the limitations of current LMs. Next, we propose augmenting a pre-trained LM with a dynamic entity memory in an end-to-end manner by using an auxiliary entity-related loss for guiding the reads and writes to the memory. We demonstrate that the dynamic entity memory increases entity coherence according to both automatic and human judgment and helps preserving entity-related information especially in settings with a limited context window. Finally, we also validate that our automatic metrics are correlated with human ratings and serve as a good indicator of the quality of generated stories.

Constrained Discrete Black-Box Optimization using Mixed-Integer Programming

- Theodore P Papalexopoulos, Christian Tjandraatmadja, Ross Anderson, Juan Pablo Vielma, David Belanger
- abstract: Discrete black-box optimization problems are challenging for model-based optimization (MBO) algorithms, such as Bayesian optimization, due to the size of the search space and the need to satisfy combinatorial constraints. In particular, these methods require repeatedly solving a complex discrete global optimization problem in the inner loop, where popular heuristic inner-loop solvers introduce approximations and are difficult to adapt to combinatorial constraints. In response, we propose NN+MILP, a general discrete MBO framework using piecewise-linear neural networks as surrogate models and mixed-integer linear programming (MILP) to optimize the acquisition function. MILP provides optimality guarantees and a versatile declarative language for domain-specific constraints. We test our approach on a range of unconstrained and constrained problems, including DNA binding, constrained binary quadratic problems from the MINLPLib benchmark, and the NAS-Bench-101 neural architecture search benchmark. NN+MILP surpasses or matches the performance of black-box algorithms tailored to the constraints at hand, with global optimization of the acquisition problem running in a few minutes using only standard software packages and hardware.

A Theoretical Comparison of Graph Neural Network Extensions

- Pál András Papp, Roger Wattenhofer
- abstract: We study and compare different Graph Neural Network extensions that increase the expressive power of GNNs beyond the Weisfeiler-Leman test. We focus on (i) GNNs based on higher order WL methods, (ii) GNNs that preprocess small substructures in the graph, (iii) GNNs that preprocess the graph up to a small radius, and (iv) GNNs that slightly perturb the graph to compute an embedding. We begin by presenting a simple improvement for this last extension that strictly increases the expressive power of this GNN variant. Then, as our main result, we compare the expressiveness of these extensions to each other through a series of example constructions that can be distinguished by one of the extensions, but not by another one. We also show negative examples that are particularly challenging for each of the extensions, and we prove several claims about the ability of these extensions to count cliques and cycles in the graph.

Validating Causal Inference Methods

- Harsh Parikh, Carlos Varjao, Louise Xu, Eric Tchetgen Tchetgen
- abstract: The fundamental challenge of drawing causal inference is that counterfactual outcomes are not fully observed for any unit. Furthermore, in observational studies, treatment assignment is likely to be confounded. Many statistical methods have emerged for causal inference under unconfoundedness conditions given pre-treatment covariates, including propensity score-based methods, prognostic score-based methods, and doubly robust methods. Unfortunately for applied researchers, there is no ‘one-size-fits-all’ causal method that can perform optimally universally. In practice, causal methods are primarily evaluated quantitatively on handcrafted simulated data. Such data-generative procedures can be of limited value because they are typically stylized models of reality. They are simplified for tractability and lack the complexities of real-world data. For applied researchers, it is critical to understand how well a method performs for the data at hand. Our work introduces a deep generative model-based framework, Credence, to validate causal inference methods. The framework’s novelty stems from its ability to generate synthetic data anchored at the empirical distribution for the

observed sample, and therefore virtually indistinguishable from the latter. The approach allows the user to specify ground truth for the form and magnitude of causal effects and confounding bias as functions of covariates. Thus simulated data sets are used to evaluate the potential performance of various causal estimation methods when applied to data similar to the observed sample. We demonstrate Credence's ability to accurately assess the relative performance of causal estimation techniques in an extensive simulation study and two real-world data applications from Lalonde and Project STAR studies.

The Unsurprising Effectiveness of Pre-Trained Vision Models for Control

- Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, Abhinav Gupta
- abstract: Recent years have seen the emergence of pre-trained representations as a powerful abstraction for AI applications in computer vision, natural language, and speech. However, policy learning for control is still dominated by a tabula-rasa learning paradigm, with visuo-motor policies often trained from scratch using data from deployment environments. In this context, we revisit and study the role of pre-trained visual representations for control, and in particular representations trained on large-scale computer vision datasets. Through extensive empirical evaluation in diverse control domains (Habitat, DeepMind Control, Adroit, Franka Kitchen), we isolate and study the importance of different representation training methods, data augmentations, and feature hierarchies. Overall, we find that pre-trained visual representations can be competitive or even better than ground-truth state representations to train control policies. This is in spite of using only out-of-domain data from standard vision datasets, without any in-domain data from the deployment environments.

Learning Symmetric Embeddings for Equivariant World Models

- Jung Yeon Park, Ondrej Biza, Linfeng Zhao, Jan-Willem Van De Meent, Robin Walters
- abstract: Incorporating symmetries can lead to highly data-efficient and generalizable models by defining equivalence classes of data samples related by transformations. However, characterizing how transformations act on input data is often difficult, limiting the applicability of equivariant models. We propose learning symmetric embedding networks (SENs) that encode an input space (e.g. images), where we do not know the effect of transformations (e.g. rotations), to a feature space that transforms in a known manner under these operations. This network can be trained end-to-end with an equivariant task network to learn an explicitly symmetric representation. We validate this approach in the context of equivariant transition models with 3 distinct forms of symmetry. Our experiments demonstrate that SENs facilitate the application of equivariant networks to data with complex symmetry representations. Moreover, doing so can yield improvements in accuracy and generalization relative to both fully-equivariant and non-equivariant baselines.

Blurs Behave Like Ensembles: Spatial Smoothings to Improve Accuracy, Uncertainty, and Robustness

- Namuk Park, Songkuk Kim
- abstract: Neural network ensembles, such as Bayesian neural networks (BNNs), have shown success in the areas of uncertainty estimation and robustness. However, a crucial challenge prohibits their use in practice. BNNs require a large number of predictions to produce reliable results, leading to a significant increase in computational cost. To alleviate this issue, we propose spatial smoothing, a method that ensembles neighboring feature map points of convolutional neural networks. By simply adding a few blur layers to the models, we empirically show that spatial smoothing improves accuracy, uncertainty estimation, and robustness of BNNs across a whole range of ensemble sizes. In particular, BNNs incorporating spatial smoothing achieve high predictive performance merely with a handful of ensembles. Moreover, this method also can be applied to canonical deterministic neural networks to improve the performances. A number of evidences suggest that the improvements can be attributed to the stabilized feature maps and the smoothing of the loss landscape. In addition, we provide a fundamental explanation for prior works {—} namely, global average pooling, pre-activation, and ReLU6 {—} by addressing them as special cases of spatial smoothing. These not only enhance accuracy, but also improve uncertainty estimation and robustness by making the loss landscape smoother in the same manner as spatial smoothing. The code is available at https://github.com/xxxnell/spatial-smoothing.

Exact Optimal Accelerated Complexity for Fixed-Point Iterations

- Jisun Park, Ernest K Ryu
- abstract: Despite the broad use of fixed-point iterations throughout applied mathematics, the optimal convergence rate of general fixed-point problems with nonexpansive nonlinear operators has not been established. This work presents an acceleration mechanism for fixed-point iterations with nonexpansive operators, contractive operators, and nonexpansive operators satisfying a Hölder-type growth condition. We then provide matching complexity lower bounds to establish the exact optimality of the acceleration mechanisms in the nonexpansive and contractive setups. Finally, we provide experiments with CT imaging, optimal transport, and decentralized optimization to demonstrate the practical effectiveness of the acceleration mechanism.

Kernel Methods for Radial Transformed Compositional Data with Many Zeros

- Junyoung Park, Changwon Yoon, Cheolwoo Park, Jeongyoun Ahn
- abstract: Compositional data analysis with a high proportion of zeros has gained increasing popularity, especially in chemometrics and human gut microbiomes research. Statistical analyses of this type of data are typically carried out via a log-ratio transformation after replacing zeros with small positive values. We should note, however, that this procedure is geometrically improper, as it causes anomalous distortions through the transformation. We propose a radial transformation that does not require zero substitutions and more importantly results in essential equivalence between domains before and after the transformation. We show that a rich class of kernels on hyperspheres can successfully define a kernel embedding for compositional data based on this equivalence. To the best of our knowledge, this is the first work that theoretically establishes the availability of the extensive library of kernel-based machine learning methods for compositional data. The applicability of the proposed approach is demonstrated with kernel principal component analysis.

Evolving Curricula with Regret-Based Environment Design

- Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, Tim Rocktäschel
- abstract: Training generally-capable agents with reinforcement learning (RL) remains a significant challenge. A promising avenue for improving the robustness of RL agents is through the use of curricula. One such class of methods frames environment design as a game between a student and a teacher, using regret-based objectives to produce environment instantiations (or levels) at the frontier of the student agent's capabilities. These methods benefit from theoretical robustness guarantees at equilibrium, yet they often struggle to find effective levels in challenging design spaces in practice. By contrast, evolutionary approaches incrementally alter environment complexity, resulting in potentially open-ended learning, but often rely on domain-specific heuristics and vast amounts of computational resources. This work proposes harnessing the power of evolution in a principled, regret-based curriculum. Our approach, which we call Adversarially Compounding Complexity by Editing Levels (ACCEL), seeks to constantly produce levels at the frontier of an agent's capabilities, resulting in curricula that start simple but become increasingly complex. ACCEL maintains the theoretical benefits of prior regret-based methods, while providing significant empirical gains in a diverse set of environments. An interactive version of this paper is available at https://accelagent.github.io.

Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps

- Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, Christophe Pallier

- abstract: Neural Language Models (NLMs) have made tremendous advances during the last years, achieving impressive performance on various linguistic tasks. Capitalizing on this, studies in neuroscience have started to use NLMs to study neural activity in the human brain during language processing. However, many questions remain unanswered regarding which factors determine the ability of a neural language model to capture brain activity (aka its 'brain score'). Here, we make first steps in this direction and examine the impact of test loss, training corpus and model architecture (comparing GloVe, LSTM, GPT-2 and BERT), on the prediction of functional Magnetic Resonance Imaging time-courses of participants listening to an audiobook. We find that (1) untrained versions of each model already explain significant amount of signal in the brain by capturing similarity in brain responses across identical words, with the untrained LSTM outperforming the transformer-based models, being less impacted by the effect of context; (2) that training NLP models improves brain scores in the same brain regions irrespective of the model's architecture; (3) that Perplexity (test loss) is not a good predictor of brain score; (4) that training data have a strong influence on the outcome and, notably, that off-the-shelf models may lack statistical power to detect brain activations. Overall, we outline the impact of model-training choices, and suggest good practices for future studies aiming at explaining the human language system using neural language models.

A new similarity measure for covariate shift with applications to nonparametric regression

- Reese Pathak, Cong Ma, Martin Wainwright
- abstract: We study covariate shift in the context of nonparametric regression. We introduce a new measure of distribution mismatch between the source and target distributions using the integrated ratio of probabilities of balls at a given radius. We use the scaling of this measure with respect to the radius to characterize the minimax rate of estimation over a family of H{ö}lder continuous functions under covariate shift. In comparison to the recently proposed notion of transfer exponent, this measure leads to a sharper rate of convergence and is more fine-grained. We accompany our theory with concrete instances of covariate shift that illustrate this sharp difference.

Align-RUDDER: Learning From Few Demonstrations by Reward Redistribution

- Vihang Patil, Markus Hofmarcher, Marius-Constantin Dinu, Matthias Dorfer, Patrick M Blies, Johannes Brandstetter, José Arjona-Medina, Sepp Hochreiter
- abstract: Reinforcement learning algorithms require many samples when solving complex hierarchical tasks with sparse and delayed rewards. For such complex tasks, the recently proposed RUDDER uses reward redistribution to leverage steps in the Q-function that are associated with accomplishing sub-tasks. However, often only few episodes with high rewards are available as demonstrations since current exploration strategies cannot discover them in reasonable time. In this work, we introduce Align-RUDDER, which utilizes a profile model for reward redistribution that is obtained from multiple sequence alignment of demonstrations. Consequently, Align-RUDDER employs reward redistribution effectively and, thereby, drastically improves learning on few demonstrations. Align-RUDDER outperforms competitors on complex artificial tasks with delayed rewards and few demonstrations. On the Minecraft ObtainDiamond task, Align-RUDDER is able to mine a diamond, though not frequently. Code is available at github.com/ml-jku/align-rudder.

POET: Training Neural Networks on Tiny Devices with Integrated Rematerialization and Paging

- Shishir G. Patil, Paras Jain, Prabal Dutta, Ion Stoica, Joseph Gonzalez
- abstract: Fine-tuning models on edge devices like mobile phones would enable privacy-preserving personalization over sensitive data. However, edge training has historically been limited to relatively small models with simple architectures because training is both memory and energy intensive. We present POET, an algorithm to enable training large neural networks on memory-scarce battery-operated edge devices. POET jointly optimizes the integrated search search spaces of rematerialization and paging, two algorithms to reduce the memory consumption of backpropagation. Given a memory budget and a run-time constraint, we formulate a mixed-integer linear program (MILP) for energy-optimal training. Our approach enables training significantly larger models on embedded devices while reducing energy consumption while not modifying mathematical correctness of backpropagation. We demonstrate that it is possible to fine-tune both ResNet-18 and BERT within the memory constraints of a Cortex-M class embedded device while outperforming current edge training methods in energy efficiency. POET is an open-source project available at https://github.com/ShishirPatil/poet

Learning to Cut by Looking Ahead: Cutting Plane Selection via Imitation Learning

- Max B Paulus, Giulia Zarpellon, Andreas Krause, Laurent Charlin, Chris Maddison
- abstract: Cutting planes are essential for solving mixed-integer linear problems (MILPs), because they facilitate bound improvements on the optimal solution value. For selecting cuts, modern solvers rely on manually designed heuristics that are tuned to gauge the potential effectiveness of cuts. We show that a greedy selection rule explicitly looking ahead to select cuts that yield the best bound improvement delivers strong decisions for cut selection – but is too expensive to be deployed in practice. In response, we propose a new neural architecture (NeuralCut) for imitation learning on the lookahead expert. Our model outperforms standard baselines for cut selection on several synthetic MILP benchmarks. Experiments on a realistic B&C solver further validate our approach, and exhibit the potential of learning methods in this setting.

Neural Network Pruning Denoises the Features and Makes Local Connectivity Emerge in Visual Tasks

- Franco Pellegrini, Giulio Biroli
- abstract: Pruning methods can considerably reduce the size of artificial neural networks without harming their performance and in some cases they can even uncover sub-networks that, when trained in isolation, match or surpass the test accuracy of their dense counterparts. Here, we characterize the inductive bias that pruning imprints in such "winning lottery tickets": focusing on visual tasks, we analyze the architecture resulting from iterative magnitude pruning of a simple fully connected network. We show that the surviving node connectivity is local in input space, and organized in patterns reminiscent of the ones found in convolutional networks. We investigate the role played by data and tasks in shaping the architecture of the pruned sub-network. We find that pruning performances, and the ability to sift out the noise and make local features emerge, improve by increasing the size of the training set, and the semantic value of the data. We also study different pruning procedures, and find that iterative magnitude pruning is particularly effective in distilling meaningful connectivity out of features present in the original task. Our results suggest the possibility to automatically discover new and efficient architectural inductive biases in other datasets and tasks.

Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding

- Yifan Peng, Siddharth Dalmia, Ian Lane, Shinji Watanabe
- abstract: Conformer has proven to be effective in many speech processing tasks. It combines the benefits of extracting local dependencies using convolutions and global dependencies using self-attention. Inspired by this, we propose a more flexible, interpretable and customizable encoder alternative, Branchformer, with parallel branches for modeling various ranged dependencies in end-to-end speech processing. In each encoder layer, one branch employs self-attention or its variant to capture long-range dependencies, while the other branch utilizes an MLP module with convolutional gating (cgMLP) to extract local relationships. We conduct experiments on several speech recognition and spoken language understanding benchmarks. Results show that our model outperforms both Transformer and cgMLP. It also matches with or outperforms state-of-the-art results achieved by Conformer. Furthermore, we show various strategies to reduce computation thanks to the two-branch architecture, including the ability to have variable inference complexity in a single trained model. The weights learned for merging branches indicate how local and global dependencies are utilized in different layers, which benefits model designing.

Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets

- Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, Jianzhu Ma
- abstract: Deep generative models have achieved tremendous success in designing novel drug molecules in recent years. A new thread of works have shown potential in advancing the specificity and success rate of in silico drug design by considering the structure of protein pockets. This setting posts fundamental computational challenges in sampling new chemical compounds that could satisfy multiple geometrical constraints imposed by pockets. Previous sampling algorithms either sample in the graph space or only consider the 3D coordinates of atoms while ignoring other detailed chemical structures such as bond types and functional groups. To address the challenge, we develop an E(3)-equivariant generative network composed of two modules: 1) a new graph neural network capturing both spatial and bonding relationships between atoms of the binding pockets and 2) a new efficient algorithm which samples new drug candidates conditioned on the pocket representations from a tractable distribution without relying on MCMC. Experimental results demonstrate that molecules sampled from Pocket2Mol achieve significantly better binding affinity and other drug properties such as drug-likeness and synthetic accessibility.

Differentiable Top-k Classification Learning

- Felix Petersen, Hilde Kuehne, Christian Borgelt, Oliver Deussen
- abstract: The top-k classification accuracy is one of the core metrics in machine learning. Here, k is conventionally a positive integer, such as 1 or 5, leading to top-1 or top-5 training objectives. In this work, we relax this assumption and optimize the model for multiple k simultaneously instead of using a single k. Leveraging recent advances in differentiable sorting and ranking, we propose a family of differentiable top-k cross-entropy classification losses. This allows training while not only considering the top-1 prediction, but also, e.g., the top-2 and top-5 predictions. We evaluate the proposed losses for fine-tuning on state-of-the-art architectures, as well as for training from scratch. We find that relaxing k not only produces better top-5 accuracies, but also leads to top-1 accuracy improvements. When fine-tuning publicly available ImageNet models, we achieve a new state-of-the-art for these models.

Multi-scale Feature Learning Dynamics: Insights for Double Descent

- Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, Guillaume Lajoie
- abstract: An intriguing phenomenon that arises from the high-dimensional learning dynamics of neural networks is the phenomenon of “double descent”. The more commonly studied aspect of this phenomenon corresponds to model-wise double descent where the test error exhibits a second descent with increasing model complexity, beyond the classical U-shaped error curve. In this work, we investigate the origins of the less studied epoch-wise double descent in which the test error undergoes two non-monotonous transitions, or descents as the training time increases. We study a linear teacher-student setup exhibiting epoch-wise double descent similar to that in deep neural networks. In this setting, we derive closed-form analytical expressions describing the generalization error in terms of low-dimensional scalar macroscopic variables. We find that double descent can be attributed to distinct features being learned at different scales: as fast-learning features overfit, slower-learning features start to fit, resulting in a second descent in test error. We validate our findings through numerical simulations where our theory accurately predicts empirical findings and remains consistent with observations in deep neural networks.

A Differential Entropy Estimator for Training Neural Networks

- Georg Pichler, Pierre Jean A. Colombo, Malik Boudiaf, Günther Koliander, Pablo Piantanida
- abstract: Mutual Information (MI) has been widely used as a loss regularizer for training neural networks. This has been particularly effective when learn disentangled or compressed representations of high dimensional data. However, differential entropy (DE), another fundamental measure of information, has not found widespread use in neural network training. Although DE offers a potentially wider range of applications than MI, off-the-shelf DE estimators are either non differentiable, computationally intractable or fail to adapt to changes in the underlying distribution. These drawbacks prevent them from being used as regularizers in neural networks training. To address shortcomings in previously proposed estimators for DE, here we introduce KNIFE, a fully parameterized, differentiable kernel-based estimator of DE. The flexibility of our approach also allows us to construct KNIFE-based estimators for conditional (on either discrete or continuous variables) DE, as well as MI. We empirically validate our method on high-dimensional synthetic data and further apply it to guide the training of neural networks for real-world tasks. Our experiments on a large variety of tasks, including visual domain adaptation, textual fair classification, and textual fine-tuning demonstrate the effectiveness of KNIFE-based estimation. Code can be found at https://github.com/g-pichler/knife.

Federated Learning with Partial Model Personalization

- Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, Lin Xiao
- abstract: We consider two federated learning algorithms for training partially personalized models, where the shared and personal parameters are updated either simultaneously or alternately on the devices. Both algorithms have been proposed in the literature, but their convergence properties are not fully understood, especially for the alternating variant. We provide convergence analyses of both algorithms in the general nonconvex setting with partial participation and delineate the regime where one dominates the other. Our experiments on real-world image, text, and speech datasets demonstrate that (a) partial personalization can obtain most of the benefits of full model personalization with a small fraction of personal parameters, and, (b) the alternating update algorithm outperforms the simultaneous update algorithm by a small but consistent margin.

Deep Networks on Toroids: Removing Symmetries Reveals the Structure of Flat Regions in the Landscape Geometry

- Fabrizio Pittorino, Antonio Ferraro, Gabriele Perugini, Christoph Feinauer, Carlo Baldassi, Riccardo Zecchina
- abstract: We systematize the approach to the investigation of deep neural network landscapes by basing it on the geometry of the space of implemented functions rather than the space of parameters. Grouping classifiers into equivalence classes, we develop a standardized parameterization in which all symmetries are removed, resulting in a toroidal topology. On this space, we explore the error landscape rather than the loss. This lets us derive a meaningful notion of the flatness of minimizers and of the geodesic paths connecting them. Using different optimization algorithms that sample minimizers with different flatness we study the mode connectivity and relative distances. Testing a variety of state-of-the-art architectures and benchmark datasets, we confirm the correlation between flatness and generalization performance; we further show that in function space flatter minima are closer to each other and that the barriers along the geodesics connecting them are small. We also find that minimizers found by variants of gradient descent can be connected by zero-error paths composed of two straight lines in parameter space, i.e. polygonal chains with a single bend. We observe similar qualitative results in neural networks with binary weights and activations, providing one of the first results concerning the connectivity in this setting. Our results hinge on symmetry removal, and are in remarkable agreement with the rich phenomenology described by some recent analytical studies performed on simple shallow models.

Geometric Multimodal Contrastive Representation Learning

- Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S. Melo, Ana Paiva, Danica Kragic
- abstract: Learning representations of multimodal data that are both informative and robust to missing modalities at test time remains a challenging problem due to the inherent heterogeneity of data obtained from different channels. To address it, we present a novel Geometric Multimodal Contrastive (GMC) representation learning method consisting of two main components: i) a two-level architecture consisting of modality-specific base encoders,

allowing to process an arbitrary number of modalities to an intermediate representation of fixed dimensionality, and a shared projection head, mapping the intermediate representations to a latent representation space; ii) a multimodal contrastive loss function that encourages the geometric alignment of the learned representations. We experimentally demonstrate that GMC representations are semantically rich and achieve state-of-the-art performance with missing modality information on three different learning problems including prediction and reinforcement learning tasks.

Constrained Offline Policy Optimization

- Nicholas Polosky, Bruno C. Da Silva, Madalina Fiterau, Jithin Jagannath
- abstract: In this work we introduce Constrained Offline Policy Optimization (COPO), an offline policy optimization algorithm for learning in MDPs with cost constraints. COPO is built upon a novel offline cost-projection method, which we formally derive and analyze. Our method improves upon the state-of-the-art in offline constrained policy optimization by explicitly accounting for distributional shift and by offering non-asymptotic confidence bounds on the cost of a policy. These formal properties are superior to those of existing techniques, which only guarantee convergence to a point estimate. We formally analyze our method and empirically demonstrate that it achieves state-of-the-art performance on discrete and continuous control problems, while offering the aforementioned improved, stronger, and more robust theoretical guarantees.

Offline Meta-Reinforcement Learning with Online Self-Supervision

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

Debiaser Beware: Pitfalls of Centering Regularized Transport Maps

- Aram-Alexandre Pooladian, Marco Cuturi, Jonathan Niles-Weed
- abstract: Estimating optimal transport (OT) maps (a.k.a. Monge maps) between two measures P and Q is a problem fraught with computational and statistical challenges. A promising approach lies in using the dual potential functions obtained when solving an entropy-regularized OT problem between samples P_n and Q_n, which can be used to recover an approximately optimal map. The negentropy penalization in that scheme introduces, however, an estimation bias that grows with the regularization strength. A well-known remedy to debias such estimates, which has gained wide popularity among practitioners of regularized OT, is to center them, by subtracting auxiliary problems involving P_n and itself, as well as Q_n and itself. We do prove that, under favorable conditions on P and Q, debiasing can yield better approximations to the Monge map. However, and perhaps surprisingly, we present a few cases in which debiasing is provably detrimental in a statistical sense, notably when the regularization strength is large or the number of samples is small. These claims are validated experimentally on synthetic and real datasets, and should reopen the debate on whether debiasing is needed when using entropic OT.

Adaptive Second Order Coresets for Data-efficient Machine Learning

- Omead Pooladzandi, David Davini, Baharan Mirzasoleiman
- abstract: Training machine learning models on massive datasets incurs substantial computational costs. To alleviate such costs, there has been a sustained effort to develop data-efficient training methods that can carefully select subsets of the training examples that generalize on par with the full training data. However, existing methods are limited in providing theoretical guarantees for the quality of the models trained on the extracted subsets, and may perform poorly in practice. We propose AdaCore, a method that leverages the geometry of the data to extract subsets of the training examples for efficient machine learning. The key idea behind our method is to dynamically approximate the curvature of the loss function via an exponentially-averaged estimate of the Hessian to select weighted subsets (coresets) that provide a close approximation of the full gradient preconditioned with the Hessian. We prove rigorous guarantees for the convergence of various first and second-order methods applied to the subsets chosen by AdaCore. Our extensive experiments show that AdaCore extracts coresets with higher quality compared to baselines and speeds up training of convex and non-convex machine learning models, such as logistic regression and neural networks, by over 2.9x over the full data and 4.5x over random subsets.

On the Practicality of Deterministic Epistemic Uncertainty

- Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, Federico Tombari
- abstract: A set of novel approaches for estimating epistemic uncertainty in deep neural networks with a single forward pass has recently emerged as a valid alternative to Bayesian Neural Networks. On the premise of informative representations, these deterministic uncertainty methods (DUMs) achieve strong performance on detecting out-of-distribution (OOD) data while adding negligible computational costs at inference time. However, it remains unclear whether DUMs are well calibrated and can seamlessly scale to real-world applications - both prerequisites for their practical deployment. To this end, we first provide a taxonomy of DUMs, and evaluate their calibration under continuous distributional shifts. Then, we extend them to semantic segmentation. We find that, while DUMs scale to realistic vision tasks and perform well on OOD detection, the practicality of current methods is undermined by poor calibration under distributional shifts.

A Simple Guard for Learned Optimizers

- Isabeau Prémont-Schwarz, Jaroslav Vítků, Jan Feyereisl
- abstract: If the trend of learned components eventually outperforming their hand-crafted version continues, learned optimizers will eventually outperform hand-crafted optimizers like SGD or Adam. Even if learned optimizers (L2Os) eventually outpace hand-crafted ones in practice however, they are still not provably convergent and might fail out of distribution. These are the questions addressed here. Currently, learned optimizers frequently outperform generic hand-crafted optimizers (such as gradient descent) at the beginning of learning but they generally plateau after some time while the generic algorithms continue to make progress and often overtake the learned algorithm as Aesop's tortoise which overtakes the hare. L2Os also still have a difficult time generalizing out of distribution. \cite{heaton_safeguarded_2020} proposed Safeguarded L2O (GL2O) which can take a learned optimizer and safeguard it with a generic learning algorithm so that by conditionally switching between the two, the resulting algorithm is provably convergent. We propose a new class of Safeguarded L2O, called Loss-Guarded L2O (LGL2O), which is both conceptually simpler and computationally less expensive. The guarding mechanism decides solely based on the expected future loss value of both optimizers. Furthermore, we show theoretical proof of LGL2O's convergence guarantee and empirical results comparing to GL2O and other baselines showing that it combines the best of both L2O and SGD and that in practice converges much better than GL2O.

Hardness and Algorithms for Robust and Sparse Optimization

- Eric Price, Sandeep Silwal, Samson Zhou
- abstract: We explore algorithms and limitations for sparse optimization problems such as sparse linear regression and robust linear regression. The goal of the sparse linear regression problem is to identify a small number of key features, while the goal of the robust linear regression problem is to identify a small number of erroneous measurements. Specifically, the sparse linear regression problem seeks a k-sparse vector $x\in\mathbb{R}^d$ to minimize $\|Ax-b\|_2$, given an input matrix $A\in\mathbb{R}^{n\times d}$ and a target vector $b\in\mathbb{R}^n$, while the robust linear regression problem seeks a set S that ignores at most k rows and a vector x to minimize $\|(Ax-b)_S\|_2$. We first show bicriteria, NP-hardness of approximation for robust regression building on the work of \cite{ODonnellWZ15} which implies a similar result for sparse regression. We further show fine-grained hardness of robust regression through a reduction from the minimum-weight k-clique conjecture. On the positive side, we give an algorithm for robust regression that achieves arbitrarily accurate additive error and uses runtime that closely matches the lower bound from the fine-grained hardness result, as well as an algorithm for sparse regression with similar runtime. Both our upper and lower bounds rely on a general reduction from robust linear regression to sparse regression that we introduce. Our algorithms, inspired by the 3SUM problem, use approximate nearest neighbor data structures and may be of independent interest for solving sparse optimization problems. For instance, we demonstrate that our techniques can also be used for the well-studied sparse PCA problem.

Nonlinear Feature Diffusion on Hypergraphs

- Konstantin Prokopchik, Austin R Benson, Francesco Tudisco
- abstract: Hypergraphs are a common model for multiway relationships in data, and hypergraph semi-supervised learning is the problem of assigning labels to all nodes in a hypergraph, given labels on just a few nodes. Diffusions and label spreading are classical techniques for semi-supervised learning in the graph setting, and there are some standard ways to extend them to hypergraphs. However, these methods are linear models, and do not offer an obvious way of incorporating node features for making predictions. Here, we develop a nonlinear diffusion process on hypergraphs that spreads both features and labels following the hypergraph structure. Even though the process is nonlinear, we show global convergence to a unique limiting point for a broad class of nonlinearities and we show that such limit is the global minimum of a new regularized semi-supervised learning loss function which aims at reducing a generalized form of variance of the nodes across the hyperedges. The limiting point serves as a node embedding from which we make predictions with a linear model. Our approach is competitive with state-of-the-art graph and hypergraph neural networks, and also takes less time to train.

Universal Joint Approximation of Manifolds and Densities by Simple Injective Flows

- Michael Puthawala, Matti Lassas, Ivan Dokmanic, Maarten De Hoop
- abstract: We study approximation of probability measures supported on n-dimensional manifolds embedded in R^m by injective flows—neural networks composed of invertible flows and injective layers. We show that in general, injective flows between R^n and R^m universally approximate measures supported on images of extendable embeddings, which are a subset of standard embeddings: when the embedding dimension m is small, topological obstructions may preclude certain manifolds as admissible targets. When the embedding dimension is sufficiently large, m >= 3n+1, we use an argument from algebraic topology known as the clean trick to prove that the topological obstructions vanish and injective flows universally approximate any differentiable embedding. Along the way we show that the studied injective flows admit efficient projections on the range, and that their optimality can be established "in reverse," resolving a conjecture made in Brehmer & Cranmer 2020.

The Teaching Dimension of Regularized Kernel Learners

- Hong Qian, Xu-Hui Liu, Chen-Xi Su, Aimin Zhou, Yang Yu
- abstract: Teaching dimension (TD) is a fundamental theoretical property for understanding machine teaching algorithms. It measures the sample complexity of teaching a target hypothesis to a learner. The TD of linear learners has been studied extensively, whereas the results of teaching non-linear learners are rare. A recent result investigates the TD of polynomial and Gaussian kernel learners. Unfortunately, the theoretical bounds therein show that the TD is high when teaching those non-linear learners. Inspired by the fact that regularization can reduce the learning complexity in machine learning, a natural question is whether the similar fact happens in machine teaching. To answer this essential question, this paper proposes a unified theoretical framework termed STARKE to analyze the TD of regularized kernel learners. On the basis of STARKE, we derive a generic result of any type of kernels. Furthermore, we disclose that the TD of regularized linear and regularized polynomial kernel learners can be strictly reduced. For regularized Gaussian kernel learners, we reveal that, although their TD is infinite, their epsilon-approximate TD can be exponentially reduced compared with that of the unregularized learners. The extensive experimental results of teaching the optimization-based learners verify the theoretical findings.

ContentVec: An Improved Self-Supervised Speech Representation by Disentangling Speakers

- Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, Shiyu Chang
- abstract: Self-supervised learning in speech involves training a speech representation network on a large-scale unannotated speech corpus, and then applying the learned representations to downstream tasks. Since the majority of the downstream tasks of SSL learning in speech largely focus on the content information in speech, the most desirable speech representations should be able to disentangle unwanted variations, such as speaker variations, from the content. However, disentangling speakers is very challenging, because removing the speaker information could easily result in a loss of content as well, and the damage of the latter usually far outweighs the benefit of the former. In this paper, we propose a new SSL method that can achieve speaker disentanglement without severe loss of content. Our approach is adapted from the HuBERT framework, and incorporates disentangling mechanisms to regularize both the teacher labels and the learned representations. We evaluate the benefit of speaker disentanglement on a set of content-related downstream tasks, and observe a consistent and notable performance advantage of our speaker-disentangled representations.

Interventional Contrastive Learning with Meta Semantic Regularizer

- Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Bing Su, Hui Xiong
- abstract: Contrastive learning (CL)-based self-supervised learning models learn visual representations in a pairwise manner. Although the prevailing CL model has achieved great progress, in this paper, we uncover an ever-overlooked phenomenon: When the CL model is trained with full images, the performance tested in full images is better than that in foreground areas; when the CL model is trained with foreground areas, the performance tested in full images is worse than that in foreground areas. This observation reveals that backgrounds in images may interfere with the model learning semantic information and their influence has not been fully eliminated. To tackle this issue, we build a Structural Causal Model (SCM) to model the background as a confounder. We propose a backdoor adjustment-based regularization method, namely Interventional Contrastive Learning with Meta Semantic Regularizer (ICL-MSR), to perform causal intervention towards the proposed SCM. ICL-MSR can be incorporated into any existing CL methods to alleviate background distractions from representation learning. Theoretically, we prove that ICL-MSR achieves a tighter error bound. Empirically, our experiments on multiple benchmark datasets demonstrate that ICL-MSR is able to improve the performances of different state-of-the-art CL methods.

Sample-Efficient Reinforcement Learning with loglog(T) Switching Cost

- Dan Qiao, Ming Yin, Ming Min, Yu-Xiang Wang

- abstract: We study the problem of reinforcement learning (RL) with low (policy) switching cost {—} a problem well-motivated by real-life RL applications in which deployments of new policies are costly and the number of policy updates must be low. In this paper, we propose a new algorithm based on stage-wise exploration and adaptive policy elimination that achieves a regret of $\widetilde{O}(\sqrt{H^4S^2AT})$ while requiring a switching cost of $O(HSA \log\log T)$. This is an exponential improvement over the best-known switching cost $O(H^2SA\log T)$ among existing methods with $\widetilde{O}(\mathrm{poly}(H,S,A)\sqrt{T})$ regret. In the above, S,A denotes the number of states and actions in an H-horizon episodic Markov Decision Process model with unknown transitions, and T is the number of steps. As a byproduct of our new techniques, we also derive a reward-free exploration algorithm with a switching cost of $O(HSA)$. Furthermore, we prove a pair of information-theoretical lower bounds which say that (1) Any no-regret algorithm must have a switching cost of $\Omega(HSA)$; (2) Any $\widetilde{O}(\sqrt{T})$ regret algorithm must incur a switching cost of $\Omega(HSA\log\log T)$. Both our algorithms are thus optimal in their switching costs.

Generalizing to Evolving Domains with Latent Structure-Aware Sequential Autoencoder

- Tiexin Qin, Shiqi Wang, Haoliang Li
- abstract: Domain generalization aims to improve the generalization capability of machine learning systems to out-of-distribution (OOD) data. Existing domain generalization techniques embark upon stationary and discrete environments to tackle the generalization issue caused by OOD data. However, many real-world tasks in non-stationary environments (e.g., self-driven car system, sensor measures) involve more complex and continuously evolving domain drift, which raises new challenges for the problem of domain generalization. In this paper, we formulate the aforementioned setting as the problem of evolving domain generalization. Specifically, we propose to introduce a probabilistic framework called Latent Structure-aware Sequential Autoencoder (LSSAE) to tackle the problem of evolving domain generalization via exploring the underlying continuous structure in the latent space of deep neural networks, where we aim to identify two major factors namely covariate shift and concept shift accounting for distribution shift in non-stationary environments. Experimental results on both synthetic and real-world datasets show that LSSAE can lead to superior performances based on the evolving domain generalization setting.

Graph Neural Architecture Search Under Distribution Shifts

- Yijian Qin, Xin Wang, Ziwei Zhang, Pengtao Xie, Wenwu Zhu
- abstract: Graph neural architecture search has shown great potentials for automatically designing graph neural network (GNN) architectures for graph classification tasks. However, when there is a distribution shift between training and testing graphs, the existing approaches fail to deal with the problem of adapting to unknown test graph structures since they only search for a fixed architecture for all graphs. To solve this problem, we propose a novel GRACES model which is able to generalize under distribution shifts through tailoring a customized GNN architecture suitable for each graph instance with unknown distribution. Specifically, we design a self-supervised disentangled graph encoder to characterize invariant factors hidden in diverse graph structures. Then, we propose a prototype-based architecture customization strategy to generate the most suitable GNN architecture weights in a continuous space for each graph instance. We further propose a customized super-network to share weights among different architectures for the sake of efficient training. Extensive experiments on both synthetic and real-world datasets demonstrate that our proposed GRACES model can adapt to diverse graph structures and achieve state-of-the-art performance for graph classification tasks under distribution shifts.

Spectral Representation of Robustness Measures for Optimization Under Input Uncertainty

- Jixiang Qing, Tom Dhaene, Ivo Couckuyt
- abstract: We study the inference of mean-variance robustness measures to quantify input uncertainty under the Gaussian Process (GP) framework. These measures are widely used in applications where the robustness of the solution is of interest, for example, in engineering design. While the variance is commonly used to characterize the robustness, Bayesian inference of the variance using GPs is known to be challenging. In this paper, we propose a Spectral Representation of Robustness Measures based on the GP's spectral representation, i.e., an analytical approach to approximately infer both robustness measures for normal and uniform input uncertainty distributions. We present two approximations based on different Fourier features and compare their accuracy numerically. To demonstrate their utility and efficacy in robust Bayesian Optimization, we integrate the analytical robustness measures in three standard acquisition functions for various robust optimization formulations. We show their competitive performance on numerical benchmarks and real-life applications.

Large-scale Stochastic Optimization of NDCG Surrogates for Deep Learning with Provable Convergence

- Zi-Hao Qiu, Quanqi Hu, Yongjian Zhong, Lijun Zhang, Tianbao Yang
- abstract: NDCG, namely Normalized Discounted Cumulative Gain, is a widely used ranking metric in information retrieval and machine learning. However, efficient and provable stochastic methods for maximizing NDCG are still lacking, especially for deep models. In this paper, we propose a principled approach to optimize NDCG and its top-K variant. First, we formulate a novel compositional optimization problem for optimizing the NDCG surrogate, and a novel bilevel compositional optimization problem for optimizing the top-K NDCG surrogate. Then, we develop efficient stochastic algorithms with provable convergence guarantees for the non-convex objectives. Different from existing NDCG optimization methods, the per-iteration complexity of our algorithms scales with the mini-batch size instead of the number of total items. To improve the effectiveness for deep learning, we further propose practical strategies by using initial warm-up and stop gradient operator. Experimental results on multiple datasets demonstrate that our methods outperform prior ranking approaches in terms of NDCG. To the best of our knowledge, this is the first time that stochastic algorithms are proposed to optimize NDCG with a provable convergence guarantee. Our proposed methods are implemented in the LibAUC library at https://libauc.org.

Latent Outlier Exposure for Anomaly Detection with Contaminated Data

- Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, Stephan Mandt
- abstract: Anomaly detection aims at identifying data points that show systematic deviations from the majority of data in an unlabeled dataset. A common assumption is that clean training data (free of anomalies) is available, which is often violated in practice. We propose a strategy for training an anomaly detector in the presence of unlabeled anomalies that is compatible with a broad class of models. The idea is to jointly infer binary labels to each datum (normal vs. anomalous) while updating the model parameters. Inspired by outlier exposure (Hendrycks et al., 2018) that considers synthetically created, labeled anomalies, we thereby use a combination of two losses that share parameters: one for the normal and one for the anomalous data. We then iteratively proceed with block coordinate updates on the parameters and the most likely (latent) labels. Our experiments with several backbone models on three image datasets, 30 tabular data sets, and a video anomaly detection benchmark showed consistent and significant improvements over the baselines.

Contrastive UCB: Provably Efficient Contrastive Self-Supervised Learning in Online Reinforcement Learning

- Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, Zhaoran Wang
- abstract: In view of its power in extracting feature representation, contrastive self-supervised learning has been successfully integrated into the practice of (deep) reinforcement learning (RL), leading to efficient policy learning on various applications. Despite its tremendous empirical successes, the understanding of contrastive learning for RL remains elusive. To narrow such a gap, we study contrastive-learning empowered RL for a class of Markov decision processes (MDPs) and Markov games (MGs) with low-rank transitions. For both models, we propose to extract the correct feature representations of the low-rank model by minimizing a contrastive loss. Moreover, under the online setting, we propose novel upper confidence bound (UCB)-type algorithms that incorporate such a contrastive loss with online RL algorithms for MDPs or MGs. We further theoretically prove that our algorithm recovers the true representations and simultaneously achieves sample efficiency in learning the optimal policy and Nash equilibrium in MDPs and MGs.

We also provide empirical studies to demonstrate the efficacy of the UCB-based contrastive learning method for RL. To the best of our knowledge, we provide the first provably efficient online RL algorithm that incorporates contrastive learning for representation learning.

Fast and Provable Nonconvex Tensor RPCA

- Haiquan Qiu, Yao Wang, Shaojie Tang, Deyu Meng, Quanming Yao
- abstract: In this paper, we study nonconvex tensor robust principal component analysis (RPCA) based on the t-SVD. We first propose an alternating projection method, i.e., APT, which converges linearly to the ground-truth under the incoherence conditions of tensors. However, as the projection to the low-rank tensor space in APT can be slow, we further propose to speedup such a process by utilizing the property of the tangent space of low-rank. The resulting algorithm, i.e., EAPT, is not only more efficient than APT but also keeps the linear convergence. Compared with existing tensor RPCA works, the proposed method, especially EAPT, is not only more effective due to the recovery guarantee and adaption in the transformed (frequency) domain but also more efficient due to faster convergence rate and lower iteration complexity. These benefits are also empirically verified both on synthetic data, and real applications, e.g., hyperspectral image denoising and video background subtraction.

Generalized Federated Learning via Sharpness Aware Minimization

- Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, Zhuo Lu
- abstract: Federated Learning (FL) is a promising framework for performing privacy-preserving, distributed learning with a set of clients. However, the data distribution among clients often exhibits non-IID, i.e., distribution shift, which makes efficient optimization difficult. To tackle this problem, many FL algorithms focus on mitigating the effects of data heterogeneity across clients by increasing the performance of the global model. However, almost all algorithms leverage Empirical Risk Minimization (ERM) to be the local optimizer, which is easy to make the global model fall into a sharp valley and increase a large deviation of parts of local clients. Therefore, in this paper, we revisit the solutions to the distribution shift problem in FL with a focus on local learning generality. To this end, we propose a general, effective algorithm, \texttt{FedSAM}, based on Sharpness Aware Minimization (SAM) local optimizer, and develop a momentum FL algorithm to bridge local and global models, \texttt{MoFedSAM}. Theoretically, we show the convergence analysis of these two algorithms and demonstrate the generalization bound of \texttt{FedSAM}. Empirically, our proposed algorithms substantially outperform existing FL studies and significantly decrease the learning deviation.

Particle Transformer for Jet Tagging

- Huilin Qu, Congqiao Li, Sitian Qian
- abstract: Jet tagging is a critical yet challenging classification task in particle physics. While deep learning has transformed jet tagging and significantly improved performance, the lack of a large-scale public dataset impedes further enhancement. In this work, we present JetClass, a new comprehensive dataset for jet tagging. The JetClass dataset consists of 100 M jets, about two orders of magnitude larger than existing public datasets. A total of 10 types of jets are simulated, including several types unexplored for tagging so far. Based on the large dataset, we propose a new Transformer-based architecture for jet tagging, called Particle Transformer (ParT). By incorporating pairwise particle interactions in the attention mechanism, ParT achieves higher tagging performance than a plain Transformer and surpasses the previous state-of-the-art, ParticleNet, by a large margin. The pre-trained ParT models, once fine-tuned, also substantially enhance the performance on two widely adopted jet tagging benchmarks. The dataset, code and models are publicly available at https://github.com/jet-universe/particle_transformer.

Winning the Lottery Ahead of Time: Efficient Early Network Pruning

- John Rachwan, Daniel Zügner, Bertrand Charpentier, Simon Geisler, Morgane Ayle, Stephan Günnemann
- abstract: Pruning, the task of sparsifying deep neural networks, received increasing attention recently. Although state-of-the-art pruning methods extract highly sparse models, they neglect two main challenges: (1) the process of finding these sparse models is often very expensive; (2) unstructured pruning does not provide benefits in terms of GPU memory, training time, or carbon emissions. We propose Early Compression via Gradient Flow Preservation (EarlyCroP), which efficiently extracts state-of-the-art sparse models before or early in training addressing challenge (1), and can be applied in a structured manner addressing challenge (2). This enables us to train sparse networks on commodity GPUs whose dense versions would be too large, thereby saving costs and reducing hardware requirements. We empirically show that EarlyCroP outperforms a rich set of baselines for many tasks (incl. classification, regression) and domains (incl. computer vision, natural language processing, and reinforcment learning). EarlyCroP leads to accuracy comparable to dense training while outperforming pruning baselines.

Convergence of Uncertainty Sampling for Active Learning

- Anant Raj, Francis Bach
- abstract: Uncertainty sampling in active learning is heavily used in practice to reduce the annotation cost. However, there has been no wide consensus on the function to be used for uncertainty estimation in binary classification tasks and convergence guarantees of the corresponding active learning algorithms are not well understood. The situation is even more challenging for multi-category classification. In this work, we propose an efficient uncertainty estimator for binary classification which we also extend to multiple classes, and provide a non-asymptotic rate of convergence for our uncertainty sampling based active learning algorithm in both cases under no-noise conditions (i.e., linearly separable data). We also extend our analysis to the noisy case and provide theoretical guarantees for our algorithm under the influence of noise in the task of binary and multi-class classification.

DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale

- Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, Yuxiong He
- abstract: As the training of giant dense models hits the boundary on the availability and capability of the hardware resources today, Mixture-of-Experts (MoE) models have become one of the most promising model architectures due to their significant training cost reduction compared to quality-equivalent dense models. Their training cost saving is demonstrated from encoder-decoder models (prior works) to a 5x saving for auto-aggressive language models (this work). However, due to the much larger model size and unique architecture, how to provide fast MoE model inference remains challenging and unsolved, limiting their practical usage. To tackle this, we present DeepSpeed-MoE, an end-to-end MoE training and inference solution, including novel MoE architecture designs and model compression techniques that reduce MoE model size by up to 3.7x, and a highly optimized inference system that provides 7.3x better latency and cost compared to existing MoE inference solutions. DeepSpeed-MoE offers an unprecedented scale and efficiency to serve massive MoE models with up to 4.5x faster and 9x cheaper inference compared to quality-equivalent dense models. We hope our innovations and systems help open a promising path to new directions in the large model landscape, a shift from dense to sparse MoE models, where training and deploying higher-quality models with fewer resources becomes more widely possible.

Fishr: Invariant Gradient Variances for Out-of-Distribution Generalization

- Alexandre Rame, Corentin Dancette, Matthieu Cord
- abstract: Learning robust models that generalize well under changes in the data distribution is critical for real-world applications. To this end, there has been a growing surge of interest to learn simultaneously from multiple training domains - while enforcing different types of invariance across those domains. Yet, all existing approaches fail to show systematic benefits under controlled evaluation protocols. In this paper, we introduce a new

regularization - named Fishr - that enforces domain invariance in the space of the gradients of the loss: specifically, the domain-level variances of gradients are matched across training domains. Our approach is based on the close relations between the gradient covariance, the Fisher Information and the Hessian of the loss: in particular, we show that Fishr eventually aligns the domain-level loss landscapes locally around the final weights. Extensive experiments demonstrate the effectiveness of Fishr for out-of-distribution generalization. Notably, Fishr improves the state of the art on the DomainBed benchmark and performs consistently better than Empirical Risk Minimization. Our code is available at https://github.com/alexrame/fishr.

A Closer Look at Smoothness in Domain Adversarial Training

- Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, Venkatesh Babu Radhakrishnan
- abstract: Domain adversarial training has been ubiquitous for achieving invariant representations and is used widely for various domain adaptation tasks. In recent times, methods converging to smooth optima have shown improved generalization for supervised learning tasks like classification. In this work, we analyze the effect of smoothness enhancing formulations on domain adversarial training, the objective of which is a combination of task loss (eg. classification, regression etc.) and adversarial terms. We find that converging to a smooth minima with respect to (w.r.t.) task loss stabilizes the adversarial training leading to better performance on target domain. In contrast to task loss, our analysis shows that converging to smooth minima w.r.t. adversarial loss leads to sub-optimal generalization on the target domain. Based on the analysis, we introduce the Smooth Domain Adversarial Training (SDAT) procedure, which effectively enhances the performance of existing domain adversarial methods for both classification and object detection tasks. Our analysis also provides insight into the extensive usage of SGD over Adam in the community for domain adversarial training.

Linear Adversarial Concept Erasure

- Shauli Ravfogel, Michael Twiton, Yoav Goldberg, Ryan D Cotterell
- abstract: Modern neural models trained on textual data rely on pre-trained representations that emerge without direct supervision. As these representations are increasingly being used in real-world applications, the inability to control their content becomes an increasingly important problem. In this work, we formulate the problem of identifying a linear subspace that corresponds to a given concept, and removing it from the representation. We formulate this problem as a constrained, linear minimax game, and show that existing solutions are generally not optimal for this task. We derive a closed-form solution for certain objectives, and propose a convex relaxation that works well for others. When evaluated in the context of binary gender removal, the method recovers a low-dimensional subspace whose removal mitigates bias by intrinsic and extrinsic evaluation. Surprisingly, we show that the method—despite being linear—is highly expressive, effectively mitigating bias in the output layers of deep, nonlinear classifiers while maintaining tractability and interpretability.

Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks

- Noam Razin, Asaf Maman, Nadav Cohen
- abstract: In the pursuit of explaining implicit regularization in deep learning, prominent focus was given to matrix and tensor factorizations, which correspond to simplified neural networks. It was shown that these models exhibit an implicit tendency towards low matrix and tensor ranks, respectively. Drawing closer to practical deep learning, the current paper theoretically analyzes the implicit regularization in hierarchical tensor factorization, a model equivalent to certain deep convolutional neural networks. Through a dynamical systems lens, we overcome challenges associated with hierarchy, and establish implicit regularization towards low hierarchical tensor rank. This translates to an implicit regularization towards locality for the associated convolutional networks. Inspired by our theory, we design explicit regularization discouraging locality, and demonstrate its ability to improve the performance of modern convolutional networks on non-local tasks, in defiance of conventional wisdom by which architectural changes are needed. Our work highlights the potential of enhancing neural networks via theoretical analysis of their implicit regularization.

One-Pass Algorithms for MAP Inference of Nonsymmetric Determinantal Point Processes

- Aravind Reddy, Ryan A. Rossi, Zhao Song, Anup Rao, Tung Mai, Nedim Lipka, Gang Wu, Eunyee Koh, Nesreen Ahmed
- abstract: In this paper, we initiate the study of one-pass algorithms for solving the maximum-a-posteriori (MAP) inference problem for Non-symmetric Determinantal Point Processes (NDPPs). In particular, we formulate streaming and online versions of the problem and provide one-pass algorithms for solving these problems. In our streaming setting, data points arrive in an arbitrary order and the algorithms are constrained to use a single-pass over the data as well as sub-linear memory, and only need to output a valid solution at the end of the stream. Our online setting has an additional requirement of maintaining a valid solution at any point in time. We design new one-pass algorithms for these problems and show that they perform comparably to (or even better than) the offline greedy algorithm while using substantially lower memory.

Universality of Winning Tickets: A Renormalization Group Perspective

- William T Redman, Tianlong Chen, Zhangyang Wang, Akshunna S. Dogra
- abstract: Foundational work on the Lottery Ticket Hypothesis has suggested an exciting corollary: winning tickets found in the context of one task can be transferred to similar tasks, possibly even across different architectures. This has generated broad interest, but methods to study this universality are lacking. We make use of renormalization group theory, a powerful tool from theoretical physics, to address this need. We find that iterative magnitude pruning, the principal algorithm used for discovering winning tickets, is a renormalization group scheme, and can be viewed as inducing a flow in parameter space. We demonstrate that ResNet-50 models with transferable winning tickets have flows with common properties, as would be expected from the theory. Similar observations are made for BERT models, with evidence that their flows are near fixed points. Additionally, we leverage our framework to study winning tickets transferred across ResNet architectures, observing that smaller models have flows with more uniform properties than larger models, complicating transfer between them.

The dynamics of representation learning in shallow, non-linear autoencoders

- Maria Refinetti, Sebastian Goldt
- abstract: Autoencoders are the simplest neural network for unsupervised learning, and thus an ideal framework for studying feature learning. While a detailed understanding of the dynamics of linear autoencoders has recently been obtained, the study of non-linear autoencoders has been hindered by the technical difficulty of handling training data with non-trivial correlations {–} a fundamental prerequisite for feature extraction. Here, we study the dynamics of feature learning in non-linear, shallow autoencoders. We derive a set of asymptotically exact equations that describe the generalisation dynamics of autoencoders trained with stochastic gradient descent (SGD) in the limit of high-dimensional inputs. These equations reveal that autoencoders learn the leading principal components of their inputs sequentially. An analysis of the long-time dynamics explains the failure of sigmoidal autoencoders to learn with tied weights, and highlights the importance of training the bias in ReLU autoencoders. Building on previous results for linear networks, we analyse a modification of the vanilla SGD algorithm which allows learning of the exact principal components. Finally, we show that our equations accurately describe the generalisation dynamics of non-linear autoencoders on realistic datasets such as CIFAR10.

Proximal Exploration for Model-guided Protein Sequence Design

- Zhizhou Ren, Jiahan Li, Fan Ding, Yuan Zhou, Jianzhu Ma, Jian Peng

- abstract: Designing protein sequences with a particular biological function is a long-lasting challenge for protein engineering. Recent advances in machine-learning-guided approaches focus on building a surrogate sequence-function model to reduce the burden of expensive in-lab experiments. In this paper, we study the exploration mechanism of model-guided sequence design. We leverage a natural property of protein fitness landscape that a concise set of mutations upon the wild-type sequence are usually sufficient to enhance the desired function. By utilizing this property, we propose Proximal Exploration (PEX) algorithm that prioritizes the evolutionary search for high-fitness mutants with low mutation counts. In addition, we develop a specialized model architecture, called Mutation Factorization Network (MuFacNet), to predict low-order mutational effects, which further improves the sample efficiency of model-guided evolution. In experiments, we extensively evaluate our method on a suite of in-silico protein sequence design tasks and demonstrate substantial improvement over baseline algorithms.

Towards Theoretical Analysis of Transformation Complexity of ReLU DNNs

- Jie Ren, Mingjie Li, Meng Zhou, Shih-Han Chan, Quanshi Zhang
- abstract: This paper aims to theoretically analyze the complexity of feature transformations encoded in piecewise linear DNNs with ReLU layers. We propose metrics to measure three types of complexities of transformations based on the information theory. We further discover and prove the strong correlation between the complexity and the disentanglement of transformations. Based on the proposed metrics, we analyze two typical phenomena of the change of the transformation complexity during the training process, and explore the ceiling of a DNN's complexity. The proposed metrics can also be used as a loss to learn a DNN with the minimum complexity, which also controls the over-fitting level of the DNN and influences adversarial robustness, adversarial transferability, and knowledge consistency. Comprehensive comparative studies have provided new perspectives to understand the DNN. The code is released at https://github.com/sjtu-XAI-lab/transformation-complexity.

Benchmarking and Analyzing Point Cloud Classification under Corruptions

- Jiawei Ren, Liang Pan, Ziwei Liu
- abstract: 3D perception, especially point cloud classification, has achieved substantial progress. However, in real-world deployment, point cloud corruptions are inevitable due to the scene complexity, sensor inaccuracy, and processing imprecision. In this work, we aim to rigorously benchmark and analyze point cloud classification under corruptions. To conduct a systematic investigation, we first provide a taxonomy of common 3D corruptions and identify the atomic corruptions. Then, we perform a comprehensive evaluation on a wide range of representative point cloud models to understand their robustness and generalizability. Our benchmark results show that although point cloud classification performance improves over time, the state-of-the-art methods are on the verge of being less robust. Based on the obtained observations, we propose several effective techniques to enhance point cloud classifier robustness. We hope our comprehensive benchmark, in-depth analysis, and proposed techniques could spark future research in robust 3D perception.

A Unified View on PAC-Bayes Bounds for Meta-Learning

- Arezou Rezazadeh
- abstract: Meta learning automatically infers an inductive bias, that includes the hyperparameter of the baselearning algorithm, by observing data from a finite number of related tasks. This paper studies PAC-Bayes bounds on meta generalization gap. The meta-generalization gap comprises two sources of generalization gaps: the environmentlevel and task-level gaps resulting from observation of a finite number of tasks and data samples per task, respectively. In this paper, by upper bounding arbitrary convex functions, which link the expected and empirical losses at the environment and also per-task levels, we obtain new PAC-Bayes bounds. Using these bounds, we develop new PAC-Bayes meta-learning algorithms. Numerical examples demonstrate the merits of the proposed novel bounds and algorithm in comparison to prior PAC-Bayes bounds for meta-learning

3PC: Three Point Compressors for Communication-Efficient Distributed Training and a Better Theory for Lazy Aggregation

- Peter Richtarik, Igor Sokolov, Elnur Gasanov, Ilyas Fatkhullin, Zhize Li, Eduard Gorbunov
- abstract: We propose and study a new class of gradient compressors for communication-efficient training—three point compressors (3PC)—as well as efficient distributed nonconvex optimization algorithms that can take advantage of them. Unlike most established approaches, which rely on a static compressor choice (e.g., TopK), our class allows the compressors to evolve throughout the training process, with the aim of improving the theoretical communication complexity and practical efficiency of the underlying methods. We show that our general approach can recover the recently proposed state-of-the-art error feedback mechanism EF21 (Richtárik et al, 2021) and its theoretical properties as a special case, but also leads to a number of new efficient methods. Notably, our approach allows us to improve upon the state-of-the-art in the algorithmic and theoretical foundations of the lazy aggregation literature (Liu et al, 2017; Lan et al, 2017). As a by-product that may be of independent interest, we provide a new and fundamental link between the lazy aggregation and error feedback literature. A special feature of our work is that we do not require the compressors to be unbiased.

Robust SDE-Based Variational Formulations for Solving Linear PDEs via Deep Learning

- Lorenz Richter, Julius Berner
- abstract: The combination of Monte Carlo methods and deep learning has recently led to efficient algorithms for solving partial differential equations (PDEs) in high dimensions. Related learning problems are often stated as variational formulations based on associated stochastic differential equations (SDEs), which allow the minimization of corresponding losses using gradient-based optimization methods. In respective numerical implementations it is therefore crucial to rely on adequate gradient estimators that exhibit low variance in order to reach convergence accurately and swiftly. In this article, we rigorously investigate corresponding numerical aspects that appear in the context of linear Kolmogorov PDEs. In particular, we systematically compare existing deep learning approaches and provide theoretical explanations for their performances. Subsequently, we suggest novel methods that can be shown to be more robust both theoretically and numerically, leading to substantial performance improvements.

Probabilistically Robust Learning: Balancing Average and Worst-case Performance

- Alexander Robey, Luiz Chamon, George J. Pappas, Hamed Hassani
- abstract: Many of the successes of machine learning are based on minimizing an averaged loss function. However, it is well-known that this paradigm suffers from robustness issues that hinder its applicability in safety-critical domains. These issues are often addressed by training against worst-case perturbations of data, a technique known as adversarial training. Although empirically effective, adversarial training can be overly conservative, leading to unfavorable trade-offs between nominal performance and robustness. To this end, in this paper we propose a framework called probabilistic robustness that bridges the gap between the accurate, yet brittle average case and the robust, yet conservative worst case by enforcing robustness to most rather than to all perturbations. From a theoretical point of view, this framework overcomes the trade-offs between the performance and the sample-complexity of worst-case and average-case learning. From a practical point of view, we propose a novel algorithm based on risk-aware optimization that effectively balances average- and worst-case performance at a considerably lower computational cost relative to adversarial training. Our results on MNIST, CIFAR-10, and SVHN illustrate the advantages of this framework on the spectrum from average- to worst-case robustness. Our code is available at: https://github.com/arobey1/advbench.

LyaNet: A Lyapunov Framework for Training Neural ODEs

- Ivan Dario Jimenez Rodriguez, Aaron Ames, Yisong Yue
- abstract: We propose a method for training ordinary differential equations by using a control-theoretic Lyapunov condition for stability. Our approach, called LyaNet, is based on a novel Lyapunov loss formulation that encourages the inference dynamics to converge quickly to the correct prediction. Theoretically, we show that minimizing Lyapunov loss guarantees exponential convergence to the correct solution and enables a novel robustness guarantee. We also provide practical algorithms, including one that avoids the cost of backpropagating through a solver or using the adjoint method. Relative to standard Neural ODE training, we empirically find that LyaNet can offer improved prediction performance, faster convergence of inference dynamics, and improved adversarial robustness. Our code is available at https://github.com/ivandariojr/LyapunovLearning.

Short-Term Plasticity Neurons Learning to Learn and Forget

- Hector Garcia Rodriguez, Qinghai Guo, Timoleon Moraitis
- abstract: Short-term plasticity (STP) is a mechanism that stores decaying memories in synapses of the cerebral cortex. In computing practice, STP has been used, but mostly in the niche of spiking neurons, even though theory predicts that it is the optimal solution to certain dynamic tasks. Here we present a new type of recurrent neural unit, the STP Neuron (STPN), which indeed turns out strikingly powerful. Its key mechanism is that synapses have a state, propagated through time by a self-recurrent connection-within-the-synapse. This formulation enables training the plasticity with backpropagation through time, resulting in a form of learning to learn and forget in the short term. The STPN outperforms all tested alternatives, i.e. RNNs, LSTMs, other models with fast weights, and differentiable plasticity. We confirm this in both supervised and reinforcement learning (RL), and in tasks such as Associative Retrieval, Maze Exploration, Atari video games, and MuJoCo robotics. Moreover, we calculate that, in neuromorphic or biological circuits, the STPN minimizes energy consumption across models, as it depresses individual synapses dynamically. Based on these, biological STP may have been a strong evolutionary attractor that maximizes both efficiency and computational power. The STPN now brings these neuromorphic advantages also to a broad spectrum of machine learning practice. Code is available in https://github.com/NeuromorphicComputing/stpn.

Function-space Inference with Sparse Implicit Processes

- Simon Rodríguez-Santana, Bryan Zaldivar, Daniel Hernandez-Lobato
- abstract: Implicit Processes (IPs) represent a flexible framework that can be used to describe a wide variety of models, from Bayesian neural networks, neural samplers and data generators to many others. IPs also allow for approximate inference in function-space. This change of formulation solves intrinsic degenerate problems of parameter-space approximate inference concerning the high number of parameters and their strong dependencies in large models. For this, previous works in the literature have attempted to employ IPs both to set up the prior and to approximate the resulting posterior. However, this has proven to be a challenging task. Existing methods that can tune the prior IP result in a Gaussian predictive distribution, which fails to capture important data patterns. By contrast, methods producing flexible predictive distributions by using another IP to approximate the posterior process cannot tune the prior IP to the observed data. We propose here the first method that can accomplish both goals. For this, we rely on an inducing-point representation of the prior IP, as often done in the context of sparse Gaussian processes. The result is a scalable method for approximate inference with IPs that can tune the prior IP parameters to the data, and that provides accurate non-Gaussian predictive distributions.

Score Matching Enables Causal Discovery of Nonlinear Additive Noise Models

- Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, Francesco Locatello
- abstract: This paper demonstrates how to recover causal graphs from the score of the data distribution in non-linear additive (Gaussian) noise models. Using score matching algorithms as a building block, we show how to design a new generation of scalable causal discovery methods. To showcase our approach, we also propose a new efficient method for approximating the score's Jacobian, enabling to recover the causal graph. Empirically, we find that the new algorithm, called SCORE, is competitive with state-of-the-art causal discovery methods while being significantly faster.

Dual Decomposition of Convex Optimization Layers for Consistent Attention in Medical Images

- Tom Ron, Tamir Hazan
- abstract: A key concern in integrating machine learning models in medicine is the ability to interpret their reasoning. Popular explainability methods have demonstrated satisfactory results in natural image recognition, yet in medical image analysis, many of these approaches provide partial and noisy explanations. Recently, attention mechanisms have shown compelling results both in their predictive performance and in their interpretable qualities. A fundamental trait of attention is that it leverages salient parts of the input which contribute to the model's prediction. To this end, our work focuses on the explanatory value of attention weight distributions. We propose a multi-layer attention mechanism that enforces consistent interpretations between attended convolutional layers using convex optimization. We apply duality to decompose the consistency constraints between the layers by reparameterizing their attention probability distributions. We further suggest learning the dual witness by optimizing with respect to our objective; thus, our implementation uses standard back-propagation, hence it is highly efficient. While preserving predictive performance, our proposed method leverages weakly annotated medical imaging data and provides complete and faithful explanations to the model's prediction.

A Consistent and Efficient Evaluation Strategy for Attribution Methods

- Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, Enkelejda Kasneci
- abstract: With a variety of local feature attribution methods being proposed in recent years, follow-up work suggested several evaluation strategies. To assess the attribution quality across different attribution techniques, the most popular among these evaluation strategies in the image domain use pixel perturbations. However, recent advances discovered that different evaluation strategies produce conflicting rankings of attribution methods and can be prohibitively expensive to compute. In this work, we present an information-theoretic analysis of evaluation strategies based on pixel perturbations. Our findings reveal that the results are strongly affected by information leakage through the shape of the removed pixels as opposed to their actual values. Using our theoretical insights, we propose a novel evaluation framework termed Remove and Debias (ROAD) which offers two contributions: First, it mitigates the impact of the confounders, which entails higher consistency among evaluation strategies. Second, ROAD does not require the computationally expensive retraining step and saves up to 99% in computational costs compared to the state-of-the-art. We release our source code at https://github.com/tleemann/road_evaluation.

Efficiently Learning the Topology and Behavior of a Networked Dynamical System Via Active Queries

- Daniel J Rosenkrantz, Abhijin Adiga, Madhav Marathe, Zirou Qiu, S S Ravi, Richard Stearns, Anil Vullikanti
- abstract: Using a discrete dynamical system model, many papers have addressed the problem of learning the behavior (i.e., the local function at each node) of a networked system through active queries, assuming that the network topology is known. We address the problem of inferring both the topology of the network and the behavior of a discrete dynamical system through active queries. We consider two query models studied in the literature, namely the batch model (where all the queries must be submitted together) and the adaptive model (where responses to previous queries can be used in formulating a new query). Our results are for systems where the state of each node is from {0,1} and the local functions are Boolean. We present algorithms to learn the topology and the behavior under both batch and adaptive query models for several classes of dynamical systems. These algorithms use only a polynomial number of queries. We also present experimental results obtained by running our query generation algorithms on synthetic and real-world networks.

Learning to Infer Structures of Network Games

- Emanuele Rossi, Federico Monti, Yan Leng, Michael Bronstein, Xiaowen Dong
- abstract: Strategic interactions between a group of individuals or organisations can be modelled as games played on networks, where a player's payoff depends not only on their actions but also on those of their neighbours. Inferring the network structure from observed game outcomes (equilibrium actions) is an important problem with numerous potential applications in economics and social sciences. Existing methods mostly require the knowledge of the utility function associated with the game, which is often unrealistic to obtain in real-world scenarios. We adopt a transformer-like architecture which correctly accounts for the symmetries of the problem and learns a mapping from the equilibrium actions to the network structure of the game without explicit knowledge of the utility function. We test our method on three different types of network games using both synthetic and real-world data, and demonstrate its effectiveness in network structure inference and superior performance over existing methods.

Direct Behavior Specification via Constrained Reinforcement Learning

- Julien Roy, Roger Girgis, Joshua Romoff, Pierre-Luc Bacon, Chris J Pal
- abstract: The standard formulation of Reinforcement Learning lacks a practical way of specifying what are admissible and forbidden behaviors. Most often, practitioners go about the task of behavior specification by manually engineering the reward function, a counter-intuitive process that requires several iterations and is prone to reward hacking by the agent. In this work, we argue that constrained RL, which has almost exclusively been used for safe RL, also has the potential to significantly reduce the amount of work spent for reward specification in applied RL projects. To this end, we propose to specify behavioral preferences in the CMDP framework and to use Lagrangian methods to automatically weigh each of these behavioral constraints. Specifically, we investigate how CMDPs can be adapted to solve goal-based tasks while adhering to several constraints simultaneously. We evaluate this framework on a set of continuous control tasks relevant to the application of Reinforcement Learning for NPC design in video games.

Constraint-based graph network simulator

- Yulia Rubanova, Alvaro Sanchez-Gonzalez, Tobias Pfaff, Peter Battaglia
- abstract: In the area of physical simulations, nearly all neural-network-based methods directly predict future states from the input states. However, many traditional simulation engines instead model the constraints of the system and select the state which satisfies them. Here we present a framework for constraint-based learned simulation, where a scalar constraint function is implemented as a graph neural network, and future predictions are computed by solving the optimization problem defined by the learned constraint. Our model achieves comparable or better accuracy to top learned simulators on a variety of challenging physical domains, and offers several unique advantages. We can improve the simulation accuracy on a larger system by applying more solver iterations at test time. We also can incorporate novel hand-designed constraints at test time and simulate new dynamics which were not present in the training data. Our constraint-based framework shows how key techniques from traditional simulation and numerical methods can be leveraged as inductive biases in machine learning simulators.

Continual Learning via Sequential Function-Space Variational Inference

- Tim G. J. Rudner, Freddie Bickford Smith, Qixuan Feng, Yee Whye Teh, Yarin Gal
- abstract: Sequential Bayesian inference over predictive functions is a natural framework for continual learning from streams of data. However, applying it to neural networks has proved challenging in practice. Addressing the drawbacks of existing techniques, we propose an optimization objective derived by formulating continual learning as sequential function-space variational inference. In contrast to existing methods that regularize neural network parameters directly, this objective allows parameters to vary widely during training, enabling better adaptation to new tasks. Compared to objectives that directly regularize neural network predictions, the proposed objective allows for more flexible variational distributions and more effective regularization. We demonstrate that, across a range of task sequences, neural networks trained via sequential function-space variational inference achieve better predictive accuracy than networks trained with related methods while depending less on maintaining a set of representative points from previous tasks.

Graph-Coupled Oscillator Networks

- T. Konstantin Rusch, Ben Chamberlain, James Rowbottom, Siddhartha Mishra, Michael Bronstein
- abstract: We propose Graph-Coupled Oscillator Networks (GraphCON), a novel framework for deep learning on graphs. It is based on discretizations of a second-order system of ordinary differential equations (ODEs), which model a network of nonlinear controlled and damped oscillators, coupled via the adjacency structure of the underlying graph. The flexibility of our framework permits any basic GNN layer (e.g. convolutional or attentional) as the coupling function, from which a multi-layer deep neural network is built up via the dynamics of the proposed ODEs. We relate the oversmoothing problem, commonly encountered in GNNs, to the stability of steady states of the underlying ODE and show that zero-Dirichlet energy steady states are not stable for our proposed ODEs. This demonstrates that the proposed framework mitigates the oversmoothing problem. Moreover, we prove that GraphCON mitigates the exploding and vanishing gradients problem to facilitate training of deep multi-layer GNNs. Finally, we show that our approach offers competitive performance with respect to the state-of-the-art on a variety of graph-based learning tasks.

Hindering Adversarial Attacks with Implicit Neural Representations

- Andrei A Rusu, Dan Andrei Calian, Sven Gowal, Raia Hadsell
- abstract: We introduce the Lossy Implicit Network Activation Coding (LINAC) defence, an input transformation which successfully hinders several common adversarial attacks on CIFAR-10 classifiers for perturbations up to 8/255 in Linf norm and 0.5 in L2 norm. Implicit neural representations are used to approximately encode pixel colour intensities in 2D images such that classifiers trained on transformed data appear to have robustness to small perturbations without adversarial training or large drops in performance. The seed of the random number generator used to initialise and train the implicit neural representation turns out to be necessary information for stronger generic attacks, suggesting its role as a private key. We devise a Parametric Bypass Approximation (PBA) attack strategy for key-based defences, which successfully invalidates an existing method in this category. Interestingly, our LINAC defence also hinders some transfer and adaptive attacks, including our novel PBA strategy. Our results emphasise the importance of a broad range of customised attacks despite apparent robustness according to standard evaluations.

Exploiting Independent Instruments: Identification and Distribution Generalization

- Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, Jonas Peters
- abstract: Instrumental variable models allow us to identify a causal function between covariates X and a response Y, even in the presence of unobserved confounding. Most of the existing estimators assume that the error term in the response Y and the hidden confounders are uncorrelated with the instruments Z. This is often motivated by a graphical separation, an argument that also justifies independence. Positing an independence restriction, however, leads to strictly stronger identifiability results. We connect to the existing literature in econometrics and provide a practical method called HSIC-X for exploiting independence that can be combined with any gradient-based learning procedure. We see that even in identifiable settings, taking into account higher moments may yield better finite sample results. Furthermore, we exploit the independence for distribution generalization. We prove that the proposed estimator is invariant to distributional shifts on the instruments and worst-case optimal whenever these shifts are sufficiently strong. These results hold even in the under-identified case where the instruments are not sufficiently rich to identify the causal function.

FedNL: Making Newton-Type Methods Applicable to Federated Learning

- Mher Safaryan, Rustem Islamov, Xun Qian, Peter Richtarik
- abstract: Inspired by recent work of Islamov et al (2021), we propose a family of Federated Newton Learn (\algname{FedNL}) methods, which we believe is a marked step in the direction of making second-order methods applicable to FL. In contrast to the aforementioned work, \algname{FedNL} employs a different Hessian learning technique which i) enhances privacy as it does not rely on the training data to be revealed to the coordinating server, ii) makes it applicable beyond generalized linear models, and iii) provably works with general contractive compression operators for compressing the local Hessians, such as Top-K or Rank-R, which are vastly superior in practice. Notably, we do not need to rely on error feedback for our methods to work with contractive compressors. Moreover, we develop \algname{FedNL-PP}, \algname{FedNL-CR} and \algname{FedNL-LS}, which are variants of \algname{FedNL} that support partial participation, and globalization via cubic regularization and line search, respectively, and \algname{FedNL-BC}, which is a variant that can further benefit from bidirectional compression of gradients and models, i.e., smart uplink gradient and smart downlink model compression. We prove local convergence rates that are independent of the condition number, the number of training data points, and compression variance. Our communication efficient Hessian learning technique provably learns the Hessian at the optimum. Finally, we perform a variety of numerical experiments that show that our \algname{FedNL} methods have state-of-the-art communication complexity when compared to key baselines.

Versatile Dueling Bandits: Best-of-both World Analyses for Learning from Relative Preferences

- Aadirupa Saha, Pierre Gaillard
- abstract: We study the problem of K-armed dueling bandit for both stochastic and adversarial environments, where the goal of the learner is to aggregate information through relative preferences of pair of decision points queried in an online sequential manner. We first propose a novel reduction from any (general) dueling bandits to multi-armed bandits which allows us to improve many existing results in dueling bandits. In particular, we give the first best-of-both world result for the dueling bandits regret minimization problem—a unified framework that is guaranteed to perform optimally for both stochastic and adversarial preferences simultaneously. Moreover, our algorithm is also the first to achieve an optimal $O(\sum_{i = 1}^K \frac{\log T}{\Delta_i})$ regret bound against the Condorcet-winner benchmark, which scales optimally both in terms of the arm-size K and the instance-specific suboptimality gaps ${\Delta_i}_{i = 1}^K$. This resolves the long standing problem of designing an instancewise gap-dependent order optimal regret algorithm for dueling bandits (with matching lower bounds up to small constant factors). We further justify the robustness of our proposed algorithm by proving its optimal regret rate under adversarially corrupted preferences—this outperforms the existing state-of-the-art corrupted dueling results by a large margin. In summary, we believe our reduction idea will find a broader scope in solving a diverse class of dueling bandits setting, which are otherwise studied separately from multi-armed bandits with often more complex solutions and worse guarantees. The efficacy of our proposed algorithms are empirically corroborated against state-of-the art dueling bandit methods.

Optimal and Efficient Dynamic Regret Algorithms for Non-Stationary Dueling Bandits

- Aadirupa Saha, Shubham Gupta
- abstract: We study the problem of dynamic regret minimization in K-armed Dueling Bandits under non-stationary or time-varying preferences. This is an online learning setup where the agent chooses a pair of items at each round and observes only a relative binary ‘win-loss’ feedback for this pair sampled from an underlying preference matrix at that round. We first study the problem of static-regret minimization for adversarial preference sequences and design an efficient algorithm with $O(\sqrt{KT})$ regret bound. We next use similar algorithmic ideas to propose an efficient and provably optimal algorithm for dynamic-regret minimization under two notions of non-stationarities. In particular, we show $\tO(\sqrt{SKT})$ and $\tO({V_T^{1/3}K^{1/3}T^{2/3}})$ dynamic-regret guarantees, respectively, with S being the total number of ‘effective-switches’ in the underlying preference relations and V_T being a measure of ‘continuous-variation’ non-stationarity. These rates are provably optimal as justified with matching lower bound guarantees. Moreover, our proposed algorithms are flexible as they can be easily ‘blackboxed’ to yield dynamic regret guarantees for other notions of dueling bandits regret, including condorcet regret, best-response bounds, and Borda regret. The complexity of these problems have not been studied prior to this work despite the practicality of non-stationary environments. Our results are corroborated with extensive simulations.

Unraveling Attention via Convex Duality: Analysis and Interpretations of Vision Transformers

- Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, Mert Pilanci
- abstract: Vision transformers using self-attention or its proposed alternatives have demonstrated promising results in many image related tasks. However, the underpinning inductive bias of attention is not well understood. To address this issue, this paper analyzes attention through the lens of convex duality. For the non-linear dot-product self-attention, and alternative mechanisms such as MLP-mixer and Fourier Neural Operator (FNO), we derive equivalent finite-dimensional convex problems that are interpretable and solvable to global optimality. The convex programs lead to block nuclear-norm regularization that promotes low rank in the latent feature and token dimensions. In particular, we show how self-attention networks implicitly clusters the tokens, based on their latent similarity. We conduct experiments for transferring a pre-trained transformer backbone for CIFAR-100 classification by fine-tuning a variety of convex attention heads. The results indicate the merits of the bias induced by attention compared with the existing MLP or linear heads.

Off-Policy Evaluation for Large Action Spaces via Embeddings

- Yuta Saito, Thorsten Joachims
- abstract: Off-policy evaluation (OPE) in contextual bandits has seen rapid adoption in real-world systems, since it enables offline evaluation of new policies using only historic log data. Unfortunately, when the number of actions is large, existing OPE estimators – most of which are based on inverse propensity score weighting – degrade severely and can suffer from extreme bias and variance. This foils the use of OPE in many applications from recommender systems to language models. To overcome this issue, we propose a new OPE estimator that leverages marginalized importance weights when action embeddings provide structure in the action space. We characterize the bias, variance, and mean squared error of the proposed estimator and analyze the conditions under which the action embedding provides statistical benefits over conventional estimators. In addition to the theoretical analysis, we find that the empirical performance improvement can be substantial, enabling reliable OPE even when existing estimators collapse due to a large number of actions.

Optimal Clipping and Magnitude-aware Differentiation for Improved Quantization-aware Training

- Charbel Sakr, Steve Dai, Rangha Venkatesan, Brian Zimmer, William Dally, Brucek Khailany
- abstract: Data clipping is crucial in reducing noise in quantization operations and improving the achievable accuracy of quantization-aware training (QAT). Current practices rely on heuristics to set clipping threshold scalars and cannot be shown to be optimal. We propose Optimally Clipped Tensors And Vectors (OCTAV), a recursive algorithm to determine MSE-optimal clipping scalars. Derived from the fast Newton-Raphson method, OCTAV finds optimal clipping scalars on the fly, for every tensor, at every iteration of the QAT routine. Thus, the QAT algorithm is formulated with provably minimum quantization noise at each step. In addition, we reveal limitations in common gradient estimation techniques in QAT and propose magnitude-aware differentiation as a remedy to further improve accuracy. Experimentally, OCTAV-enabled QAT achieves state-of-the-art accuracy on multiple tasks. These include training-from-scratch and retraining ResNets and MobileNets on ImageNet, and Squad fine-tuning using BERT models, where OCTAV-enabled QAT consistently preserves accuracy at low precision (4-to-6-bits). Our results require no modifications to the baseline training recipe, except for the insertion of quantization operations where appropriate.

A Convergence Theory for SVGD in the Population Limit under Talagrand’s Inequality T1

- Adil Salim, Lukang Sun, Peter Richtarik

- abstract: Stein Variational Gradient Descent (SVGD) is an algorithm for sampling from a target density which is known up to a multiplicative constant. Although SVGD is a popular algorithm in practice, its theoretical study is limited to a few recent works. We study the convergence of SVGD in the population limit, (i.e., with an infinite number of particles) to sample from a non-logconcave target distribution satisfying Talagrand's inequality T1. We first establish the convergence of the algorithm. Then, we establish a dimension-dependent complexity bound in terms of the Kernelized Stein Discrepancy (KSD). Unlike existing works, we do not assume that the KSD is bounded along the trajectory of the algorithm. Our approach relies on interpreting SVGD as a gradient descent over a space of probability measures.

FITNESS: (Fine Tune on New and Similar Samples) to detect anomalies in streams with drift and outliers

- Abishek Sankararaman, Balakrishnan Narayanaswamy, Vikramank Y Singh, Zhao Song
- abstract: Technology improvements have made it easier than ever to collect diverse telemetry at high resolution from any cyber or physical system, for both monitoring and control. In the domain of monitoring, anomaly detection has become an important problem in many research areas ranging from IoT and sensor networks to devOps. These systems operate in real, noisy and non-stationary environments. A fundamental question is then, 'How to quickly spot anomalies in a data-stream, and differentiate them from either sudden or gradual drifts in the normal behaviour?' Although several heuristics have been proposed for detecting anomalies on streams, no known method has formalized the desiderata and rigorously proven that they can be achieved. We begin by formalizing the problem as a sequential estimation task. We propose \name, (\textbf{Fi}ne \textbf{T}une on \textbf{Ne}w and \textbf{S}imilar \textbf{S}amples), a flexible framework for detecting anomalies on data streams. We show that in the case when the data stream has a gaussian distribution, FITNESS is provably both robust and adaptive. The core of our method is to fine-tune the anomaly detection system only on recent, similar examples, before predicting an anomaly score. We prove that this is sufficient for robustness and adaptivity. We further experimentally demonstrate that \name;{is} flexible in practice, i.e., it can convert existing offline AD algorithms in to robust and adaptive online ones.

The Algebraic Path Problem for Graph Metrics

- Enrique Fita Sanmartín, Sebastian Damrich, Fred Hamprecht
- abstract: Finding paths with optimal properties is a foundational problem in computer science. The notions of shortest paths (minimal sum of edge costs), minimax paths (minimal maximum edge weight), reliability of a path and many others all arise as special cases of the "algebraic path problem" (APP). Indeed, the APP formalizes the relation between different semirings such as min-plus, min-max and the distances they induce. We here clarify, for the first time, the relation between the potential distance and the log-semiring. We also define a new unifying family of algebraic structures that include all above-mentioned path problems as well as the commute cost and others as special or limiting cases. The family comprises not only semirings but also strong bimonoids (that is, semirings without distributivity). We call this new and very general distance the "log-norm distance". Finally, we derive some sufficient conditions which ensure that the APP associated with a semiring defines a metric over an arbitrary graph.

LSB: Local Self-Balancing MCMC in Discrete Spaces

- Emanuele Sansone
- abstract: We present the Local Self-Balancing sampler (LSB), a local Markov Chain Monte Carlo (MCMC) method for sampling in purely discrete domains, which is able to autonomously adapt to the target distribution and to reduce the number of target evaluations required to converge. LSB is based on (i) a parametrization of locally balanced proposals, (ii) an objective function based on mutual information and (iii) a self-balancing learning procedure, which minimises the proposed objective to update the proposal parameters. Experiments on energy-based models and Markov networks show that LSB converges using a smaller number of queries to the oracle distribution compared to recent local MCMC samplers.

PoF: Post-Training of Feature Extractor for Improving Generalization

- Ikuro Sato, Yamada Ryota, Masayuki Tanaka, Nakamasa Inoue, Rei Kawakami
- abstract: It has been intensively investigated that the local shape, especially flatness, of the loss landscape near a minimum plays an important role for generalization of deep models. We developed a training algorithm called PoF: Post-Training of Feature Extractor that updates the feature extractor part of an already-trained deep model to search a flatter minimum. The characteristics are two-fold: 1) Feature extractor is trained under parameter perturbations in the higher-layer parameter space, based on observations that suggest flattening higher-layer parameter space, and 2) the perturbation range is determined in a data-driven manner aiming to reduce a part of test loss caused by the positive loss curvature. We provide a theoretical analysis that shows the proposed algorithm implicitly reduces the target Hessian components as well as the loss. Experimental results show that PoF improved model performance against baseline methods on both CIFAR-10 and CIFAR-100 datasets for only 10-epoch post-training, and on SVHN dataset for 50-epoch post-training.

Re-evaluating Word Mover's Distance

- Ryoma Sato, Makoto Yamada, Hisashi Kashima
- abstract: The word mover's distance (WMD) is a fundamental technique for measuring the similarity of two documents. As the crux of WMD, it can take advantage of the underlying geometry of the word space by employing an optimal transport formulation. The original study on WMD reported that WMD outperforms classical baselines such as bag-of-words (BOW) and TF-IDF by significant margins in various datasets. In this paper, we point out that the evaluation in the original study could be misleading. We re-evaluate the performances of WMD and the classical baselines and find that the classical baselines are competitive with WMD if we employ an appropriate preprocessing, i.e., L1 normalization. In addition, we introduce an analogy between WMD and L1-normalized BOW and find that not only the performance of WMD but also the distance values resemble those of BOW in high dimensional spaces.

Understanding Contrastive Learning Requires Incorporating Inductive Biases

- Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, Akshay Krishnamurthy
- abstract: Contrastive learning is a popular form of self-supervised learning that encourages augmentations (views) of the same input to have more similar representations compared to augmentations of different inputs. Recent attempts to theoretically explain the success of contrastive learning on downstream classification tasks prove guarantees depending on properties of augmentations and the value of contrastive loss of representations. We demonstrate that such analyses, that ignore inductive biases of the function class and training algorithm, cannot adequately explain the success of contrastive learning, even provably leading to vacuous guarantees in some settings. Extensive experiments on image and text domains highlight the ubiquity of this problem – different function classes and algorithms behave very differently on downstream tasks, despite having the same augmentations and contrastive losses. Theoretical analysis is presented for the class of linear representations, where incorporating inductive biases of the function class allows contrastive learning to work with less stringent conditions compared to prior analyses.

The Neural Race Reduction: Dynamics of Abstraction in Gated Networks

- Andrew Saxe, Shagun Sodhani, Sam Jay Lewallen
- abstract: Our theoretical understanding of deep learning has not kept pace with its empirical success. While network architecture is known to be critical, we do not yet understand its effect on learned representations and network behavior, or how this architecture should reflect task structure.In this work, we

begin to address this gap by introducing the Gated Deep Linear Network framework that schematizes how pathways of information flow impact learning dynamics within an architecture. Crucially, because of the gating, these networks can compute nonlinear functions of their input. We derive an exact reduction and, for certain cases, exact solutions to the dynamics of learning. Our analysis demonstrates that the learning dynamics in structured networks can be conceptualized as a neural race with an implicit bias towards shared representations, which then govern the model's ability to systematically generalize, multi-task, and transfer. We validate our key insights on naturalistic datasets and with relaxed assumptions. Taken together, our work gives rise to general hypotheses relating neural architecture to learning and provides a mathematical approach towards understanding the design of more complex architectures and the role of modularity and compositionality in solving real-world problems. The code and results are available at https://www.saxelab.org/gated-dln.

Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Lojasiewicz Condition and Local Smoothness

- Kevin Scaman, Cedric Malherbe, Ludovic Dos Santos
- abstract: Training over-parameterized neural networks involves the empirical minimization of highly non-convex objective functions. Recently, a large body of works provided theoretical evidence that, despite this non-convexity, properly initialized over-parameterized networks can converge to a zero training loss through the introduction of the Polyak-Lojasiewicz condition. However, these analyses are restricted to quadratic losses such as mean square error, and tend to indicate fast exponential convergence rates that are seldom observed in practice. In this work, we propose to extend these results by analyzing stochastic gradient descent under more generic Lojasiewicz conditions that are applicable to any convex loss function, thus extending the current theory to a larger panel of losses commonly used in practice such as cross-entropy. Moreover, our analysis provides high-probability bounds on the approximation error under sub-Gaussian gradient noise and only requires the local smoothness of the objective function, thus making it applicable to deep neural networks in realistic settings.

An Asymptotic Test for Conditional Independence using Analytic Kernel Embeddings

- Meyer Scetbon, Laurent Meunier, Yaniv Romano
- abstract: We propose a new conditional dependence measure and a statistical test for conditional independence. The measure is based on the difference between analytic kernel embeddings of two well-suited distributions evaluated at a finite set of locations. We obtain its asymptotic distribution under the null hypothesis of conditional independence and design a consistent statistical test from it. We conduct a series of experiments showing that our new test outperforms state-of-the-art methods both in terms of type-I and type-II errors even in the high dimensional setting.

Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs

- Meyer Scetbon, Gabriel Peyré, Marco Cuturi
- abstract: The ability to align points across two related yet incomparable point clouds (e.g. living in different spaces) plays an important role in machine learning. The Gromov-Wasserstein (GW) framework provides an increasingly popular answer to such problems, by seeking a low-distortion, geometry-preserving assignment between these points. As a non-convex, quadratic generalization of optimal transport (OT), GW is NP-hard. While practitioners often resort to solving GW approximately as a nested sequence of entropy-regularized OT problems, the cubic complexity (in the number n of samples) of that approach is a roadblock. We show in this work how a recent variant of the OT problem that restricts the set of admissible couplings to those having a low-rank factorization is remarkably well suited to the resolution of GW: when applied to GW, we show that this approach is not only able to compute a stationary point of the GW problem in time $O(n^2)$, but also uniquely positioned to benefit from the knowledge that the initial cost matrices are low-rank, to yield a linear time $O(n)$ GW approximation. Our approach yields similar results, yet orders of magnitude faster computation than the SoTA entropic GW approaches, on both simulated and real data.

Streaming Inference for Infinite Feature Models

- Rylan Schaeffer, Yilun Du, Gabrielle K Liu, Ila Fiete
- abstract: Unsupervised learning from a continuous stream of data is arguably one of the most common and most challenging problems facing intelligent agents. One class of unsupervised models, collectively termed feature models, attempts unsupervised discovery of latent features underlying the data and includes common models such as PCA, ICA, and NMF. However, if the data arrives in a continuous stream, determining the number of features is a significant challenge and the number may grow with time. In this work, we make feature models significantly more applicable to streaming data by imbuing them with the ability to create new features, online, in a probabilistic and principled manner. To achieve this, we derive a novel recursive form of the Indian Buffet Process, which we term the Recursive IBP (R-IBP). We demonstrate that R-IBP can be be used as a prior for feature models to efficiently infer a posterior over an unbounded number of latent features, with quasilinear average time complexity and logarithmic average space complexity. We compare R-IBP to existing offline sampling and variational baselines in two feature models (Linear Gaussian and Factor Analysis) and demonstrate on synthetic and real data that R-IBP achieves comparable or better performance in significantly less time.

Modeling Irregular Time Series with Continuous Recurrent Units

- Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, Maja Rudolph
- abstract: Recurrent neural networks (RNNs) are a popular choice for modeling sequential data. Modern RNN architectures assume constant time-intervals between observations. However, in many datasets (e.g. medical records) observation times are irregular and can carry important information. To address this challenge, we propose continuous recurrent units (CRUs) {–} a neural architecture that can naturally handle irregular intervals between observations. The CRU assumes a hidden state, which evolves according to a linear stochastic differential equation and is integrated into an encoder-decoder framework. The recursive computations of the CRU can be derived using the continuous-discrete Kalman filter and are in closed form. The resulting recurrent architecture has temporal continuity between hidden states and a gating mechanism that can optimally integrate noisy observations. We derive an efficient parameterization scheme for the CRU that leads to a fast implementation f-CRU. We empirically study the CRU on a number of challenging datasets and find that it can interpolate irregular time series better than methods based on neural ordinary differential equations.

Structure Preserving Neural Networks: A Case Study in the Entropy Closure of the Boltzmann Equation

- Steffen Schotthöfer, Tianbai Xiao, Martin Frank, Cory Hauck
- abstract: In this paper, we explore applications of deep learning in statistical physics. We choose the Boltzmann equation as a typical example, where neural networks serve as a closure to its moment system. We present two types of neural networks to embed the convexity of entropy and to preserve the minimum entropy principle and intrinsic mathematical structures of the moment system of the Boltzmann equation. We derive an error bound for the generalization gap of convex neural networks which are trained in Sobolev norm and use the results to construct data sampling methods for neural network training. Numerical experiments demonstrate that the neural entropy closure is significantly faster than classical optimizers while maintaining sufficient accuracy.

Improving Robustness against Real-World and Worst-Case Distribution Shifts through Decision Region Quantification

- Leo Schwinn, Leon Bungert, An Nguyen, René Raab, Falk Pulsmeyer, Doina Precup, Bjoern Eskofier, Dario Zanca

- abstract: The reliability of neural networks is essential for their use in safety-critical applications. Existing approaches generally aim at improving the robustness of neural networks to either real-world distribution shifts (e.g., common corruptions and perturbations, spatial transformations, and natural adversarial examples) or worst-case distribution shifts (e.g., optimized adversarial examples). In this work, we propose the Decision Region Quantification (DRQ) algorithm to improve the robustness of any differentiable pre-trained model against both real-world and worst-case distribution shifts in the data. DRQ analyzes the robustness of local decision regions in the vicinity of a given data point to make more reliable predictions. We theoretically motivate the DRQ algorithm by showing that it effectively smooths spurious local extrema in the decision surface. Furthermore, we propose an implementation using targeted and untargeted adversarial attacks. An extensive empirical evaluation shows that DRQ increases the robustness of adversarially and non-adversarially trained models against real-world and worst-case distribution shifts on several computer vision benchmark datasets.

Symmetric Machine Theory of Mind

- Melanie Sclar, Graham Neubig, Yonatan Bisk
- abstract: Theory of mind, the ability to model others' thoughts and desires, is a cornerstone of human social intelligence. This makes it an important challenge for the machine learning community, but previous works mainly attempt to design agents that model the "mental state" of others as passive observers or in specific predefined roles, such as in speaker-listener scenarios. In contrast, we propose to model machine theory of mind in a more general symmetric scenario. We introduce a multi-agent environment SymmToM where, like in real life, all agents can speak, listen, see other agents, and move freely through the world. Effective strategies to maximize an agent's reward require it to develop a theory of mind. We show that reinforcement learning agents that model the mental states of others achieve significant performance improvements over agents with no such theory of mind model. Importantly, our best agents still fail to achieve performance comparable to agents with access to the gold-standard mental state of other agents, demonstrating that the modeling of theory of mind in multi-agent scenarios is very much an open challenge.

Data-SUITE: Data-centric identification of in-distribution incongruous examples

- Nabeel Seedat, Jonathan Crabbé, Mihaela van der Schaar
- abstract: Systematic quantification of data quality is critical for consistent model performance. Prior works have focused on out-of-distribution data. Instead, we tackle an understudied yet equally important problem of characterizing incongruous regions of in-distribution (ID) data, which may arise from feature space heterogeneity. To this end, we propose a paradigm shift with Data-SUITE: a data-centric AI framework to identify these regions, independent of a task-specific model. Data-SUITE leverages copula modeling, representation learning, and conformal prediction to build feature-wise confidence interval estimators based on a set of training instances. These estimators can be used to evaluate the congruence of test instances with respect to the training set, to answer two practically useful questions: (1) which test instances will be reliably predicted by a model trained with the training instances? and (2) can we identify incongruous regions of the feature space so that data owners understand the data's limitations or guide future data collection? We empirically validate Data-SUITE's performance and coverage guarantees and demonstrate on cross-site medical data, biased data, and data with concept drift, that Data-SUITE best identifies ID regions where a downstream model may be reliable (independent of said model). We also illustrate how these identified regions can provide insights into datasets and highlight their limitations.

Continuous-Time Modeling of Counterfactual Outcomes Using Neural Controlled Differential Equations

- Nabeel Seedat, Fergus Imrie, Alexis Bellot, Zhaozhi Qian, Mihaela van der Schaar
- abstract: Estimating counterfactual outcomes over time has the potential to unlock personalized healthcare by assisting decision-makers to answer "what-if" questions. Existing causal inference approaches typically consider regular, discrete-time intervals between observations and treatment decisions and hence are unable to naturally model irregularly sampled data, which is the common setting in practice. To handle arbitrary observation patterns, we interpret the data as samples from an underlying continuous-time process and propose to model its latent trajectory explicitly using the mathematics of controlled differential equations. This leads to a new approach, the Treatment Effect Neural Controlled Differential Equation (TE-CDE), that allows the potential outcomes to be evaluated at any time point. In addition, adversarial training is used to adjust for time-dependent confounding which is critical in longitudinal settings and is an added challenge not encountered in conventional time series. To assess solutions to this problem, we propose a controllable simulation environment based on a model of tumor growth for a range of scenarios with irregular sampling reflective of a variety of clinical scenarios. TE-CDE consistently outperforms existing approaches in all scenarios with irregular sampling.

Neural Tangent Kernel Beyond the Infinite-Width Limit: Effects of Depth and Initialization

- Mariia Seleznova, Gitta Kutyniok
- abstract: Neural Tangent Kernel (NTK) is widely used to analyze overparametrized neural networks due to the famous result by Jacot et al. (2018): in the infinite-width limit, the NTK is deterministic and constant during training. However, this result cannot explain the behavior of deep networks, since it generally does not hold if depth and width tend to infinity simultaneously. In this paper, we study the NTK of fully-connected ReLU networks with depth comparable to width. We prove that the NTK properties depend significantly on the depth-to-width ratio and the distribution of parameters at initialization. In fact, our results indicate the importance of the three phases in the hyperparameter space identified in Poole et al. (2016): ordered, chaotic and the edge of chaos (EOC). We derive exact expressions for the NTK dispersion in the infinite-depth-and-width limit in all three phases and conclude that the NTK variability grows exponentially with depth at the EOC and in the chaotic phase but not in the ordered phase. We also show that the NTK of deep networks may stay constant during training only in the ordered phase and discuss how the structure of the NTK matrix changes during training.

Reinforcement Learning with Action-Free Pre-Training from Videos

- Younggyo Seo, Kimin Lee, Stephen L James, Pieter Abbeel
- abstract: Recent unsupervised pre-training methods have shown to be effective on language and vision domains by learning useful representations for multiple downstream tasks. In this paper, we investigate if such unsupervised pre-training methods can also be effective for vision-based reinforcement learning (RL). To this end, we introduce a framework that learns representations useful for understanding the dynamics via generative pre-training on videos. Our framework consists of two phases: we pre-train an action-free latent video prediction model, and then utilize the pre-trained representations for efficiently learning action-conditional world models on unseen environments. To incorporate additional action inputs during fine-tuning, we introduce a new architecture that stacks an action-conditional latent prediction model on top of the pre-trained action-free prediction model. Moreover, for better exploration, we propose a video-based intrinsic bonus that leverages pre-trained representations. We demonstrate that our framework significantly improves both final performances and sample-efficiency of vision-based RL in a variety of manipulation and locomotion tasks. Code is available at \url{https://github.com/younggyoseo/apv}.

Efficient Model-based Multi-agent Reinforcement Learning via Optimistic Equilibrium Computation

- Pier Giuseppe Sessa, Maryam Kamgarpour, Andreas Krause
- abstract: We consider model-based multi-agent reinforcement learning, where the environment transition model is unknown and can only be learned via expensive interactions with the environment. We propose H-MARL (Hallucinated Multi-Agent Reinforcement Learning), a novel sample-efficient algorithm that can efficiently balance exploration, i.e., learning about the environment, and exploitation, i.e., achieve good equilibrium performance in the underlying general-sum Markov game. H-MARL builds high-probability confidence intervals around the unknown transition model and sequentially updates them based on newly observed data. Using these, it constructs an optimistic hallucinated game for the agents for which equilibrium policies are computed at each round. We consider general statistical models (e.g., Gaussian processes, deep ensembles, etc.) and policy classes (e.g., deep neural

networks), and theoretically analyze our approach by bounding the agents' dynamic regret. Moreover, we provide a convergence rate to the equilibria of the underlying Markov game. We demonstrate our approach experimentally on an autonomous driving simulation benchmark. H-MARL learns successful equilibrium policies after a few interactions with the environment and can significantly improve the performance compared to non-optimistic exploration methods.

Selective Regression under Fairness Criteria

- Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, Gregory W Wornell
- abstract: Selective regression allows abstention from prediction if the confidence to make an accurate prediction is not sufficient. In general, by allowing a reject option, one expects the performance of a regression model to increase at the cost of reducing coverage (i.e., by predicting on fewer samples). However, as we show, in some cases, the performance of a minority subgroup can decrease while we reduce the coverage, and thus selective regression can magnify disparities between different sensitive subgroups. Motivated by these disparities, we propose new fairness criteria for selective regression requiring the performance of every subgroup to improve with a decrease in coverage. We prove that if a feature representation satisfies the sufficiency criterion or is calibrated for mean and variance, then the proposed fairness criteria is met. Further, we introduce two approaches to mitigate the performance disparity across subgroups: (a) by regularizing an upper bound of conditional mutual information under a Gaussian assumption and (b) by regularizing a contrastive loss for conditional mean and conditional variance prediction. The effectiveness of these approaches is demonstrated on synthetic and real-world datasets.

Utility Theory for Sequential Decision Making

- Mehran Shakerinava, Siamak Ravanbakhsh
- abstract: The von Neumann-Morgenstern (VNM) utility theorem shows that under certain axioms of rationality, decision-making is reduced to maximizing the expectation of some utility function. We extend these axioms to increasingly structured sequential decision making settings and identify the structure of the corresponding utility functions. In particular, we show that memoryless preferences lead to a utility in the form of a per transition reward and multiplicative factor on the future return. This result motivates a generalization of Markov Decision Processes (MDPs) with this structure on the agent's returns, which we call Affine-Reward MDPs. A stronger constraint on preferences is needed to recover the commonly used cumulative sum of scalar rewards in MDPs. A yet stronger constraint simplifies the utility function for goal-seeking agents in the form of a difference in some function of states that we call potential functions. Our necessary and sufficient conditions demystify the reward hypothesis that underlies the design of rational agents in reinforcement learning by adding an axiom to the VNM rationality axioms and motivates new directions for AI research involving sequential decision making.

Translating Robot Skills: Learning Unsupervised Skill Correspondences Across Robots

- Tanmay Shankar, Yixin Lin, Aravind Rajeswaran, Vikash Kumar, Stuart Anderson, Jean Oh
- abstract: In this paper, we explore how we can endow robots with the ability to learn correspondences between their own skills, and those of morphologically different robots in different domains, in an entirely unsupervised manner. We make the insight that different morphological robots use similar task strategies to solve similar tasks. Based on this insight, we frame learning skill correspondences as a problem of matching distributions of sequences of skills across robots. We then present an unsupervised objective that encourages a learnt skill translation model to match these distributions across domains, inspired by recent advances in unsupervised machine translation. Our approach is able to learn semantically meaningful correspondences between skills across multiple robot-robot and human-robot domain pairs despite being completely unsupervised. Further, the learnt correspondences enable the transfer of task strategies across robots and domains. We present dynamic visualizations of our results at https://sites.google.com/view/translatingrobotskills/home.

A State-Distribution Matching Approach to Non-Episodic Reinforcement Learning

- Archit Sharma, Rehaan Ahmad, Chelsea Finn
- abstract: While reinforcement learning (RL) provides a framework for learning through trial and error, translating RL algorithms into the real world has remained challenging. A major hurdle to real-world application arises from the development of algorithms in an episodic setting where the environment is reset after every trial, in contrast with the continual and non-episodic nature of the real-world encountered by embodied agents such as humans and robots. Enabling agents to learn behaviors autonomously in such non-episodic environments requires that the agent to be able to conduct its own trials. Prior works have considered an alternating approach where a forward policy learns to solve the task and the backward policy learns to reset the environment, but what initial state distribution should the backward policy reset the agent to? Assuming access to a few demonstrations, we propose a new method, MEDAL, that trains the backward policy to match the state distribution in the provided demonstrations. This keeps the agent close to the task-relevant states, allowing for a mix of easy and difficult starting states for the forward policy. Our experiments show that MEDAL matches or outperforms prior methods on three sparse-reward continuous control tasks from the EARL benchmark, with 40% gains on the hardest task, while making fewer assumptions than prior works.

Content Addressable Memory Without Catastrophic Forgetting by Heteroassociation with a Fixed Scaffold

- Sugandha Sharma, Sarthak Chandra, Ila Fiete
- abstract: Content-addressable memory (CAM) networks, so-called because stored items can be recalled by partial or corrupted versions of the items, exhibit near-perfect recall of a small number of information-dense patterns below capacity and a 'memory cliff' beyond, such that inserting a single additional pattern results in catastrophic loss of all stored patterns. We propose a novel CAM architecture, Memory Scaffold with Heteroassociation (MESH), that factorizes the problems of internal attractor dynamics and association with external content to generate a CAM continuum without a memory cliff: Small numbers of patterns are stored with complete information recovery matching standard CAMs, while inserting more patterns still results in partial recall of every pattern, with a graceful trade-off between pattern number and pattern richness. Motivated by the architecture of the Entorhinal-Hippocampal memory circuit in the brain, MESH is a tripartite architecture with pairwise interactions that uses a predetermined set of internally stabilized states together with heteroassociation between the internal states and arbitrary external patterns. We show analytically and experimentally that for any number of stored patterns, MESH nearly saturates the total information bound (given by the number of synapses) for CAM networks, outperforming all existing CAM models.

Federated Minimax Optimization: Improved Convergence Analyses and Algorithms

- Pranay Sharma, Rohan Panda, Gauri Joshi, Pramod Varshney
- abstract: In this paper, we consider nonconvex minimax optimization, which is gaining prominence in many modern machine learning applications, such as GANs. Large-scale edge-based collection of training data in these applications calls for communication-efficient distributed optimization algorithms, such as those used in federated learning, to process the data. In this paper, we analyze local stochastic gradient descent ascent (SGDA), the local-update version of the SGDA algorithm. SGDA is the core algorithm used in minimax optimization, but it is not well-understood in a distributed setting. We prove that Local SGDA has order-optimal sample complexity for several classes of nonconvex-concave and nonconvex-nonconcave minimax problems, and also enjoys linear speedup with respect to the number of clients. We provide a novel and tighter analysis, which improves the convergence and communication guarantees in the existing literature. For nonconvex-PL and nonconvex-one-point-concave functions, we improve the existing complexity

results for centralized minimax problems. Furthermore, we propose a momentum-based local-update algorithm, which has the same convergence guarantees, but outperforms Local SGDA as demonstrated in our experiments.

DNS: Determinantal Point Process Based Neural Network Sampler for Ensemble Reinforcement Learning

- Hassam Sheikh, Kizza Frisbee, Mariano Phielipp
- abstract: The application of an ensemble of neural networks is becoming an imminent tool for advancing state-of-the-art deep reinforcement learning algorithms. However, training these large numbers of neural networks in the ensemble has an exceedingly high computation cost which may become a hindrance in training large-scale systems. In this paper, we propose DNS: a Determinantal Point Process based Neural Network Sampler that specifically uses k-DPP to sample a subset of neural networks for backpropagation at every training step thus significantly reducing the training time and computation cost. We integrated DNS in REDQ for continuous control tasks and evaluated on MuJoCo environments. Our experiments show that DNS augmented REDQ matches the baseline REDQ in terms of average cumulative reward and achieves this using less than 50% computation when measured in FLOPS. The code is available at https://github.com/IntelLabs/DNS

Instance Dependent Regret Analysis of Kernelized Bandits

- Shubhanshu Shekhar, Tara Javidi
- abstract: We study the problem of designing an adaptive strategy for querying a noisy zeroth-order-oracle to efficiently learn about the optimizer of an unknown function f. To make the problem tractable, we assume that f lies in the reproducing kernel Hilbert space (RKHS) associated with a known kernel K, with its norm bounded by $M<\infty$. Prior results, working in a minimax framework, have characterized the worst-case (over all functions in the problem class) limits on regret achievable by any algorithm, and have constructed algorithms with matching (modulo polylogarithmic factors) worst-case performance for the Matern family of kernels. These results suffer from two drawbacks. First, the minimax lower bound gives limited information about the limits of regret achievable by commonly used algorithms on a specific problem instance f. Second, the existing upper bound analysis fails to adapt to easier problem instances within the function class. Our work takes steps to address both these issues. First, we derive instance-dependent regret lower bounds for algorithms with uniformly (over the function class) vanishing normalized cumulative regret. Our result, valid for several practically relevant kernelized bandits algorithms, such as, GP-UCB, GP-TS and SupKernelUCB, identifies a fundamental complexity measure associated with every problem instance. We then address the second issue, by proposing a new minimax near-optimal algorithm that also adapts to easier problem instances.

Data Augmentation as Feature Manipulation

- Ruoqi Shen, Sebastien Bubeck, Suriya Gunasekar
- abstract: Data augmentation is a cornerstone of the machine learning pipeline, yet its theoretical underpinnings remain unclear. Is it merely a way to artificially augment the data set size? Or is it about encouraging the model to satisfy certain invariances? In this work we consider another angle, and we study the effect of data augmentation on the dynamic of the learning process. We find that data augmentation can alter the relative importance of various features, effectively making certain informative but hard to learn features more likely to be captured in the learning process. Importantly, we show that this effect is more pronounced for non-linear models, such as neural networks. Our main contribution is a detailed analysis of data augmentation on the learning dynamic for a two layer convolutional neural network in the recently proposed multi-view model by Z. Allen-Zhu and Y. Li. We complement this analysis with further experimental evidence that data augmentation can be viewed as a form of feature manipulation.

Metric-Fair Active Learning

- Jie Shen, Nan Cui, Jing Wang
- abstract: Active learning has become a prevalent technique for designing label-efficient algorithms, where the central principle is to only query and fit “informative” labeled instances. It is, however, known that an active learning algorithm may incur unfairness due to such instance selection procedure. In this paper, we henceforth study metric-fair active learning of homogeneous halfspaces, and show that under the distribution-dependent PAC learning model, fairness and label efficiency can be achieved simultaneously. We further propose two extensions of our main results: 1) we show that it is possible to make the algorithm robust to the adversarial noise – one of the most challenging noise models in learning theory; and 2) it is possible to significantly improve the label complexity when the underlying halfspace is sparse.

PDO-s3DCNNs: Partial Differential Operator Based Steerable 3D CNNs

- Zhengyang Shen, Tao Hong, Qi She, Jinwen Ma, Zhouchen Lin
- abstract: Steerable models can provide very general and flexible equivariance by formulating equivariance requirements in the language of representation theory and feature fields, which has been recognized to be effective for many vision tasks. However, deriving steerable models for 3D rotations is much more difficult than that in the 2D case, due to more complicated mathematics of 3D rotations. In this work, we employ partial differential operators (PDOs) to model 3D filters, and derive general steerable 3D CNNs, which are called PDO-s3DCNNs. We prove that the equivariant filters are subject to linear constraints, which can be solved efficiently under various conditions. As far as we know, PDO-s3DCNNs are the most general steerable CNNs for 3D rotations, in the sense that they cover all common subgroups of SO(3) and their representations, while existing methods can only be applied to specific groups and representations. Extensive experiments show that our models can preserve equivariance well in the discrete domain, and outperform previous works on SHREC’17 retrieval and ISBI 2012 segmentation tasks with a low network complexity.

Connect, Not Collapse: Explaining Contrastive Learning for Unsupervised Domain Adaptation

- Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z. Haochen, Tengyu Ma, Percy Liang
- abstract: We consider unsupervised domain adaptation (UDA), where labeled data from a source domain (e.g., photos) and unlabeled data from a target domain (e.g., sketches) are used to learn a classifier for the target domain. Conventional UDA methods (e.g., domain adversarial training) learn domain-invariant features to generalize from the source domain to the target domain. In this paper, we show that contrastive pre-training, which learns features on unlabeled source and target data and then fine-tunes on labeled source data, is competitive with strong UDA methods. However, we find that contrastive pre-training does not learn domain-invariant features, diverging from conventional UDA intuitions. We show theoretically that contrastive pre-training can learn features that vary subtantially across domains but still generalize to the target domain, by disentangling domain and class information. We empirically validate our theory on benchmark vision datasets.

Constrained Optimization with Dynamic Bound-scaling for Effective NLP Backdoor Defense

- Guangyu Shen, Yingqi Liu, Guanhong Tao, Qiuling Xu, Zhuo Zhang, Shengwei An, Shiqing Ma, Xiangyu Zhang
- abstract: Modern language models are vulnerable to backdoor attacks. An injected malicious token sequence (i.e., a trigger) can cause the compromised model to misbehave, raising security concerns. Trigger inversion is a widely-used technique for scanning backdoors in vision models. It can- not be directly applied to NLP models due to their discrete nature. In this paper, we develop a novel optimization method for NLP backdoor inversion. We leverage a dynamically reducing temperature coefficient in the softmax function to provide changing loss landscapes to the optimizer such that the process gradually focuses on the ground truth trigger, which is denoted as a one-hot value in a convex hull. Our method also features a temperature rollback

mechanism to step away from local optimals, exploiting the observation that local optimals can be easily determined in NLP trigger inversion (while not in general optimization). We evaluate the technique on over 1600 models (with roughly half of them having injected backdoors) on 3 prevailing NLP tasks, with 4 different backdoor attacks and 7 architectures. Our results show that the technique is able to effectively and efficiently detect and remove backdoors, outperforming 5 baseline methods. The code is available at https: //github.com/PurduePAML/DBS.

Staged Training for Transformer Language Models

- Sheng Shen, Pete Walsh, Kurt Keutzer, Jesse Dodge, Matthew Peters, Iz Beltagy
- abstract: The current standard approach to scaling transformer language models trains each model size from a different random initialization. As an alternative, we consider a staged training setup that begins with a small model and incrementally increases the amount of compute used for training by applying a "growth operator" to increase the model depth and width. By initializing each stage with the output of the previous one, the training process effectively re-uses the compute from prior stages and becomes more efficient. Our growth operators each take as input the entire training state (including model parameters, optimizer state, learning rate schedule, etc.) and output a new training state from which training continues. We identify two important properties of these growth operators, namely that they preserve both the loss and the “training dynamics” after applying the operator. While the loss-preserving property has been discussed previously, to the best of our knowledge this work is the first to identify the importance of preserving the training dynamics (the rate of decrease of the loss during training). To find the optimal schedule for stages, we use the scaling laws from (Kaplan et al., 2020) to find a precise schedule that gives the most compute saving by starting a new stage when training efficiency starts decreasing. We empirically validate our growth operators and staged training for autoregressive language models, showing up to 22% compute savings compared to a strong baseline trained from scratch. Our code is available at https://github.com/allenai/staged-training.

Deep Network Approximation in Terms of Intrinsic Parameters

- Zuowei Shen, Haizhao Yang, Shijun Zhang
- abstract: One of the arguments to explain the success of deep learning is the powerful approximation capacity of deep neural networks. Such capacity is generally accompanied by the explosive growth of the number of parameters, which, in turn, leads to high computational costs. It is of great interest to ask whether we can achieve successful deep learning with a small number of learnable parameters adapting to the target function. From an approximation perspective, this paper shows that the number of parameters that need to be learned can be significantly smaller than people typically expect. First, we theoretically design ReLU networks with a few learnable parameters to achieve an attractive approximation. We prove by construction that, for any Lipschitz continuous function f on $[0,1]^d$ with a Lipschitz constant $\lambda>0$, a ReLU network with $n+2$ intrinsic parameters (those depending on f) can approximate f with an exponentially small error $5 \lambda \sqrt{d} \, 2^{-n}$. Such a result is generalized to generic continuous functions. Furthermore, we show that the idea of learning a small number of parameters to achieve a good approximation can be numerically observed. We conduct several experiments to verify that training a small part of parameters can also achieve good results for classification problems if other parameters are pre-specified or pre-trained from a related problem.

Gradient-Free Method for Heavily Constrained Nonconvex Optimization

- Wanli Shi, Hongchang Gao, Bin Gu
- abstract: Zeroth-order (ZO) method has been shown to be a powerful method for solving the optimization problem where explicit expression of the gradients is difficult or infeasible to obtain. Recently, due to the practical value of the constrained problems, a lot of ZO Frank-Wolfe or projected ZO methods have been proposed. However, in many applications, we may have a very large number of nonconvex white/black-box constraints, which makes the existing zeroth-order methods extremely inefficient (or even not working) since they need to inquire function value of all the constraints and project the solution to the complicated feasible set. In this paper, to solve the nonconvex problem with a large number of white/black-box constraints, we proposed a doubly stochastic zeroth-order gradient method (DSZOG) with momentum method and adaptive step size. Theoretically, we prove DSZOG can converge to the ϵ-stationary point of the constrained problem. Experimental results in two applications demonstrate the superiority of our method in terms of training time and accuracy compared with other ZO methods for the constrained problem.

Global Optimization of K-Center Clustering

- Mingfei Shi, Kaixun Hua, Jiayang Ren, Yankai Cao
- abstract: k-center problem is a well-known clustering method and can be formulated as a mixed-integer nonlinear programming problem. This work provides a practical global optimization algorithm for this task based on a reduced-space spatial branch and bound scheme. This algorithm can guarantee convergence to the global optimum by only branching on the centers of clusters, which is independent of the dataset’s cardinality. In addition, a set of feasibility-based bounds tightening techniques are proposed to narrow down the domain of centers and significantly accelerate the convergence. To demonstrate the capacity of this algorithm, we present computational results on 32 datasets. Notably, for the dataset with 14 million samples and 3 features, the serial implementation of the algorithm can converge to an optimality gap of 0.1% within 2 hours. Compared with a heuristic method, the global optimum obtained by our algorithm can reduce the objective function on average by 30.4%.

Pessimistic Q-Learning for Offline Reinforcement Learning: Towards Optimal Sample Complexity

- Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Yuejie Chi
- abstract: Offline or batch reinforcement learning seeks to learn a near-optimal policy using history data without active exploration of the environment. To counter the insufficient coverage and sample scarcity of many offline datasets, the principle of pessimism has been recently introduced to mitigate high bias of the estimated values. While pessimistic variants of model-based algorithms (e.g., value iteration with lower confidence bounds) have been theoretically investigated, their model-free counterparts — which do not require explicit model estimation — have not been adequately studied, especially in terms of sample efficiency. To address this inadequacy, we study a pessimistic variant of Q-learning in the context of finite-horizon Markov decision processes, and characterize its sample complexity under the single-policy concentrability assumption which does not require the full coverage of the state-action space. In addition, a variance-reduced pessimistic Q-learning algorithm is proposed to achieve near-optimal sample complexity. Altogether, this work highlights the efficiency of model-free algorithms in offline RL when used in conjunction with pessimism and variance reduction.

Adversarial Masking for Self-Supervised Learning

- Yuge Shi, N Siddharth, Philip Torr, Adam R Kosiorek
- abstract: We propose ADIOS, a masked image model (MIM) framework for self-supervised learning, which simultaneously learns a masking function and an image encoder using an adversarial objective. The image encoder is trained to minimise the distance between representations of the original and that of a masked image. The masking function, conversely, aims at maximising this distance. ADIOS consistently improves on state-of-the-art self-supervised learning (SSL) methods on a variety of tasks and datasets—including classification on ImageNet100 and STL10, transfer learning on CIFAR10/100, Flowers102 and iNaturalist, as well as robustness evaluated on the backgrounds challenge (Xiao et al., 2021)—while generating semantically meaningful masks. Unlike modern MIM models such as MAE, BEiT and iBOT, ADIOS does not rely on the image-patch tokenisation construction of Vision Transformers, and can be implemented with convolutional backbones. We further demonstrate that the masks learned by ADIOS are more effective in improving representation learning of SSL methods than masking schemes used in popular MIM models.

Visual Attention Emerges from Recurrent Sparse Reconstruction

- Baifeng Shi, Yale Song, Neel Joshi, Trevor Darrell, Xin Wang
- abstract: Visual attention helps achieve robust perception under noise, corruption, and distribution shifts in human vision, which are areas where modern neural networks still fall short. We present VARS, Visual Attention from Recurrent Sparse reconstruction, a new attention formulation built on two prominent features of the human visual attention mechanism: recurrency and sparsity. Related features are grouped together via recurrent connections between neurons, with salient objects emerging via sparse regularization. VARS adopts an attractor network with recurrent connections that converges toward a stable pattern over time. Network layers are represented as ordinary differential equations (ODEs), formulating attention as a recurrent attractor network that equivalently optimizes the sparse reconstruction of input using a dictionary of "templates" encoding underlying patterns of data. We show that self-attention is a special case of VARS with a single-step optimization and no sparsity constraint. VARS can be readily used as a replacement for self-attention in popular vision transformers, consistently improving their robustness across various benchmarks.

A Minimax Learning Approach to Off-Policy Evaluation in Confounded Partially Observable Markov Decision Processes

- Chengchun Shi, Masatoshi Uehara, Jiawei Huang, Nan Jiang
- abstract: We consider off-policy evaluation (OPE) in Partially Observable Markov Decision Processes (POMDPs), where the evaluation policy depends only on observable variables and the behavior policy depends on unobservable latent variables. Existing works either assume no unmeasured confounders, or focus on settings where both the observation and the state spaces are tabular. In this work, we first propose novel identification methods for OPE in POMDPs with latent confounders, by introducing bridge functions that link the target policy's value and the observed data distribution. We next propose minimax estimation methods for learning these bridge functions, and construct three estimators based on these estimated bridge functions, corresponding to a value function-based estimator, a marginalized importance sampling estimator, and a doubly-robust estimator. Our proposal permits general function approximation and is thus applicable to settings with continuous or large observation/state spaces. The nonasymptotic and asymptotic properties of the proposed estimators are investigated in detail. A Python implementation of our proposal is available at https://github.com/jiaweihhuang/ Confounded-POMDP-Exp.

Robust Group Synchronization via Quadratic Programming

- Yunpeng Shi, Cole M Wyeth, Gilad Lerman
- abstract: We propose a novel quadratic programming formulation for estimating the corruption levels in group synchronization, and use these estimates to solve this problem. Our objective function exploits the cycle consistency of the group and we thus refer to our method as detection and estimation of structural consistency (DESC). This general framework can be extended to other algebraic and geometric structures. Our formulation has the following advantages: it can tolerate corruption as high as the information-theoretic bound, it does not require a good initialization for the estimates of group elements, it has a simple interpretation, and under some mild conditions the global minimum of our objective function exactly recovers the corruption levels. We demonstrate the competitive accuracy of our approach on both synthetic and real data experiments of rotation averaging.

Log-Euclidean Signatures for Intrinsic Distances Between Unaligned Datasets

- Tal Shnitzer, Mikhail Yurochkin, Kristjan Greenewald, Justin M Solomon
- abstract: The need for efficiently comparing and representing datasets with unknown alignment spans various fields, from model analysis and comparison in machine learning to trend discovery in collections of medical datasets. We use manifold learning to compare the intrinsic geometric structures of different datasets by comparing their diffusion operators, symmetric positive-definite (SPD) matrices that relate to approximations of the continuous Laplace-Beltrami operator from discrete samples. Existing methods typically assume known data alignment and compare such operators in a pointwise manner. Instead, we exploit the Riemannian geometry of SPD matrices to compare these operators and define a new theoretically-motivated distance based on a lower bound of the log-Euclidean metric. Our framework facilitates comparison of data manifolds expressed in datasets with different sizes, numbers of features, and measurement modalities. Our log-Euclidean signature (LES) distance recovers meaningful structural differences, outperforming competing methods in various application domains.

Scalable Computation of Causal Bounds

- Madhumitha Shridharan, Garud Iyengar
- abstract: We consider the problem of computing bounds for causal inference problems with unobserved confounders, where identifiability does not hold. Existing non-parametric approaches for computing such bounds use linear programming (LP) formulations that quickly become intractable for existing solvers because the size of the LP grows exponentially in the number of edges in the underlying causal graph. We show that this LP can be significantly pruned by carefully considering the structure of the causal query, allowing us to compute bounds for significantly larger causal inference problems as compared to what is possible using existing techniques. This pruning procedure also allows us to compute the bounds in closed form for a special class of causal graphs and queries, which includes a well-studied family of problems where multiple confounded treatments influence an outcome. We also propose a very efficient greedy heuristic that produces very high quality bounds, and scales to problems that are several orders of magnitude larger than those for which the pruned LP can be solved.

Bit Prioritization in Variational Autoencoders via Progressive Coding

- Rui Shu, Stefano Ermon
- abstract: The hierarchical variational autoencoder (HVAE) is a popular generative model used for many representation learning tasks. However, its application to image synthesis often yields models with poor sample quality. In this work, we treat image synthesis itself as a hierarchical representation learning problem and regularize an HVAE toward representations that improve the model's image synthesis performance. We do so by leveraging the progressive coding hypothesis, which claims hierarchical latent variable models that are good at progressive lossy compression will generate high-quality samples. To test this hypothesis, we first show empirically that conventionally-trained HVAEs are not good progressive coders. We then propose a simple method that constrains the hierarchical representations to prioritize the encoding of information beneficial for lossy compression, and show that this modification leads to improved sample quality. Our work lends further support to the progressive coding hypothesis and demonstrates that this hypothesis should be exploited when designing variational autoencoders.

Fair Representation Learning through Implicit Path Alignment

- Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, Christian Gagné
- abstract: We consider a fair representation learning perspective, where optimal predictors, on top of the data representation, are ensured to be invariant with respect to different sub-groups. Specifically, we formulate this intuition as a bi-level optimization, where the representation is learned in the outer-loop, and invariant optimal group predictors are updated in the inner-loop. Moreover, the proposed bi-level objective is demonstrated to fulfill the sufficiency rule, which is desirable in various practical scenarios but was not commonly studied in the fair learning. Besides, to avoid the high computational and memory cost of differentiating in the inner-loop of bi-level objective, we propose an implicit path alignment algorithm, which only relies on the solution of inner optimization and the implicit differentiation rather than the exact optimization path. We further analyze the error gap of the

implicit approach and empirically validate the proposed method in both classification and regression settings. Experimental results show the consistently better trade-off in prediction performance and fairness measurement.

Faster Algorithms for Learning Convex Functions

- Ali Siahkamari, Durmus Alp Emre Acar, Christopher Liao, Kelly L Geyer, Venkatesh Saligrama, Brian Kulis
- abstract: The task of approximating an arbitrary convex function arises in several learning problems such as convex regression, learning with a difference of convex (DC) functions, and learning Bregman or f-divergences. In this paper, we develop and analyze an approach for solving a broad range of convex function learning problems that is faster than state-of-the-art approaches. Our approach is based on a 2-block ADMM method where each block can be computed in closed form. For the task of convex Lipschitz regression, we establish that our proposed algorithm converges with iteration complexity of $ O(n\sqrt{d}/\epsilon)$ for a dataset $\bm X \in \mathbb R^{n\times d}$ and $\epsilon > 0$. Combined with per-iteration computation complexity, our method converges with the rate $O(n^3 d^{1.5}/\epsilon+n^2 d^{2.5}/\epsilon+n d^3/\epsilon)$. This new rate improves the state of the art rate of $O(n^5d^2/\epsilon)$ if $d = o(n^4)$. Further we provide similar solvers for DC regression and Bregman divergence learning. Unlike previous approaches, our method is amenable to the use of GPUs. We demonstrate on regression and metric learning experiments that our approach is over 100 times faster than existing approaches on some data sets, and produces results that are comparable to state of the art.

Coin Flipping Neural Networks

- Yuval Sieradzki, Nitzan Hodos, Gal Yehuda, Assaf Schuster
- abstract: We show that neural networks with access to randomness can outperform deterministic networks by using amplification. We call such networks Coin-Flipping Neural Networks, or CFNNs. We show that a CFNN can approximate the indicator of a d-dimensional ball to arbitrary accuracy with only 2 layers and O(1) neurons, where a 2-layer deterministic network was shown to require Omega(e^d) neurons, an exponential improvement. We prove a highly non-trivial result, that for almost any classification problem, there exists a trivially simple network that solves it given a sufficiently powerful generator for the network's weights. Combining these results we conjecture that for most classification problems, there is a CFNN which solves them with higher accuracy or fewer neurons than any deterministic network. Finally, we verify our proofs experimentally using novel CFNN architectures on CIFAR10 and CIFAR100, reaching an improvement of 9.25% from the baseline.

Reverse Engineering the Neural Tangent Kernel

- James Benjamin Simon, Sajant Anand, Mike Deweese
- abstract: The development of methods to guide the design of neural networks is an important open challenge for deep learning theory. As a paradigm for principled neural architecture design, we propose the translation of high-performing kernels, which are better-understood and amenable to first-principles design, into equivalent network architectures, which have superior efficiency, flexibility, and feature learning. To this end, we constructively prove that, with just an appropriate choice of activation function, any positive-semidefinite dot-product kernel can be realized as either the NNGP or neural tangent kernel of a fully-connected neural network with only one hidden layer. We verify our construction numerically and demonstrate its utility as a design tool for finite fully-connected networks in several experiments.

Demystifying the Adversarial Robustness of Random Transformation Defenses

- Chawin Sitawarin, Zachary J Golan-Strieb, David Wagner
- abstract: Neural networks' lack of robustness against attacks raises concerns in security-sensitive settings such as autonomous vehicles. While many countermeasures may look promising, only a few withstand rigorous evaluation. Defenses using random transformations (RT) have shown impressive results, particularly BaRT (Raff et al., 2019) on ImageNet. However, this type of defense has not been rigorously evaluated, leaving its robustness properties poorly understood. Their stochastic properties make evaluation more challenging and render many proposed attacks on deterministic models inapplicable. First, we show that the BPDA attack (Athalye et al., 2018a) used in BaRT's evaluation is ineffective and likely overestimates its robustness. We then attempt to construct the strongest possible RT defense through the informed selection of transformations and Bayesian optimization for tuning their parameters. Furthermore, we create the strongest possible attack to evaluate our RT defense. Our new attack vastly outperforms the baseline, reducing the accuracy by 83% compared to the 19% reduction by the commonly used EoT attack ($4.3\times$ improvement). Our result indicates that the RT defense on the Imagenette dataset (a ten-class subset of ImageNet) is not robust against adversarial examples. Extending the study further, we use our new attack to adversarially train RT defense (called AdvRT), resulting in a large robustness gain. Code is available at https://github.com/wagnergroup/demystify-random-transform.

Smoothed Adversarial Linear Contextual Bandits with Knapsacks

- Vidyashankar Sivakumar, Shiliang Zuo, Arindam Banerjee
- abstract: Many bandit problems are characterized by the learner making decisions under constraints. The learner in Linear Contextual Bandits with Knapsacks (LinCBwK) receives a resource consumption vector in addition to a scalar reward in each time step which are both linear functions of the context corresponding to the chosen arm. For a fixed time horizon T, the goal of the learner is to maximize rewards while ensuring resource consumptions do not exceed a pre-specified budget. We present algorithms and characterize regret for LinCBwK in the smoothed setting where base context vectors are assumed to be perturbed by Gaussian noise. We consider both the stochastic and adversarial settings for the base contexts, and our analysis of stochastic LinCBwK can be viewed as a warm-up to the more challenging adversarial LinCBwK. For the stochastic setting, we obtain $O(\sqrt{T})$ additive regret bounds compared to the best context dependent fixed policy. The analysis combines ideas for greedy parameter estimation in \cite{kmrw18, siwb20} and the primal-dual paradigm first explored in \cite{agde17, agde14}. Our main contribution is an algorithm with $O(\log T)$ competitive ratio relative to the best context dependent fixed policy for the adversarial setting. The algorithm for the adversarial setting employs ideas from the primal-dual framework \cite{agde17, agde14} and a novel adaptation of the doubling trick \cite{isss19}.

GenLabel: Mixup Relabeling using Generative Models

- Jy-Yong Sohn, Liang Shang, Hongxu Chen, Jaekyun Moon, Dimitris Papailiopoulos, Kangwook Lee
- abstract: Mixup is a data augmentation method that generates new data points by mixing a pair of input data. While mixup generally improves the prediction performance, it sometimes degrades the performance. In this paper, we first identify the main causes of this phenomenon by theoretically and empirically analyzing the mixup algorithm. To resolve this, we propose GenLabel, a simple yet effective relabeling algorithm designed for mixup. In particular, GenLabel helps the mixup algorithm correctly label mixup samples by learning the class-conditional data distribution using generative models. Via theoretical and empirical analysis, we show that mixup, when used together with GenLabel, can effectively resolve the aforementioned phenomenon, improving the accuracy of mixup-trained model.

Communicating via Markov Decision Processes

- Samuel Sokota, Christian A Schroeder De Witt, Maximilian Igl, Luisa M Zintgraf, Philip Torr, Martin Strohmeier, Zico Kolter, Shimon Whiteson, Jakob Foerster

- abstract: We consider the problem of communicating exogenous information by means of Markov decision process trajectories. This setting, which we call a Markov coding game (MCG), generalizes both source coding and a large class of referential games. MCGs also isolate a problem that is important in decentralized control settings in which cheap-talk is not available—namely, they require balancing communication with the associated cost of communicating. We contribute a theoretically grounded approach to MCGs based on maximum entropy reinforcement learning and minimum entropy coupling that we call MEME. Due to recent breakthroughs in approximation algorithms for minimum entropy coupling, MEME is not merely a theoretical algorithm, but can be applied to practical settings. Empirically, we show both that MEME is able to outperform a strong baseline on small MCGs and that MEME is able to achieve strong performance on extremely large MCGs. To the latter point, we demonstrate that MEME is able to losslessly communicate binary images via trajectories of Cartpole and Pong, while simultaneously achieving the maximal or near maximal expected returns, and that it is even capable of performing well in the presence of actuator noise.

The Multivariate Community Hawkes Model for Dependent Relational Events in Continuous-time Networks

- Hadeel Soliman, Lingfei Zhao, Zhipeng Huang, Subhadeep Paul, Kevin S Xu
- abstract: The stochastic block model (SBM) is one of the most widely used generative models for network data. Many continuous-time dynamic network models are built upon the same assumption as the SBM: edges or events between all pairs of nodes are conditionally independent given the block or community memberships, which prevents them from reproducing higher-order motifs such as triangles that are commonly observed in real networks. We propose the multivariate community Hawkes (MULCH) model, an extremely flexible community-based model for continuous-time networks that introduces dependence between node pairs using structured multivariate Hawkes processes. We fit the model using a spectral clustering and likelihood-based local refinement procedure. We find that our proposed MULCH model is far more accurate than existing models both for predictive and generative tasks.

Disentangling Sources of Risk for Distributional Multi-Agent Reinforcement Learning

- Kyunghwan Son, Junsu Kim, Sungsoo Ahn, Roben D Delos Reyes, Yung Yi, Jinwoo Shin
- abstract: In cooperative multi-agent reinforcement learning, the outcomes of agent-wise policies are highly stochastic due to the two sources of risk: (a) random actions taken by teammates and (b) random transition and rewards. Although the two sources have very distinct characteristics, existing frameworks are insufficient to control the risk-sensitivity of agent-wise policies in a disentangled manner. To this end, we propose Disentangled RIsk-sensitive Multi-Agent reinforcement learning (DRIMA) to separately access the risk sources. For example, our framework allows an agent to be optimistic with respect to teammates (who can prosocially adapt) but more risk-neutral with respect to the environment (which does not adapt). Our experiments demonstrate that DRIMA significantly outperforms prior state-of-the-art methods across various scenarios in the StarCraft Multi-agent Challenge environment. Notably, DRIMA shows robust performance where prior methods learn only a highly suboptimal policy, regardless of reward shaping, exploration scheduling, and noisy (random or adversarial) agents.

TAM: Topology-Aware Margin Loss for Class-Imbalanced Node Classification

- Jaeyun Song, Joonhyung Park, Eunho Yang
- abstract: Learning unbiased node representations under class-imbalanced graph data is challenging due to interactions between adjacent nodes. Existing studies have in common that they compensate the minor class nodes ‘as a group’ according to their overall quantity (ignoring node connections in graph), which inevitably increase the false positive cases for major nodes. We hypothesize that the increase in these false positive cases is highly affected by the label distribution around each node and confirm it experimentally. In addition, in order to handle this issue, we propose Topology-Aware Margin (TAM) to reflect local topology on the learning objective. Our method compares the connectivity pattern of each node with the class-averaged counter-part and adaptively adjusts the margin accordingly based on that. Our method consistently exhibits superiority over the baselines on various node classification benchmark datasets with representative GNN architectures.

A General Recipe for Likelihood-free Bayesian Optimization

- Jiaming Song, Lantao Yu, Willie Neiswanger, Stefano Ermon
- abstract: The acquisition function, a critical component in Bayesian optimization (BO), can often be written as the expectation of a utility function under a surrogate model. However, to ensure that acquisition functions are tractable to optimize, restrictions must be placed on the surrogate model and utility function. To extend BO to a broader class of models and utilities, we propose likelihood-free BO (LFBO), an approach based on likelihood-free inference. LFBO directly models the acquisition function without having to separately perform inference with a probabilistic surrogate model. We show that computing the acquisition function in LFBO can be reduced to optimizing a weighted classification problem, which extends an existing likelihood-free density ratio estimation method related to probability of improvement (PI). By choosing the utility function for expected improvement (EI), LFBO outperforms the aforementioned method, as well as various state-of-the-art black-box optimization methods on several real-world optimization problems. LFBO can also leverage composite structures of the objective function, which further improves its regret by several orders of magnitude.

Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis

- Sho Sonoda, Isao Ishikawa, Masahiro Ikeda
- abstract: Neural network on Riemannian symmetric space such as hyperbolic space and the manifold of symmetric positive definite (SPD) matrices is an emerging subject of research in geometric deep learning. Based on the well-established framework of the Helgason-Fourier transform on the noncompact symmetric space, we present a fully-connected network and its associated ridgelet transform on the noncompact symmetric space, covering the hyperbolic neural network (HNN) and the SPDNet as special cases. The ridgelet transform is an analysis operator of a depth-2 continuous network spanned by neurons, namely, it maps an arbitrary given function to the weights of a network. Thanks to the coordinate-free reformulation, the role of nonlinear activation functions is revealed to be a wavelet function. Moreover, the reconstruction formula is applied to present a constructive proof of the universality of finite networks on symmetric spaces.

Saute RL: Almost Surely Safe Reinforcement Learning Using State Augmentation

- Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, Haitham Ammar
- abstract: Satisfying safety constraints almost surely (or with probability one) can be critical for the deployment of Reinforcement Learning (RL) in real-life applications. For example, plane landing and take-off should ideally occur with probability one. We address the problem by introducing Safety Augmented (Saute) Markov Decision Processes (MDPs), where the safety constraints are eliminated by augmenting them into the state-space and reshaping the objective. We show that Saute MDP satisfies the Bellman equation and moves us closer to solving Safe RL with constraints satisfied almost surely. We argue that Saute MDP allows viewing the Safe RL problem from a different perspective enabling new features. For instance, our approach has a plug-and-play nature, i.e., any RL algorithm can be "Sauteed”. Additionally, state augmentation allows for policy generalization across safety constraints. We finally show that Saute RL algorithms can outperform their state-of-the-art counterparts when constraint satisfaction is of high importance.

Lightweight Projective Derivative Codes for Compressed Asynchronous Gradient Descent

- Pedro J Soto, Ilia Ilmer, Haibin Guan, Jun Li

- abstract: Coded distributed computation has become common practice for performing gradient descent on large datasets to mitigate stragglers and other faults. This paper proposes a novel algorithm that encodes the partial derivatives themselves and furthermore optimizes the codes by performing lossy compression on the derivative codewords by maximizing the information contained in the codewords while minimizing the information between the codewords. The utility of this application of coding theory is a geometrical consequence of the observed fact in optimization research that noise is tolerable, sometimes even helpful, in gradient descent based learning algorithms since it helps avoid overfitting and local minima. This stands in contrast with much current conventional work on distributed coded computation which focuses on recovering all of the data from the workers. A second further contribution is that the low-weight nature of the coding scheme allows for asynchronous gradient updates since the code can be iteratively decoded; i.e., a worker's task can immediately be updated into the larger gradient. The directional derivative is always a linear function of the direction vectors; thus, our framework is robust since it can apply linear coding techniques to general machine learning frameworks such as deep neural networks.

Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders

- Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, Andrew Gordon Wilson
- abstract: Bayesian optimization (BayesOpt) is a gold standard for query-efficient continuous optimization. However, its adoption for drug design has been hindered by the discrete, high-dimensional nature of the decision variables. We develop a new approach (LaMBO) which jointly trains a denoising autoencoder with a discriminative multi-task Gaussian process head, allowing gradient-based optimization of multi-objective acquisition functions in the latent space of the autoencoder. These acquisition functions allow LaMBO to balance the explore-exploit tradeoff over multiple design rounds, and to balance objective tradeoffs by optimizing sequences at many different points on the Pareto frontier. We evaluate LaMBO on two small-molecule design tasks, and introduce new tasks optimizing in silico and in vitro properties of large-molecule fluorescent proteins. In our experiments LaMBO outperforms genetic optimizers and does not require a large pretraining corpus, demonstrating that BayesOpt is practical and effective for biological sequence design.

3D Infomax improves GNNs for Molecular Property Prediction

- Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, Pietro Lió
- abstract: Molecular property prediction is one of the fastest-growing applications of deep learning with critical real-world impacts. Although the 3D molecular graph structure is necessary for models to achieve strong performance on many tasks, it is infeasible to obtain 3D structures at the scale required by many real-world applications. To tackle this issue, we propose to use existing 3D molecular datasets to pre-train a model to reason about the geometry of molecules given only their 2D molecular graphs. Our method, called 3D Infomax, maximizes the mutual information between learned 3D summary vectors and the representations of a graph neural network (GNN). During fine-tuning on molecules with unknown geometry, the GNN is still able to produce implicit 3D information and uses it for downstream tasks. We show that 3D Infomax provides significant improvements for a wide range of properties, including a 22% average MAE reduction on QM9 quantum mechanical properties. Moreover, the learned representations can be effectively transferred between datasets in different molecular spaces.

EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction

- Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Dr.Regina Barzilay, Tommi Jaakkola
- abstract: Predicting how a drug-like molecule binds to a specific protein target is a core problem in drug discovery. An extremely fast computational binding method would enable key applications such as fast virtual screening or drug engineering. Existing methods are computationally expensive as they rely on heavy candidate sampling coupled with scoring, ranking, and fine-tuning steps. We challenge this paradigm with EquiBind, an SE(3)-equivariant geometric deep learning model performing direct-shot prediction of both i) the receptor binding location (blind docking) and ii) the ligand's bound pose and orientation. EquiBind achieves significant speed-ups and better quality compared to traditional and recent baselines. Further, we show extra improvements when coupling it with existing fine-tuning techniques at the cost of increased running time. Finally, we propose a novel and fast fine-tuning model that adjusts torsion angles of a ligand's rotatable bonds based on closed form global minima of the von Mises angular distance to a given input atomic point cloud, avoiding previous expensive differential evolution strategies for energy minimization.

Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks

- Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correira, Antonia Adler, Kristian Kersting
- abstract: Model inversion attacks (MIAs) aim to create synthetic images that reflect the class-wise characteristics from a target classifier's private training data by exploiting the model's learned knowledge. Previous research has developed generative MIAs that use generative adversarial networks (GANs) as image priors tailored to a specific target model. This makes the attacks time- and resource-consuming, inflexible, and susceptible to distributional shifts between datasets. To overcome these drawbacks, we present Plug & Play Attacks, which relax the dependency between the target model and image prior, and enable the use of a single GAN to attack a wide range of targets, requiring only minor adjustments to the attack. Moreover, we show that powerful MIAs are possible even with publicly available pre-trained GANs and under strong distributional shifts, for which previous approaches fail to produce meaningful results. Our extensive evaluation confirms the improved robustness and flexibility of Plug & Play Attacks and their ability to create high-quality images revealing sensitive class characteristics.

Scaling-up Diverse Orthogonal Convolutional Networks by a Paraunitary Framework

- Jiahao Su, Wonmin Byeon, Furong Huang
- abstract: Enforcing orthogonality in convolutional neural networks is a remedy for gradient vanishing/exploding problems and sensitivity to perturbation. Many previous approaches for orthogonal convolutions enforce orthogonality on its flattened kernel, which, however, do not lead to the orthogonality of the operation. Some recent approaches consider orthogonality for standard convolutional layers and propose specific classes of their realizations. In this work, we propose a theoretical framework that establishes the equivalence between diverse orthogonal convolutional layers in the spatial domain and the paraunitary systems in the spectral domain. Since 1D paraunitary systems admit a complete factorization, we can parameterize any separable orthogonal convolution as a composition of spatial filters. As a result, our framework endows high expressive power to various convolutional layers while maintaining their exact orthogonality. Furthermore, our layers are memory and computationally efficient for deep networks compared to previous designs. Our versatile framework, for the first time, enables the study of architectural designs for deep orthogonal networks, such as choices of skip connection, initialization, stride, and dilation. Consequently, we scale up orthogonal networks to deep architectures, including ResNet and ShuffleNet, substantially outperforming their shallower counterparts. Finally, we show how to construct residual flows, a flow-based generative model that requires strict Lipschitzness, using our orthogonal networks. Our code will be publicly available at https://github.com/umd-huang-lab/ortho-conv

Divergence-Regularized Multi-Agent Actor-Critic

- Kefan Su, Zongqing Lu
- abstract: Entropy regularization is a popular method in reinforcement learning (RL). Although it has many advantages, it alters the RL objective and makes the converged policy deviate from the optimal policy of the original Markov Decision Process (MDP). Though divergence regularization has been proposed to settle this problem, it cannot be trivially applied to cooperative multi-agent reinforcement learning (MARL). In this paper, we investigate divergence regularization in cooperative MARL and propose a novel off-policy cooperative MARL framework, divergence-regularized multi-agent actor-critic (DMAC). Theoretically, we derive the update rule of DMAC which is naturally off-policy, guarantees the monotonic policy improvement and convergence in both the original MDP and the divergence-regularized MDP, and is not biased by the regularization. We also give a bound of the discrepancy between the converged policy and the optimal policy in the original MDP. DMAC is a flexible framework and can be combined with many

existing MARL algorithms. Empirically, we evaluate DMAC in a didactic stochastic game and StarCraft Multi-Agent Challenge and show that DMAC substantially improves the performance of existing MARL algorithms.

Influence-Augmented Local Simulators: a Scalable Solution for Fast Deep RL in Large Networked Systems

- Miguel Suau, Jinke He, Matthijs T. J. Spaan, Frans Oliehoek
- abstract: Learning effective policies for real-world problems is still an open challenge for the field of reinforcement learning (RL). The main limitation being the amount of data needed and the pace at which that data can be obtained. In this paper, we study how to build lightweight simulators of complicated systems that can run sufficiently fast for deep RL to be applicable. We focus on domains where agents interact with a reduced portion of a larger environment while still being affected by the global dynamics. Our method combines the use of local simulators with learned models that mimic the influence of the global system. The experiments reveal that incorporating this idea into the deep RL workflow can considerably accelerate the training process and presents several opportunities for the future.

Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images

- Rakshith Subramanyam, Vivek Narayanaswamy, Mark Naufel, Andreas Spanias, Jayaraman J. Thiagarajan
- abstract: Inverting an image onto the latent space of pre-trained generators, e.g., StyleGAN-v2, has emerged as a popular strategy to leverage strong image priors for ill-posed restoration. Several studies have showed that this approach is effective at inverting images similar to the data used for training. However, with out-of-distribution (OOD) data that the generator has not been exposed to, existing inversion techniques produce sub-optimal results. In this paper, we propose SPHInX (StyleGAN with Projection Heads for Inverting X), an approach for accurately embedding OOD images onto the StyleGAN latent space. SPHInX optimizes a style projection head using a novel training strategy that imposes a vicinal regularization in the StyleGAN latent space. To further enhance OOD inversion, SPHInX can additionally optimize a content projection head and noise variables in every layer. Our empirical studies on a suite of OOD data show that, in addition to producing higher quality reconstructions over the state-of-the-art inversion techniques, SPHInX is effective for ill-posed restoration tasks while offering semantic editing capabilities.

Continuous-Time Analysis of Accelerated Gradient Methods via Conservation Laws in Dilated Coordinate Systems

- Jaewook J Suh, Gyumin Roh, Ernest K Ryu
- abstract: We analyze continuous-time models of accelerated gradient methods through deriving conservation laws in dilated coordinate systems. Namely, instead of analyzing the dynamics of $X(t)$, we analyze the dynamics of $W(t)=t^\alpha(X(t)-X_c)$ for some α and X_c and derive a conserved quantity, analogous to physical energy, in this dilated coordinate system. Through this methodology, we recover many known continuous-time analyses in a streamlined manner and obtain novel continuous-time analyses for OGM-G, an acceleration mechanism for efficiently reducing gradient magnitude that is distinct from that of Nesterov. Finally, we show that a semi-second-order symplectic Euler discretization in the dilated coordinate system leads to an $\mathcal{O}(1/k^2)$ rate on the standard setup of smooth convex minimization, without any further assumptions such as infinite differentiability.

Do Differentiable Simulators Give Better Policy Gradients?

- Hyung Ju Suh, Max Simchowitz, Kaiqing Zhang, Russ Tedrake
- abstract: Differentiable simulators promise faster computation time for reinforcement learning by replacing zeroth-order gradient estimates of a stochastic objective with an estimate based on first-order gradients. However, it is yet unclear what factors decide the performance of the two estimators on complex landscapes that involve long-horizon planning and control on physical systems, despite the crucial relevance of this question for the utility of differentiable simulators. We show that characteristics of certain physical systems, such as stiffness or discontinuities, may compromise the efficacy of the first-order estimator, and analyze this phenomenon through the lens of bias and variance. We additionally propose an α-order gradient estimator, with $\alpha \in [0,1]$, which correctly utilizes exact gradients to combine the efficiency of first-order estimates with the robustness of zero-order methods. We demonstrate the pitfalls of traditional estimators and the advantages of the α-order estimator on some numerical examples.

Intriguing Properties of Input-Dependent Randomized Smoothing

- Peter Súkeník, Aleksei Kuvshinov, Stephan Günnemann
- abstract: Randomized smoothing is currently considered the state-of-the-art method to obtain certifiably robust classifiers. Despite its remarkable performance, the method is associated with various serious problems such as “certified accuracy waterfalls”, certification vs. accuracy trade-off, or even fairness issues. Input-dependent smoothing approaches have been proposed with intention of overcoming these flaws. However, we demonstrate that these methods lack formal guarantees and so the resulting certificates are not justified. We show that in general, the input-dependent smoothing suffers from the curse of dimensionality, forcing the variance function to have low semi-elasticity. On the other hand, we provide a theoretical and practical framework that enables the usage of input-dependent smoothing even in the presence of the curse of dimensionality, under strict restrictions. We present one concrete design of the smoothing variance function and test it on CIFAR10 and MNIST. Our design mitigates some of the problems of classical smoothing and is formally underlined, yet further improvement of the design is still necessary.

Cliff Diving: Exploring Reward Surfaces in Reinforcement Learning Environments

- Ryan Sullivan, Jordan K Terry, Benjamin Black, John P Dickerson
- abstract: Visualizing optimization landscapes has resulted in many fundamental insights in numeric optimization, specifically regarding novel improvements to optimization techniques. However, visualizations of the objective that reinforcement learning optimizes (the "reward surface") have only ever been generated for a small number of narrow contexts. This work presents reward surfaces and related visualizations of 27 of the most widely used reinforcement learning environments in Gym for the first time. We also explore reward surfaces in the policy gradient direction and show for the first time that many popular reinforcement learning environments have frequent "cliffs" (sudden large drops in expected reward). We demonstrate that A2C often "dives off" these cliffs into low reward regions of the parameter space while PPO avoids them, confirming a popular intuition for PPO’s improved performance over previous methods. We additionally introduce a highly extensible library that allows researchers to easily generate these visualizations in the future. Our findings provide new intuition to explain the successes and failures of modern RL methods, and our visualizations concretely characterize several failure modes of reinforcement learning agents in novel ways.

AGNAS: Attention-Guided Micro and Macro-Architecture Search

- Zihao Sun, Yu Hu, Shun Lu, Longxing Yang, Jilin Mei, Yinhe Han, Xiaowei Li
- abstract: Micro- and macro-architecture search have emerged as two popular NAS paradigms recently. Existing methods leverage different search strategies for searching micro- and macro- architectures. When using architecture parameters to search for micro-structure such as normal cell and reduction cell, the architecture parameters can not fully reflect the corresponding operation importance. When searching for the macro-structure chained by pre-defined blocks, many sub-networks need to be sampled for evaluation, which is very time-consuming. To address the two issues, we propose a new search paradigm, that is, leverage the attention mechanism to guide the micro- and macro-architecture search, namely AGNAS. Specifically, we introduce an attention module and plug it behind each candidate operation or each candidate block. We utilize the attention weights to represent the importance of the relevant operations for the micro search or the importance of the relevant blocks for the macro search. Experimental results show that AGNAS can

achieve 2.46% test error on CIFAR-10 in the DARTS search space, and 23.4% test error when directly searching on ImageNet in the ProxylessNAS search space. AGNAS also achieves optimal performance on NAS-Bench-201, outperforming state-of-the-art approaches. The source code can be available at https://github.com/Sunzh1996/AGNAS.

Adaptive Random Walk Gradient Descent for Decentralized Optimization

- Tao Sun, Dongsheng Li, Bao Wang
- abstract: In this paper, we study the adaptive step size random walk gradient descent with momentum for decentralized optimization, in which the training samples are drawn dependently with each other. We establish theoretical convergence rates of the adaptive step size random walk gradient descent with momentum for both convex and nonconvex settings. In particular, we prove that adaptive random walk algorithms perform as well as the non-adaptive method for dependent data in general cases but achieve acceleration when the stochastic gradients are “sparse”. Moreover, we study the zeroth-order version of adaptive random walk gradient descent and provide corresponding convergence results. All assumptions used in this paper are mild and general, making our results applicable to many machine learning problems.

MAE-DET: Revisiting Maximum Entropy Principle in Zero-Shot NAS for Efficient Object Detection

- Zhenhong Sun, Ming Lin, Xiuyu Sun, Zhiyu Tan, Hao Li, Rong Jin
- abstract: In object detection, the detection backbone consumes more than half of the overall inference cost. Recent researches attempt to reduce this cost by optimizing the backbone architecture with the help of Neural Architecture Search (NAS). However, existing NAS methods for object detection require hundreds to thousands of GPU hours of searching, making them impractical in fast-paced research and development. In this work, we propose a novel zero-shot NAS method to address this issue. The proposed method, named MAE-DET, automatically designs efficient detection backbones via the Maximum Entropy Principle without training network parameters, reducing the architecture design cost to nearly zero yet delivering the state-of-the-art (SOTA) performance. Under the hood, MAE-DET maximizes the differential entropy of detection backbones, leading to a better feature extractor for object detection under the same computational budgets. After merely one GPU day of fully automatic design, MAE-DET innovates SOTA detection backbones on multiple detection benchmark datasets with little human intervention. Comparing to ResNet-50 backbone, MAE-DET is $+2.0%$ better in mAP when using the same amount of FLOPs/parameters, and is 1.54 times faster on NVIDIA V100 at the same mAP. Code and pre-trained models are available here (https://github.com/alibaba/lightweight-neural-architecture-search).

Out-of-Distribution Detection with Deep Nearest Neighbors

- Yiyou Sun, Yifei Ming, Xiaojin Zhu, Yixuan Li
- abstract: Out-of-distribution (OOD) detection is a critical task for deploying machine learning models in the open world. Distance-based methods have demonstrated promise, where testing samples are detected as OOD if they are relatively far away from in-distribution (ID) data. However, prior methods impose a strong distributional assumption of the underlying feature space, which may not always hold. In this paper, we explore the efficacy of non-parametric nearest-neighbor distance for OOD detection, which has been largely overlooked in the literature. Unlike prior works, our method does not impose any distributional assumption, hence providing stronger flexibility and generality. We demonstrate the effectiveness of nearest-neighbor-based OOD detection on several benchmarks and establish superior performance. Under the same model trained on ImageNet-1k, our method substantially reduces the false positive rate (FPR@TPR95) by 24.77% compared to a strong baseline SSD+, which uses a parametric approach Mahalanobis distance in detection. Code is available: https://github.com/deeplearning-wisc/knn-ood.

Black-Box Tuning for Language-Model-as-a-Service

- Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, Xipeng Qiu
- abstract: Extremely large pre-trained language models (PTMs) such as GPT-3 are usually released as a service. It allows users to design task-specific prompts to query the PTMs through some black-box APIs. In such a scenario, which we call Language-Model-as-a-Service (LMaaS), the gradients of PTMs are usually unavailable. Can we optimize the task prompts by only accessing the model inference APIs? This paper proposes the black-box tuning framework to optimize the continuous prompt prepended to the input text via derivative-free optimization. Instead of optimizing in the original high-dimensional prompt space, which is intractable for traditional derivative-free optimization, we perform optimization in a randomly generated subspace due to the low intrinsic dimensionality of large PTMs. The experimental results show that the black-box tuning with RoBERTa on a few labeled samples not only significantly outperforms manual prompt and GPT-3’s in-context learning, but also surpasses the gradient-based counterparts, i.e., prompt tuning and full model tuning.

Correlated Quantization for Distributed Mean Estimation and Optimization

- Ananda Theertha Suresh, Ziteng Sun, Jae Ro, Felix Yu
- abstract: We study the problem of distributed mean estimation and optimization under communication constraints. We propose a correlated quantization protocol whose error guarantee depends on the deviation of data points instead of their absolute range. The design doesn’t need any prior knowledge on the concentration property of the dataset, which is required to get such dependence in previous works. We show that applying the proposed protocol as a sub-routine in distributed optimization algorithms leads to better convergence rates. We also prove the optimality of our protocol under mild assumptions. Experimental results show that our proposed algorithm outperforms existing mean estimation protocols on a diverse set of tasks.

Causal Imitation Learning under Temporally Correlated Noise

- Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, Steven Wu
- abstract: We develop algorithms for imitation learning from policy data that was corrupted by temporally correlated noise in expert actions. When noise affects multiple timesteps of recorded data, it can manifest as spurious correlations between states and actions that a learner might latch on to, leading to poor policy performance. To break up these spurious correlations, we apply modern variants of the instrumental variable regression (IVR) technique of econometrics, enabling us to recover the underlying policy without requiring access to an interactive expert. In particular, we present two techniques, one of a generative-modeling flavor (DoubIL) that can utilize access to a simulator, and one of a game-theoretic flavor (ResiduIL) that can be run entirely offline. We find both of our algorithms compare favorably to behavioral cloning on simulated control tasks.

Being Properly Improper

- Tyler Sypherd, Richard Nock, Lalitha Sankar
- abstract: Properness for supervised losses stipulates that the loss function shapes the learning algorithm towards the true posterior of the data generating distribution. Unfortunately, data in modern machine learning can be corrupted or twisted in many ways. Hence, optimizing a proper loss function on twisted data could perilously lead the learning algorithm towards the twisted posterior, rather than to the desired clean posterior. Many papers cope with specific twists (e.g., label/feature/adversarial noise), but there is a growing need for a unified and actionable understanding atop properness. Our chief theoretical contribution is a generalization of the properness framework with a notion called twist-properness, which delineates loss functions with the ability to "untwist" the twisted posterior into the clean posterior. Notably, we show that a nontrivial extension of a loss function called alpha-loss, which was first introduced in information theory, is twist-proper. We study the twist-proper alpha-loss under a novel boosting algorithm, called PILBoost, and

provide formal and experimental results for this algorithm. Our overarching practical conclusion is that the twist-proper alpha-loss outperforms the proper log-loss on several variants of twisted data.

Distributionally-Aware Kernelized Bandit Problems for Risk Aversion

- Sho Takemori
- abstract: The kernelized bandit problem is a theoretically justified framework and has solid applications to various fields. Recently, there is a growing interest in generalizing the problem to the optimization of risk-averse metrics such as Conditional Value-at-Risk (CVaR) or Mean-Variance (MV). However, due to the model assumption, most existing methods need explicit design of environment random variables and can incur large regret because of possible high dimensionality of them. To address the issues, in this paper, we model environments using a family of the output distributions (or more precisely, probability kernel) and Kernel Mean Embeddings (KME), and provide novel UCB-type algorithms for CVaR and MV. Moreover, we provide algorithm-independent lower bounds for CVaR in the case of Matérn kernels, and propose a nearly optimal algorithm. Furthermore, we empirically verify our theoretical result in synthetic environments, and demonstrate that our proposed method significantly outperforms a baseline in many cases.

Sequential and Parallel Constrained Max-value Entropy Search via Information Lower Bound

- Shion Takeno, Tomoyuki Tamura, Kazuki Shitara, Masayuki Karasuyama
- abstract: Max-value entropy search (MES) is one of the state-of-the-art approaches in Bayesian optimization (BO). In this paper, we propose a novel variant of MES for constrained problems, called Constrained MES via Information lower BOund (CMES-IBO), that is based on a Monte Carlo (MC) estimator of a lower bound of a mutual information (MI). Unlike existing studies, our MI is defined so that uncertainty with respect to feasibility can be incorporated. We derive a lower bound of the MI that guarantees non-negativity, while a constrained counterpart of conventional MES can be negative. We further provide theoretical analysis that assures the low-variability of our estimator which has never been investigated for any existing information-theoretic BO. Moreover, using the conditional MI, we extend CMES-IBO to the parallel setting while maintaining the desirable properties. We demonstrate the effectiveness of CMES-IBO by several benchmark functions and real-world problems.

SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization

- Yuhta Takida, Takashi Shibuya, Weihsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, Yuki Mitsufuji
- abstract: One noted issue of vector-quantized variational autoencoder (VQ-VAE) is that the learned discrete representation uses only a fraction of the full capacity of the codebook, also known as codebook collapse. We hypothesize that the training scheme of VQ-VAE, which involves some carefully designed heuristics, underlies this issue. In this paper, we propose a new training scheme that extends the standard VAE via novel stochastic dequantization and quantization, called stochastically quantized variational autoencoder (SQ-VAE). In SQ-VAE, we observe a trend that the quantization is stochastic at the initial stage of the training but gradually converges toward a deterministic quantization, which we call self-annealing. Our experiments show that SQ-VAE improves codebook utilization without using common heuristics. Furthermore, we empirically show that SQ-VAE is superior to VAE and VQ-VAE in vision- and speech-related tasks.

A Tree-based Model Averaging Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources

- Xiaoqing Tan, Chung-Chou H. Chang, Ling Zhou, Lu Tang
- abstract: Accurately estimating personalized treatment effects within a study site (e.g., a hospital) has been challenging due to limited sample size. Furthermore, privacy considerations and lack of resources prevent a site from leveraging subject-level data from other sites. We propose a tree-based model averaging approach to improve the estimation accuracy of conditional average treatment effects (CATE) at a target site by leveraging models derived from other potentially heterogeneous sites, without them sharing subject-level data. To our best knowledge, there is no established model averaging approach for distributed data with a focus on improving the estimation of treatment effects. Specifically, under distributed data networks, our framework provides an interpretable tree-based ensemble of CATE estimators that joins models across study sites, while actively modeling the heterogeneity in data sources through site partitioning. The performance of this approach is demonstrated by a real-world study of the causal effects of oxygen therapy on hospital survival rate and backed up by comprehensive simulation results.

N-Penetrate: Active Learning of Neural Collision Handler for Complex 3D Mesh Deformations

- Qingyang Tan, Zherong Pan, Breannan Smith, Takaaki Shiratori, Dinesh Manocha
- abstract: We present a robust learning algorithm to detect and handle collisions in 3D deforming meshes. We first train a neural network to detect collisions and then use a numerical optimization algorithm to resolve penetrations guided by the network. Our learned collision handler can resolve collisions for unseen, high-dimensional meshes with thousands of vertices. To obtain stable network performance in such large and unseen spaces, we apply active learning by progressively inserting new collision data based on the network inferences. We automatically label these new data using an analytical collision detector and progressively fine-tune our detection networks. We evaluate our method for collision handling of complex, 3D meshes coming from several datasets with different shapes and topologies, including datasets corresponding to dressed and undressed human poses, cloth simulations, and human hand poses acquired using multi-view capture systems.

Biased Gradient Estimate with Drastic Variance Reduction for Meta Reinforcement Learning

- Yunhao Tang
- abstract: Despite the empirical success of meta reinforcement learning (meta-RL), there are still a number poorly-understood discrepancies between theory and practice. Critically, biased gradient estimates are almost always implemented in practice, whereas prior theory on meta-RL only establishes convergence under unbiased gradient estimates. In this work, we investigate such a discrepancy. In particular, (1) We show that unbiased gradient estimates have variance $\Theta(N)$ which linearly depends on the sample size N of the inner loop updates; (2) We propose linearized score function (LSF) gradient estimates, which have bias $\mathcal{O}(1/\sqrt{N})$ and variance $\mathcal{O}(1/N)$; (3) We show that most empirical prior work in fact implements variants of the LSF gradient estimates. This implies that practical algorithms "accidentally" introduce bias to achieve better performance; (4) We establish theoretical guarantees for the LSF gradient estimates in meta-RL regarding its convergence to stationary points, showing better dependency on N than prior work when N is large.

Rethinking Graph Neural Networks for Anomaly Detection

- Jianheng Tang, Jiajin Li, Ziqi Gao, Jia Li
- abstract: Graph Neural Networks (GNNs) are widely applied for graph anomaly detection. As one of the key components for GNN design is to select a tailored spectral filter, we take the first step towards analyzing anomalies via the lens of the graph spectrum. Our crucial observation is the existence of anomalies will lead to the ‘right-shift’ phenomenon, that is, the spectral energy distribution concentrates less on low frequencies and more on high frequencies. This fact motivates us to propose the Beta Wavelet Graph Neural Network (BWGNN). Indeed, BWGNN has spectral and spatial localized band-pass filters to better handle the ‘right-shift’ phenomenon in anomalies. We demonstrate the effectiveness of BWGNN on four large-scale anomaly detection datasets. Our code and data are released at https://github.com/squareRoot3/Rethinking-Anomaly-Detection.

Deep Safe Incomplete Multi-view Clustering: Theorem and Algorithm

- Huayi Tang, Yong Liu
- abstract: Incomplete multi-view clustering is a significant but challenging task. Although jointly imputing incomplete samples and conducting clustering has been shown to achieve promising performance, learning from both complete and incomplete data may be worse than learning only from complete data, particularly when imputed views are semantic inconsistent with missing views. To address this issue, we propose a novel framework to reduce the clustering performance degradation risk from semantic inconsistent imputed views. Concretely, by the proposed bi-level optimization framework, missing views are dynamically imputed from the learned semantic neighbors, and imputed samples are automatically selected for training. In theory, the empirical risk of the model is no higher than learning only from complete data, and the model is never worse than learning only from complete data in terms of expected risk with high probability. Comprehensive experiments demonstrate that the proposed method achieves superior performance and efficient safe incomplete multi-view clustering.

Virtual Homogeneity Learning: Defending against Data Heterogeneity in Federated Learning

- Zhenheng Tang, Yonggang Zhang, Shaohuai Shi, Xin He, Bo Han, Xiaowen Chu
- abstract: In federated learning (FL), model performance typically suffers from client drift induced by data heterogeneity, and mainstream works focus on correcting client drift. We propose a different approach named virtual homogeneity learning (VHL) to directly “rectify” the data heterogeneity. In particular, VHL conducts FL with a virtual homogeneous dataset crafted to satisfy two conditions: containing no private information and being separable. The virtual dataset can be generated from pure noise shared across clients, aiming to calibrate the features from the heterogeneous clients. Theoretically, we prove that VHL can achieve provable generalization performance on the natural distribution. Empirically, we demonstrate that VHL endows FL with drastically improved convergence speed and generalization performance. VHL is the first attempt towards using a virtual dataset to address data heterogeneity, offering new and effective means to FL.

Cross-Space Active Learning on Graph Convolutional Networks

- Yufei Tao, Hao Wu, Shiyuan Deng
- abstract: This paper formalizes cross-space active learning on a graph convolutional network (GCN). The objective is to attain the most accurate hypothesis available in any of the instance spaces generated by the GCN. Subject to the objective, the challenge is to minimize the label cost, measured in the number of vertices whose labels are requested. Our study covers both budget algorithms which terminate after a designated number of label requests, and verifiable algorithms which terminate only after having found an accurate hypothesis. A new separation in label complexity between the two algorithm types is established. The separation is unique to GCNs.

FedNest: Federated Bilevel, Minimax, and Compositional Optimization

- Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, Samet Oymak
- abstract: Standard federated optimization methods successfully apply to stochastic problems with single-level structure. However, many contemporary ML problems - including adversarial robustness, hyperparameter tuning, actor-critic - fall under nested bilevel programming that subsumes minimax and compositional optimization. In this work, we propose FedNest: A federated alternating stochastic gradient method to address general nested problems. We establish provable convergence rates for FedNest in the presence of heterogeneous data and introduce variations for bilevel, minimax, and compositional optimization. FedNest introduces multiple innovations including federated hypergradient computation and variance reduction to address inner-level heterogeneity. We complement our theory with experiments on hyperparameter & hyper-representation learning and minimax optimization that demonstrate the benefits of our method in practice.

Efficient Distributionally Robust Bayesian Optimization with Worst-case Sensitivity

- Sebastian Shenghong Tay, Chuan Sheng Foo, Urano Daisuke, Richalynn Leong, Bryan Kian Hsiang Low
- abstract: In distributionally robust Bayesian optimization (DRBO), an exact computation of the worst-case expected value requires solving an expensive convex optimization problem. We develop a fast approximation of the worst-case expected value based on the notion of worst-case sensitivity that caters to arbitrary convex distribution distances. We provide a regret bound for our novel DRBO algorithm with the fast approximation, and empirically show it is competitive with that using the exact worst-case expected value while incurring significantly less computation time. In order to guide the choice of distribution distance to be used with DRBO, we show that our approximation implicitly optimizes an objective close to an interpretable risk-sensitive value.

LIDL: Local Intrinsic Dimension Estimation Using Approximate Likelihood

- Piotr Tempczyk, Rafał Michaluk, Lukasz Garncarek, Przemysław Spurek, Jacek Tabor, Adam Golinski
- abstract: Most of the existing methods for estimating the local intrinsic dimension of a data distribution do not scale well to high dimensional data. Many of them rely on a non-parametric nearest neighbours approach which suffers from the curse of dimensionality. We attempt to address that challenge by proposing a novel approach to the problem: Local Intrinsic Dimension estimation using approximate Likelihood (LIDL). Our method relies on an arbitrary density estimation method as its subroutine, and hence tries to sidestep the dimensionality challenge by making use of the recent progress in parametric neural methods for likelihood estimation. We carefully investigate the empirical properties of the proposed method, compare them with our theoretical predictions, show that LIDL yields competitive results on the standard benchmarks for this problem, and that it scales to thousands of dimensions. What is more, we anticipate this approach to improve further with the continuing advances in the density estimation literature.

LCANets: Lateral Competition Improves Robustness Against Corruption and Attack

- Michael Teti, Garrett Kenyon, Ben Migliori, Juston Moore
- abstract: Although Convolutional Neural Networks (CNNs) achieve high accuracy on image recognition tasks, they lack robustness against realistic corruptions and fail catastrophically when deliberately attacked. Previous CNNs with representations similar to primary visual cortex (V1) were more robust to adversarial attacks on images than current adversarial defense techniques, but they required training on large-scale neural recordings or handcrafting neuroscientific models. Motivated by evidence that neural activity in V1 is sparse, we develop a class of hybrid CNNs, called LCANets, which feature a frontend that performs sparse coding via local lateral competition. We demonstrate that LCANets achieve competitive clean accuracy to standard CNNs on action and image recognition tasks and significantly greater accuracy under various image corruptions. We also perform the first adversarial attacks with full knowledge of a sparse coding CNN layer by attacking LCANets with white-box and black-box attacks, and we show that, contrary to previous hypotheses, sparse coding layers are not very robust to white-box attacks. Finally, we propose a way to use sparse coding layers as a plug-and-play robust frontend by showing that they significantly increase the robustness of adversarially-trained CNNs over corruptions and attacks.

Reverse Engineering ℓ_p attacks: A block-sparse optimization approach with recovery guarantees

- Darshan Thaker, Paris Giampouras, Rene Vidal

- abstract: Deep neural network-based classifiers have been shown to be vulnerable to imperceptible perturbations to their input, such as ℓ_p-bounded norm adversarial attacks. This has motivated the development of many defense methods, which are then broken by new attacks, and so on. This paper focuses on a different but related problem of reverse engineering adversarial attacks. Specifically, given an attacked signal, we study conditions under which one can determine the type of attack (ℓ_1, ℓ_2 or ℓ_∞) and recover the clean signal. We pose this problem as a block-sparse recovery problem, where both the signal and the attack are assumed to lie in a union of subspaces that includes one subspace per class and one subspace per attack type. We derive geometric conditions on the subspaces under which any attacked signal can be decomposed as the sum of a clean signal plus an attack. In addition, by determining the subspaces that contain the signal and the attack, we can also classify the signal and determine the attack type. Experiments on digit and face classification demonstrate the effectiveness of the proposed approach.

Generalised Policy Improvement with Geometric Policy Composition

- Shantanu Thakoor, Mark Rowland, Diana Borsa, Will Dabney, Remi Munos, Andre Barreto
- abstract: We introduce a method for policy improvement that interpolates between the greedy approach of value-based reinforcement learning (RL) and the full planning approach typical of model-based RL. The new method builds on the concept of a geometric horizon model (GHM, also known as a \gamma-model), which models the discounted state-visitation distribution of a given policy. We show that we can evaluate any non-Markov policy that switches between a set of base Markov policies with fixed probability by a careful composition of the base policy GHMs, without any additional learning. We can then apply generalised policy improvement (GPI) to collections of such non-Markov policies to obtain a new Markov policy that will in general outperform its precursors. We provide a thorough theoretical analysis of this approach, develop applications to transfer and standard RL, and empirically demonstrate its effectiveness over standard GPI on a challenging deep RL continuous control task. We also provide an analysis of GHM training methods, proving a novel convergence result regarding previously proposed methods and showing how to train these models stably in deep RL settings.

Algorithms for the Communication of Samples

- Lucas Theis, Noureldin Y Ahmed
- abstract: The efficient communication of noisy data has applications in several areas of machine learning, such as neural compression or differential privacy, and is also known as reverse channel coding or the channel simulation problem. Here we propose two new coding schemes with practical advantages over existing approaches. First, we introduce ordered random coding (ORC) which uses a simple trick to reduce the coding cost of previous approaches. This scheme further illuminates a connection between schemes based on importance sampling and the so-called Poisson functional representation. Second, we describe a hybrid coding scheme which uses dithered quantization to more efficiently communicate samples from distributions with bounded support.

Consistent Polyhedral Surrogates for Top-k Classification and Variants

- Anish Thilagar, Rafael Frongillo, Jessica J Finocchiaro, Emma Goodwill
- abstract: Top-k classification is a generalization of multiclass classification used widely in information retrieval, image classification, and other extreme classification settings. Several hinge-like (piecewise-linear) surrogates have been proposed for the problem, yet all are either non-convex or inconsistent. For the proposed hinge-like surrogates that are convex (i.e., polyhedral), we apply the recent embedding framework of Finocchiaro et al. (2019; 2022) to determine the prediction problem for which the surrogate is consistent. These problems can all be interpreted as variants of top-k classification, which may be better aligned with some applications. We leverage this analysis to derive constraints on the conditional label distributions under which these proposed surrogates become consistent for top-k. It has been further suggested that every convex hinge-like surrogate must be inconsistent for top-k. Yet, we use the same embedding framework to give the first consistent polyhedral surrogate for this problem.

On the Finite-Time Complexity and Practical Computation of Approximate Stationarity Concepts of Lipschitz Functions

- Lai Tian, Kaiwen Zhou, Anthony Man-Cho So
- abstract: We report a practical finite-time algorithmic scheme to compute approximately stationary points for nonconvex nonsmooth Lipschitz functions. In particular, we are interested in two kinds of approximate stationarity notions for nonconvex nonsmooth problems, i.e., Goldstein approximate stationarity (GAS) and near-approximate stationarity (NAS). For GAS, our scheme removes the unrealistic subgradient selection oracle assumption in (Zhang et al., 2020, Assumption 1) and computes GAS with the same finite-time complexity. For NAS, Davis & Drusvyatskiy (2019) showed that ρ-weakly convex functions admit finite-time computation, while Tian & So (2021) provided the matching impossibility results of dimension-free finite-time complexity for first-order methods. Complement to these developments, in this paper, we isolate a new class of functions that could be Clarke irregular (and thus not weakly convex anymore) and show that our new algorithmic scheme can compute NAS points for functions in that class within finite time. To demonstrate the wide applicability of our new theoretical framework, we show that ρ-margin SVM, 1-layer, and 2-layer ReLU neural networks, all being Clarke irregular, satisfy our new conditions.

From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

- Daniil Tiapkin, Denis Belomestny, Eric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, Pierre Menard
- abstract: We propose the Bayes-UCBVI algorithm for reinforcement learning in tabular, stage-dependent, episodic Markov decision process: a natural extension of the Bayes-UCB algorithm by Kaufmann et al. 2012 for multi-armed bandits. Our method uses the quantile of a Q-value function posterior as upper confidence bound on the optimal Q-value function. For Bayes-UCBVI, we prove a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ where H is the length of one episode, S is the number of states, A the number of actions, T the number of episodes, that matches the lower-bound of $\Omega(\sqrt{H^3SAT})$ up to poly-$\log$ terms in H,S,A,T for a large enough T. To the best of our knowledge, this is the first algorithm that obtains an optimal dependence on the horizon H (and S) without the need of an involved Bernstein-like bonus or noise. Crucial to our analysis is a new fine-grained anti-concentration bound for a weighted Dirichlet sum that can be of independent interest. We then explain how Bayes-UCBVI can be easily extended beyond the tabular setting, exhibiting a strong link between our algorithm and Bayesian bootstrap (Rubin,1981).

Nonparametric Sparse Tensor Factorization with Hierarchical Gamma Processes

- Conor Tillinghast, Zheng Wang, Shandian Zhe
- abstract: We propose a nonparametric factorization approach for sparsely observed tensors. The sparsity does not mean zero-valued entries are massive or dominated. Rather, it implies the observed entries are very few, and even fewer with the growth of the tensor; this is ubiquitous in practice. Compared with the existent works, our model not only leverages the structural information underlying the observed entry indices, but also provides extra interpretability and flexibility {—} it can simultaneously estimate a set of location factors about the intrinsic properties of the tensor nodes, and another set of sociability factors reflecting their extrovert activity in interacting with others; users are free to choose a trade-off between the two types of factors. Specifically, we use hierarchical Gamma processes and Poisson random measures to construct a tensor-valued process, which can freely sample the two types of factors to generate tensors and always guarantees an asymptotic sparsity. We then normalize the tensor process to obtain hierarchical Dirichlet processes to sample each observed entry index, and use a Gaussian process to sample the entry value as a nonlinear function of the factors, so as to capture both the sparse structure properties and complex node relationships. For efficient inference, we use Dirichlet process properties over finite sample partitions, density transformations, and random features to develop a stochastic variational estimation algorithm. We demonstrate the advantage of our method in several benchmark datasets.

Deciphering Lasso-based Classification Through a Large Dimensional Analysis of the Iterative Soft-Thresholding Algorithm

- Malik Tiomoko, Ekkehard Schnoor, Mohamed El Amine Seddik, Igor Colin, Aladin Virmaux
- abstract: This paper proposes a theoretical analysis of a Lasso-based classification algorithm. Leveraging on a realistic regime where the dimension of the data p and their number n are of the same order of magnitude, the theoretical classification error is derived as a function of the data statistics. As a result, insights into the functioning of the Lasso in classification and its differences with competing algorithms are highlighted. Our work is based on an original novel analysis of the Iterative Soft-Thresholding Algorithm (ISTA), which may be of independent interest beyond the particular problem studied here and may be adapted to similar iterative schemes. A theoretical optimization of the model's hyperparameters is also provided, which allows for the data- and time-consuming cross-validation to be avoided. Finally, several applications on synthetic and real data are provided to validate the theoretical study and justify its impact in the design and understanding of algorithms of practical interest.

Extended Unconstrained Features Model for Exploring Deep Neural Collapse

- Tom Tirer, Joan Bruna
- abstract: The modern strategy for training deep neural networks for classification tasks includes optimizing the network's weights even after the training error vanishes to further push the training loss toward zero. Recently, a phenomenon termed "neural collapse" (NC) has been empirically observed in this training procedure. Specifically, it has been shown that the learned features (the output of the penultimate layer) of within-class samples converge to their mean, and the means of different classes exhibit a certain tight frame structure, which is also aligned with the last layer's weights. Recent papers have shown that minimizers with this structure emerge when optimizing a simplified "unconstrained features model" (UFM) with a regularized cross-entropy loss. In this paper, we further analyze and extend the UFM. First, we study the UFM for the regularized MSE loss, and show that the minimizers' features can have a more delicate structure than in the cross-entropy case. This affects also the structure of the weights. Then, we extend the UFM by adding another layer of weights as well as ReLU nonlinearity to the model and generalize our previous results. Finally, we empirically demonstrate the usefulness of our nonlinear extended UFM in modeling the NC phenomenon that occurs with practical networks.

Object Permanence Emerges in a Random Walk along Memory

- Pavel Tokmakov, Allan Jabri, Jie Li, Adrien Gaidon
- abstract: This paper proposes a self-supervised objective for learning representations that localize objects under occlusion - a property known as object permanence. A central question is the choice of learning signal in cases of total occlusion. Rather than directly supervising the locations of invisible objects, we propose a self-supervised objective that requires neither human annotation, nor assumptions about object dynamics. We show that object permanence can emerge by optimizing for temporal coherence of memory: we fit a Markov walk along a space-time graph of memories, where the states in each time step are non-Markovian features from a sequence encoder. This leads to a memory representation that stores occluded objects and predicts their motion, to better localize them. The resulting model outperforms existing approaches on several datasets of increasing complexity and realism, despite requiring minimal supervision, and hence being broadly applicable.

Generic Coreset for Scalable Learning of Monotonic Kernels: Logistic Regression, Sigmoid and more

- Elad Tolochinksy, Ibrahim Jubran, Dan Feldman
- abstract: Coreset (or core-set) is a small weighted subset Q of an input set P with respect to a given monotonic function $f:\mathbb{R}\to\mathbb{R}$ that provably approximates its fitting loss $\sum_{p\in P}f(p\cdot x)$ to any given $x\in\mathbb{R}^d$. Using Q we can obtain an approximation of x^* that minimizes this loss, by running existing optimization algorithms on Q. In this work we provide: (i) A lower bound which proves that there are sets with no coresets smaller than $n=|P|$ for general monotonic loss functions. (ii) A proof that, with an additional common regularization term and under a natural assumption that holds e.g. for logistic regression and the sigmoid activation functions, a small coreset exists for any input P. (iii) A generic coreset construction algorithm that computes such a small coreset Q in $O(nd+n\log n)$ time, and (iv) Experimental results with open-source code which demonstrate that our coresets are effective and are much smaller in practice than predicted in theory.

Failure and success of the spectral bias prediction for Laplace Kernel Ridge Regression: the case of low-dimensional data

- Umberto M Tomasini, Antonio Sclocchi, Matthieu Wyart
- abstract: Recently, several theories including the replica method made predictions for the generalization error of Kernel Ridge Regression. In some regimes, they predict that the method has a 'spectral bias': decomposing the true function $f^$ *on the eigenbasis of the kernel, it fits well the coefficients associated with the O(P) largest eigenvalues, where P is the size of the training set. This prediction works very well on benchmark data sets such as images, yet the assumptions these approaches make on the data are never satisfied in practice. To clarify when the spectral bias prediction holds, we first focus on a one-dimensional model where rigorous results are obtained and then use scaling arguments to generalize and test our findings in higher dimensions. Our predictions include the classification case $f(x)=$sign(x_1) with a data distribution that vanishes at the decision boundary $p(x)\sim x_1^{\chi}$. For $\chi>0$ and a Laplace kernel, we find that (i) there exists a cross-over ridge $\lambda^{d,\chi}(P)\sim P^{-\frac{1}{d+\chi}}$ such that for $\lambda\gg \lambda^_{d,\chi}(P)$, the replica method applies, but not for $\lambda\ll\lambda^{d,\chi}(P)$,* (ii) in the ridge-less case, spectral bias predicts the correct training curve exponent only in the limit $d\rightarrow\infty$.

Quantifying and Learning Linear Symmetry-Based Disentanglement

- Loek Tonnaer, Luis Armando Perez Rey, Vlado Menkovski, Mike Holenderski, Jim Portegies
- abstract: The definition of Linear Symmetry-Based Disentanglement (LSBD) formalizes the notion of linearly disentangled representations, but there is currently no metric to quantify LSBD. Such a metric is crucial to evaluate LSBD methods and to compare them to previous understandings of disentanglement. We propose D_LSBD, a mathematically sound metric to quantify LSBD, and provide a practical implementation for SO(2) groups. Furthermore, from this metric we derive LSBD-VAE, a semi-supervised method to learn LSBD representations. We demonstrate the utility of our metric by showing that (1) common VAE-based disentanglement methods don't learn LSBD representations, (2) LSBD-VAE, as well as other recent methods, can learn LSBD representations needing only limited supervision on transformations, and (3) various desirable properties expressed by existing disentanglement metrics are also achieved by LSBD representations.

A Temporal-Difference Approach to Policy Gradient Estimation

- Samuele Tosatto, Andrew Patterson, Martha White, Rupam Mahmood
- abstract: The policy gradient theorem (Sutton et al., 2000) prescribes the usage of a cumulative discounted state distribution under the target policy to approximate the gradient. Most algorithms based on this theorem, in practice, break this assumption, introducing a distribution shift that can cause the convergence to poor solutions. In this paper, we propose a new approach of reconstructing the policy gradient from the start state without requiring a particular sampling strategy. The policy gradient calculation in this form can be simplified in terms of a gradient critic, which can be recursively estimated due to a new Bellman equation of gradients. By using temporal-difference updates of the gradient critic from an off-policy data stream, we develop the first estimator that side-steps the distribution shift issue in a model-free way. We prove that, under certain realizability conditions, our estimator is unbiased regardless of the sampling strategy. We empirically show that our technique achieves a superior bias-variance trade-off and performance in presence of off-policy samples.

Simple and near-optimal algorithms for hidden stratification and multi-group learning

- Christopher J Tosh, Daniel Hsu
- abstract: Multi-group agnostic learning is a formal learning criterion that is concerned with the conditional risks of predictors within subgroups of a population. The criterion addresses recent practical concerns such as subgroup fairness and hidden stratification. This paper studies the structure of solutions to the multi-group learning problem, and provides simple and near-optimal algorithms for the learning problem.

Design-Bench: Benchmarks for Data-Driven Offline Model-Based Optimization

- Brandon Trabucco, Xinyang Geng, Aviral Kumar, Sergey Levine
- abstract: Black-box model-based optimization (MBO) problems, where the goal is to find a design input that maximizes an unknown objective function, are ubiquitous in a wide range of domains, such as the design of proteins, DNA sequences, aircraft, and robots. Solving model-based optimization problems typically requires actively querying the unknown objective function on design proposals, which means physically building the candidate molecule, aircraft, or robot, testing it, and storing the result. This process can be expensive and time consuming, and one might instead prefer to optimize for the best design using only the data one already has. This setting—called offline MBO—poses substantial and different algorithmic challenges than more commonly studied online techniques. A number of recent works have demonstrated success with offline MBO for high-dimensional optimization problems using high-capacity deep neural networks. However, the lack of standardized benchmarks in this emerging field is making progress difficult to track. To address this, we present Design-Bench, a benchmark for offline MBO with a unified evaluation protocol and reference implementations of recent methods. Our benchmark includes a suite of diverse and realistic tasks derived from real-world optimization problems in biology, materials science, and robotics that present distinct challenges for offline MBO. Our benchmark and reference implementations are released at github.com/rail-berkeley/design-bench and github.com/rail-berkeley/design-baselines.

AnyMorph: Learning Transferable Polices By Inferring Agent Morphology

- Brandon Trabucco, Mariano Phielipp, Glen Berseth
- abstract: The prototypical approach to reinforcement learning involves training policies tailored to a particular agent from scratch for every new morphology. Recent work aims to eliminate the re-training of policies by investigating whether a morphology-agnostic policy, trained on a diverse set of agents with similar task objectives, can be transferred to new agents with unseen morphologies without re-training. This is a challenging problem that required previous approaches to use hand-designed descriptions of the new agent's morphology. Instead of hand-designing this description, we propose a data-driven method that learns a representation of morphology directly from the reinforcement learning objective. Ours is the first reinforcement learning algorithm that can train a policy to generalize to new agent morphologies without requiring a description of the agent's morphology in advance. We evaluate our approach on the standard benchmark for agent-agnostic control, and improve over the current state of the art in zero-shot generalization to new agents. Importantly, our method attains good performance without an explicit description of morphology.

Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them

- Florian Tramer
- abstract: Making classifiers robust to adversarial examples is challenging. Thus, many works tackle the seemingly easier task of detecting perturbed inputs. We show a barrier towards this goal. We prove a hardness reduction between detection and classification of adversarial examples: given a robust detector for attacks at distance ϵ (in some metric), we show how to build a similarly robust (but inefficient) classifier for attacks at distance $\epsilon/2$. Our reduction is computationally inefficient, but preserves the data complexity of the original detector. The reduction thus cannot be directly used to build practical classifiers. Instead, it is a useful sanity check to test whether empirical detection results imply something much stronger than the authors presumably anticipated (namely a highly robust and data-efficient classifier). To illustrate, we revisit 14 empirical detector defenses published over the past years. For $12/14$ defenses, we show that the claimed detection results imply an inefficient classifier with robustness far beyond the state-of-the-art— thus casting some doubts on the results' validity. Finally, we show that our reduction applies in both directions: a robust classifier for attacks at distance $\epsilon/2$ implies an inefficient robust detector at distance ϵ. Thus, we argue that robust classification and robust detection should be regarded as (near)-equivalent problems, if we disregard their computational complexity.

Nesterov Accelerated Shuffling Gradient Method for Convex Optimization

- Trang H Tran, Katya Scheinberg, Lam M Nguyen
- abstract: In this paper, we propose Nesterov Accelerated Shuffling Gradient (NASG), a new algorithm for the convex finite-sum minimization problems. Our method integrates the traditional Nesterov's acceleration momentum with different shuffling sampling schemes. We show that our algorithm has an improved rate of $\Ocal(1/T)$ using unified shuffling schemes, where T is the number of epochs. This rate is better than that of any other shuffling gradient methods in convex regime. Our convergence analysis does not require an assumption on bounded domain or a bounded gradient condition. For randomized shuffling schemes, we improve the convergence bound further. When employing some initial condition, we show that our method converges faster near the small neighborhood of the solution. Numerical simulations demonstrate the efficiency of our algorithm.

A Completely Tuning-Free and Robust Approach to Sparse Precision Matrix Estimation

- Chau Tran, Guo Yu
- abstract: Despite the vast literature on sparse Gaussian graphical models, current methods either are asymptotically tuning-free (which still require fine-tuning in practice) or hinge on computationally expensive methods (e.g., cross-validation) to determine the proper level of regularization. We propose a completely tuning-free approach for estimating sparse Gaussian graphical models. Our method uses model-agnostic regularization parameters to estimate each column of the target precision matrix and enjoys several desirable properties. Computationally, our estimator can be computed efficiently by linear programming. Theoretically, the proposed estimator achieves minimax optimal convergence rates under various norms. We further propose a second-stage enhancement with non-convex penalties which possesses the strong oracle property. Through comprehensive numerical studies, our methods demonstrate favorable statistical performance. Remarkably, our methods exhibit strong robustness to the violation of the Gaussian assumption and significantly outperform competing methods in the heavy-tailed settings.

Tackling covariate shift with node-based Bayesian neural networks

- Trung Q Trinh, Markus Heinonen, Luigi Acerbi, Samuel Kaski
- abstract: Bayesian neural networks (BNNs) promise improved generalization under covariate shift by providing principled probabilistic representations of epistemic uncertainty. However, weight-based BNNs often struggle with high computational complexity of large-scale architectures and datasets. Node-based BNNs have recently been introduced as scalable alternatives, which induce epistemic uncertainty by multiplying each hidden node with latent random variables, while learning a point-estimate of the weights. In this paper, we interpret these latent noise variables as implicit representations of simple and domain-agnostic data perturbations during training, producing BNNs that perform well under covariate shift due to input corruptions. We observe that the diversity of the implicit corruptions depends on the entropy of the latent variables, and propose a straightforward approach to increase the entropy of these variables during training. We evaluate the method on out-of-distribution image classification benchmarks, and show improved uncertainty

estimation of node-based BNNs under covariate shift due to input perturbations. As a side effect, the method also provides robustness against noisy training labels.

Fenrir: Physics-Enhanced Regression for Initial Value Problems

- Filip Tronarp, Nathanael Bosch, Philipp Hennig
- abstract: We show how probabilistic numerics can be used to convert an initial value problem into a Gauss–Markov process parametrised by the dynamics of the initial value problem. Consequently, the often difficult problem of parameter estimation in ordinary differential equations is reduced to hyper-parameter estimation in Gauss–Markov regression, which tends to be considerably easier. The method's relation and benefits in comparison to classical numerical integration and gradient matching approaches is elucidated. In particular, the method can, in contrast to gradient matching, handle partial observations, and has certain routes for escaping local optima not available to classical numerical integration. Experimental results demonstrate that the method is on par or moderately better than competing approaches.

Interpretable Off-Policy Learning via Hyperbox Search

- Daniel Tschernutter, Tobias Hatt, Stefan Feuerriegel
- abstract: Personalized treatment decisions have become an integral part of modern medicine. Thereby, the aim is to make treatment decisions based on individual patient characteristics. Numerous methods have been developed for learning such policies from observational data that achieve the best outcome across a certain policy class. Yet these methods are rarely interpretable. However, interpretability is often a prerequisite for policy learning in clinical practice. In this paper, we propose an algorithm for interpretable off-policy learning via hyperbox search. In particular, our policies can be represented in disjunctive normal form (i.e., OR-of-ANDs) and are thus intelligible. We prove a universal approximation theorem that shows that our policy class is flexible enough to approximate any measurable function arbitrarily well. For optimization, we develop a tailored column generation procedure within a branch-and-bound framework. Using a simulation study, we demonstrate that our algorithm outperforms state-of-the-art methods from interpretable off-policy learning in terms of regret. Using real-word clinical data, we perform a user study with actual clinical experts, who rate our policies as highly interpretable.

FriendlyCore: Practical Differentially Private Aggregation

- Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, Uri Stemmer
- abstract: Differentially private algorithms for common metric aggregation tasks, such as clustering or averaging, often have limited practicality due to their complexity or to the large number of data points that is required for accurate results. We propose a simple and practical tool $\mathsf{FriendlyCore}$ that takes a set of points ${\cal D}$ from an unrestricted (pseudo) metric space as input. When ${\cal D}$ has effective diameter r, $\mathsf{FriendlyCore}$ returns a "stable" subset ${\cal C} \subseteq {\cal D}$ that includes all points, except possibly few outliers, and is guaranteed to have diameter r. $\mathsf{FriendlyCore}$ can be used to preprocess the input before privately aggregating it, potentially simplifying the aggregation or boosting its accuracy. Surprisingly, $\mathsf{FriendlyCore}$ is light-weight with no dependence on the dimension. We empirically demonstrate its advantages in boosting the accuracy of mean estimation and clustering tasks such as k-means and k-GMM, outperforming tailored methods.

Pairwise Conditional Gradients without Swap Steps and Sparser Kernel Herding

- Kazuma K Tsuji, Ken'Ichiro Tanaka, Sebastian Pokutta
- abstract: The Pairwise Conditional Gradients (PCG) algorithm is a powerful extension of the Frank-Wolfe algorithm leading to particularly sparse solutions, which makes PCG very appealing for problems such as sparse signal recovery, sparse regression, and kernel herding. Unfortunately, PCG exhibits so-called swap steps that might not provide sufficient primal progress. The number of these bad steps is bounded by a function in the dimension and as such known guarantees do not generalize to the infinite-dimensional case, which would be needed for kernel herding. We propose a new variant of PCG, the so-called Blended Pairwise Conditional Gradients (BPCG). This new algorithm does not exhibit any swap steps, is very easy to implement, and does not require any internal gradient alignment procedures. The convergence rate of BPCG is basically that of PCG if no drop steps would occur and as such is no worse than PCG but improves and provides new rates in many cases. Moreover, we observe in the numerical experiments that BPCG's solutions are much sparser than those of PCG. We apply BPCG to the kernel herding setting, where we derive nice quadrature rules and provide numerical results demonstrating the performance of our method.

Prototype Based Classification from Hierarchy to Fairness

- Mycal Tucker, Julie A. Shah
- abstract: Artificial neural nets can represent and classify many types of high-dimensional data but are often tailored to particular applications – e.g., for "fair" or "hierarchical" classification. Once an architecture has been selected, it is often difficult for humans to adjust models for a new task; for example, a hierarchical classifier cannot be easily transformed into a fair classifier that shields a protected field. Our contribution in this work is a new neural network architecture, the concept subspace network (CSN), which generalizes existing specialized classifiers to produce a unified model capable of learning a spectrum of multi-concept relationships. We demonstrate that CSNs reproduce state-of-the-art results in fair classification when enforcing concept independence, may be transformed into hierarchical classifiers, or may even reconcile fairness and hierarchy within a single classifier. The CSN is inspired by and matches the performance of existing prototype-based classifiers that promote interpretability.

Consensus Multiplicative Weights Update: Learning to Learn using Projector-based Game Signatures

- Nelson Vadori, Rahul Savani, Thomas Spooner, Sumitra Ganesh
- abstract: Cheung and Piliouras (2020) recently showed that two variants of the Multiplicative Weights Update method - OMWU and MWU - display opposite convergence properties depending on whether the game is zero-sum or cooperative. Inspired by this work and the recent literature on learning to optimize for single functions, we introduce a new framework for learning last-iterate convergence to Nash Equilibria in games, where the update rule's coefficients (learning rates) along a trajectory are learnt by a reinforcement learning policy that is conditioned on the nature of the game: the game signature. We construct the latter using a new decomposition of two-player games into eight components corresponding to commutative projection operators, generalizing and unifying recent game concepts studied in the literature. We compare the performance of various update rules when their coefficients are learnt, and show that the RL policy is able to exploit the game signature across a wide range of game types. In doing so, we introduce CMWU, a new algorithm that extends consensus optimization to the constrained case, has local convergence guarantees for zero-sum bimatrix games, and show that it enjoys competitive performance on both zero-sum games with constant coefficients and across a spectrum of games when its coefficients are learnt.

Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech

- Aditya R Vaidya, Shailee Jain, Alexander Huth
- abstract: Self-supervised language models are very effective at predicting high-level cortical responses during language comprehension. However, the best current models of lower-level auditory processing in the human brain rely on either hand-constructed acoustic filters or representations from supervised

audio neural networks. In this work, we capitalize on the progress of self-supervised speech representation learning (SSL) to create new state-of-the-art models of the human auditory system. Compared against acoustic baselines, phonemic features, and supervised models, representations from the middle layers of self-supervised models (APC, wav2vec, wav2vec 2.0, and HuBERT) consistently yield the best prediction performance for fMRI recordings within the auditory cortex (AC). Brain areas involved in low-level auditory processing exhibit a preference for earlier SSL model layers, whereas higher-level semantic areas prefer later layers. We show that these trends are due to the models' ability to encode information at multiple linguistic levels (acoustic, phonetic, and lexical) along their representation depth. Overall, these results show that self-supervised models effectively capture the hierarchy of information relevant to different stages of speech processing in human cortex.

Path-Gradient Estimators for Continuous Normalizing Flows

- Lorenz Vaitl, Kim Andrea Nicoli, Shinichi Nakajima, Pan Kessel
- abstract: Recent work has established a path-gradient estimator for simple variational Gaussian distributions and has argued that the path-gradient is particularly beneficial in the regime in which the variational distribution approaches the exact target distribution. In many applications, this regime can however not be reached by a simple Gaussian variational distribution. In this work, we overcome this crucial limitation by proposing a path-gradient estimator for the considerably more expressive variational family of continuous normalizing flows. We outline an efficient algorithm to calculate this estimator and establish its superior performance empirically.

Improved Convergence Rates for Sparse Approximation Methods in Kernel-Based Learning

- Sattar Vakili, Jonathan Scarlett, Da-Shan Shiu, Alberto Bernacchia
- abstract: Kernel-based models such as kernel ridge regression and Gaussian processes are ubiquitous in machine learning applications for regression and optimization. It is well known that a major downside for kernel-based models is the high computational cost; given a dataset of n samples, the cost grows as $\mathcal{O}(n^3)$. Existing sparse approximation methods can yield a significant reduction in the computational cost, effectively reducing the actual cost down to as low as $\mathcal{O}(n)$ in certain cases. Despite this remarkable empirical success, significant gaps remain in the existing results for the analytical bounds on the error due to approximation. In this work, we provide novel confidence intervals for the Nyström method and the sparse variational Gaussian process approximation method, which we establish using novel interpretations of the approximate (surrogate) posterior variance of the models. Our confidence intervals lead to improved performance bounds in both regression and optimization problems.

EDEN: Communication-Efficient and Robust Distributed Mean Estimation for Federated Learning

- Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, Michael Mitzenmacher
- abstract: Distributed Mean Estimation (DME) is a central building block in federated learning, where clients send local gradients to a parameter server for averaging and updating the model. Due to communication constraints, clients often use lossy compression techniques to compress the gradients, resulting in estimation inaccuracies. DME is more challenging when clients have diverse network conditions, such as constrained communication budgets and packet losses. In such settings, DME techniques often incur a significant increase in the estimation error leading to degraded learning performance. In this work, we propose a robust DME technique named EDEN that naturally handles heterogeneous communication budgets and packet losses. We derive appealing theoretical guarantees for EDEN and evaluate it empirically. Our results demonstrate that EDEN consistently improves over state-of-the-art DME techniques.

Towards Noise-adaptive, Problem-adaptive (Accelerated) Stochastic Gradient Descent

- Sharan Vaswani, Benjamin Dubois-Taine, Reza Babanezhad
- abstract: We aim to make stochastic gradient descent (SGD) adaptive to (i) the noise σ^2 in the stochastic gradients and (ii) problem-dependent constants. When minimizing smooth, strongly-convex functions with condition number κ, we prove that T iterations of SGD with exponentially decreasing step-sizes and knowledge of the smoothness can achieve an $\tilde{O} \left(\exp \left(\nicefrac{-T}{\kappa} \right) + \nicefrac{\sigma^2}{T} \right)$ rate, without knowing σ^2. In order to be adaptive to the smoothness, we use a stochastic line-search (SLS) and show (via upper and lower-bounds) that SGD with SLS converges at the desired rate, but only to a neighbourhood of the solution. On the other hand, we prove that SGD with an offline estimate of the smoothness converges to the minimizer. However, its rate is slowed down proportional to the estimation error. Next, we prove that SGD with Nesterov acceleration and exponential step-sizes (referred to as ASGD) can achieve the near-optimal $\tilde{O} \left(\exp \left(\nicefrac{-T}{\sqrt{\kappa}} \right) + \nicefrac{\sigma^2}{T} \right)$ rate, without knowledge of σ^2. When used with offline estimates of the smoothness and strong-convexity, ASGD still converges to the solution, albeit at a slower rate. Finally, we empirically demonstrate the effectiveness of exponential step-sizes coupled with a novel variant of SLS.

Correlation Clustering via Strong Triadic Closure Labeling: Fast Approximation Algorithms and Practical Lower Bounds

- Nate Veldt
- abstract: Correlation clustering is a widely studied framework for clustering based on pairwise similarity and dissimilarity scores, but its best approximation algorithms rely on impractical linear programming relaxations. We present faster approximation algorithms that avoid these relaxations, for two well-studied special cases: cluster editing and cluster deletion. We accomplish this by drawing new connections to edge labeling problems related to the principle of strong triadic closure. This leads to faster and more practical linear programming algorithms, as well as extremely scalable combinatorial techniques, including the first combinatorial approximation algorithm for cluster deletion. In practice, our algorithms produce approximate solutions that nearly match the best algorithms in quality, while scaling to problems that are orders of magnitude larger.

The CLRS Algorithmic Reasoning Benchmark

- Petar Veličković, Adrià Puigdomènech Badia, David Budden, Razvan Pascanu, Andrea Banino, Misha Dashevskiy, Raia Hadsell, Charles Blundell
- abstract: Learning representations of algorithms is an emerging area of machine learning, seeking to bridge concepts from neural networks with classical algorithms. Several important works have investigated whether neural networks can effectively reason like algorithms, typically by learning to execute them. The common trend in the area, however, is to generate targeted kinds of algorithmic data to evaluate specific hypotheses, making results hard to transfer across publications, and increasing the barrier of entry. To consolidate progress and work towards unified evaluation, we propose the CLRS Algorithmic Reasoning Benchmark, covering classical algorithms from the Introduction to Algorithms textbook. Our benchmark spans a variety of algorithmic reasoning procedures, including sorting, searching, dynamic programming, graph algorithms, string algorithms and geometric algorithms. We perform extensive experiments to demonstrate how several popular algorithmic reasoning baselines perform on these tasks, and consequently, highlight links to several open challenges. Our library is readily available at https://github.com/deepmind/clrs.

Bregman Power k-Means for Clustering Exponential Family Data

- Adithya Vellal, Saptarshi Chakraborty, Jason Q Xu
- abstract: Recent progress in center-based clustering algorithms combats poor local minima by implicit annealing through a family of generalized means. These methods are variations of Lloyd's celebrated k-means algorithm, and are most appropriate for spherical clusters such as those arising from Gaussian data. In this paper, we bridge these new algorithmic advances to classical work on hard clustering under Bregman divergences, which enjoy a bijection to

exponential family distributions and are thus well-suited for clustering objects arising from a breadth of data generating mechanisms. The elegant properties of Bregman divergences allow us to maintain closed form updates in a simple and transparent algorithm, and moreover lead to new theoretical arguments for establishing finite sample bounds that relax the bounded support assumption made in the existing state of the art. Additionally, we consider thorough empirical analyses on simulated experiments and a case study on rainfall data, finding that the proposed method outperforms existing peer methods in a variety of non-Gaussian data settings.

Estimation in Rotationally Invariant Generalized Linear Models via Approximate Message Passing

- Ramji Venkataramanan, Kevin Kögler, Marco Mondelli
- abstract: We consider the problem of signal estimation in generalized linear models defined via rotationally invariant design matrices. Since these matrices can have an arbitrary spectral distribution, this model is well suited for capturing complex correlation structures which often arise in applications. We propose a novel family of approximate message passing (AMP) algorithms for signal estimation, and rigorously characterize their performance in the high-dimensional limit via a state evolution recursion. Our rotationally invariant AMP has complexity of the same order as the existing AMP derived under the restrictive assumption of a Gaussian design; our algorithm also recovers this existing AMP as a special case. Numerical results showcase a performance close to Vector AMP (which is conjectured to be Bayes-optimal in some settings), but obtained with a much lower complexity, as the proposed algorithm does not require a computationally expensive singular value decomposition.

Bayesian Optimization under Stochastic Delayed Feedback

- Arun Verma, Zhongxiang Dai, Bryan Kian Hsiang Low
- abstract: Bayesian optimization (BO) is a widely-used sequential method for zeroth-order optimization of complex and expensive-to-compute black-box functions. The existing BO methods assume that the function evaluation (feedback) is available to the learner immediately or after a fixed delay. Such assumptions may not be practical in many real-life problems like online recommendations, clinical trials, and hyperparameter tuning where feedback is available after a random delay. To benefit from the experimental parallelization in these problems, the learner needs to start new function evaluations without waiting for delayed feedback. In this paper, we consider the BO under stochastic delayed feedback problem. We propose algorithms with sub-linear regret guarantees that efficiently address the dilemma of selecting new function queries while waiting for randomly delayed feedback. Building on our results, we also make novel contributions to batch BO and contextual Gaussian process bandits. Experiments on synthetic and real-life datasets verify the performance of our algorithms.

VarScene: A Deep Generative Model for Realistic Scene Graph Synthesis

- Tathagat Verma, Abir De, Yateesh Agrawal, Vishwa Vinay, Soumen Chakrabarti
- abstract: Scene graphs are powerful abstractions that capture relationships between objects in images by modeling objects as nodes and relationships as edges. Generation of realistic synthetic scene graphs has applications like scene synthesis and data augmentation for supervised learning. Existing graph generative models are predominantly targeted toward molecular graphs, leveraging the limited vocabulary of atoms and bonds and also the well-defined semantics of chemical compounds. In contrast, scene graphs have much larger object and relation vocabularies, and their semantics are latent. To address this challenge, we propose a variational autoencoder for scene graphs, which is optimized for the maximum mean discrepancy (MMD) between the ground truth scene graph distribution and distribution of the generated scene graphs. Our method views a scene graph as a collection of star graphs and encodes it into a latent representation of the underlying stars. The decoder generates scene graphs by learning to sample the component stars and edges between them. Our experiments show that our method is able to mimic the underlying scene graph generative process more accurately than several state-of-the-art baselines.

Calibrated Learning to Defer with One-vs-All Classifiers

- Rajeev Verma, Eric Nalisnick
- abstract: The learning to defer (L2D) framework has the potential to make AI systems safer. For a given input, the system can defer the decision to a human if the human is more likely than the model to take the correct action. We study the calibration of L2D systems, investigating if the probabilities they output are sound. We find that Mozannar & Sontag's (2020) multiclass framework is not calibrated with respect to expert correctness. Moreover, it is not even guaranteed to produce valid probabilities due to its parameterization being degenerate for this purpose. We propose an L2D system based on one-vs-all classifiers that is able to produce calibrated probabilities of expert correctness. Furthermore, our loss function is also a consistent surrogate for multiclass L2D, like Mozannar & Sontag's (2020). Our experiments verify that not only is our system calibrated, but this benefit comes at no cost to accuracy. Our model's accuracy is always comparable (and often superior) to Mozannar & Sontag's (2020) model's in tasks ranging from hate speech detection to galaxy classification to diagnosis of skin lesions.

Regret Bounds for Stochastic Shortest Path Problems with Linear Function Approximation

- Daniel Vial, Advait Parulekar, Sanjay Shakkottai, R Srikant
- abstract: We propose an algorithm that uses linear function approximation (LFA) for stochastic shortest path (SSP). Under minimal assumptions, it obtains sublinear regret, is computationally efficient, and uses stationary policies. To our knowledge, this is the first such algorithm in the LFA literature (for SSP or other formulations). Our algorithm is a special case of a more general one, which achieves regret square root in the number of episodes given access to a computation oracle.

On Implicit Bias in Overparameterized Bilevel Optimization

- Paul Vicol, Jonathan P Lorraine, Fabian Pedregosa, David Duvenaud, Roger B Grosse
- abstract: Many problems in machine learning involve bilevel optimization (BLO), including hyperparameter optimization, meta-learning, and dataset distillation. Bilevel problems involve inner and outer parameters, each optimized for its own objective. Often, at least one of the two levels is underspecified and there are multiple ways to choose among equivalent optima. Inspired by recent studies of the implicit bias induced by optimization algorithms in single-level optimization, we investigate the implicit bias of different gradient-based algorithms for jointly optimizing the inner and outer parameters. We delineate two standard BLO methods—cold-start and warm-start BLO—and show that the converged solution or long-run behavior depends to a large degree on these and other algorithmic choices, such as the hypergradient approximation. We also show that the solutions from warm-start BLO can encode a surprising amount of information about the outer objective, even when the outer optimization variables are low-dimensional. We believe that implicit bias deserves as central a role in the study of bilevel optimization as it has attained in the study of single-level neural net optimization.

Multiclass learning with margin: exponential rates with no bias-variance trade-off

- Stefano Vigogna, Giacomo Meanti, Ernesto De Vito, Lorenzo Rosasco
- abstract: We study the behavior of error bounds for multiclass classification under suitable margin conditions. For a wide variety of methods we prove that the classification error under a hard-margin condition decreases exponentially fast without any bias-variance trade-off. Different convergence rates can be obtained in correspondence of different margin assumptions. With a self-contained and instructive analysis we are able to generalize known results from the binary to the multiclass setting.

Addressing Optimism Bias in Sequence Modeling for Reinforcement Learning

- Adam R Villaflor, Zhe Huang, Swapnil Pande, John M Dolan, Jeff Schneider
- abstract: Impressive results in natural language processing (NLP) based on the Transformer neural network architecture have inspired researchers to explore viewing offline reinforcement learning (RL) as a generic sequence modeling problem. Recent works based on this paradigm have achieved state-of-the-art results in several of the mostly deterministic offline Atari and D4RL benchmarks. However, because these methods jointly model the states and actions as a single sequencing problem, they struggle to disentangle the effects of the policy and world dynamics on the return. Thus, in adversarial or stochastic environments, these methods lead to overly optimistic behavior that can be dangerous in safety-critical systems like autonomous driving. In this work, we propose a method that addresses this optimism bias by explicitly disentangling the policy and world models, which allows us at test time to search for policies that are robust to multiple possible futures in the environment. We demonstrate our method's superior performance on a variety of autonomous driving tasks in simulation.

Bayesian Nonparametrics for Offline Skill Discovery

- Valentin Villecroze, Harry Braviner, Panteha Naderian, Chris Maddison, Gabriel Loaiza-Ganem
- abstract: Skills or low-level policies in reinforcement learning are temporally extended actions that can speed up learning and enable complex behaviours. Recent work in offline reinforcement learning and imitation learning has proposed several techniques for skill discovery from a set of expert trajectories. While these methods are promising, the number K of skills to discover is always a fixed hyperparameter, which requires either prior knowledge about the environment or an additional parameter search to tune it. We first propose a method for offline learning of options (a particular skill framework) exploiting advances in variational inference and continuous relaxations. We then highlight an unexplored connection between Bayesian nonparametrics and offline skill discovery, and show how to obtain a nonparametric version of our model. This version is tractable thanks to a carefully structured approximate posterior with a dynamically-changing number of options, removing the need to specify K. We also show how our nonparametric extension can be applied in other skill frameworks, and empirically demonstrate that our method can outperform state-of-the-art offline skill learning algorithms across a variety of environments.

Hermite Polynomial Features for Private Data Generation

- Margarita Vinaroz, Mohammad-Amin Charusaie, Frederik Harder, Kamil Adamczewski, Mi Jung Park
- abstract: Kernel mean embedding is a useful tool to compare probability measures. Despite its usefulness, kernel mean embedding considers infinite-dimensional features, which are challenging to handle in the context of differentially private data generation. A recent work, DP-MERF (Harder et al., 2021), proposes to approximate the kernel mean embedding of data distribution using finite-dimensional random features, which yields an analytically tractable sensitivity of approximate kernel mean embedding. However, the required number of random features in DP-MERF is excessively high, often ten thousand to a hundred thousand, which worsens the sensitivity of the approximate kernel mean embedding. To improve the sensitivity, we propose to replace random features with Hermite polynomial features. Unlike the random features, the Hermite polynomial features are ordered, where the features at the low orders contain more information on the distribution than those at the high orders. Hence, a relatively low order of Hermite polynomial features can more accurately approximate the mean embedding of the data distribution compared to a significantly higher number of random features. As a result, the Hermite polynomial features help us to improve the privacy-accuracy trade-off compared to DP-MERF, as demonstrated on several heterogeneous tabular datasets, as well as several image benchmark datasets.

What Can Linear Interpolation of Neural Network Loss Landscapes Tell Us?

- Tiffany J Vlaar, Jonathan Frankle
- abstract: Studying neural network loss landscapes provides insights into the nature of the underlying optimization problems. Unfortunately, loss landscapes are notoriously difficult to visualize in a human-comprehensible fashion. One common way to address this problem is to plot linear slices of the landscape, for example from the initial state of the network to the final state after optimization. On the basis of this analysis, prior work has drawn broader conclusions about the difficulty of the optimization problem. In this paper, we put inferences of this kind to the test, systematically evaluating how linear interpolation and final performance vary when altering the data, choice of initialization, and other optimizer and architecture design choices. Further, we use linear interpolation to study the role played by individual layers and substructures of the network. We find that certain layers are more sensitive to the choice of initialization, but that the shape of the linear path is not indicative of the changes in test accuracy of the model. Our results cast doubt on the broader intuition that the presence or absence of barriers when interpolating necessarily relates to the success of optimization.

Multirate Training of Neural Networks

- Tiffany J Vlaar, Benedict Leimkuhler
- abstract: We propose multirate training of neural networks: partitioning neural network parameters into "fast" and "slow" parts which are trained on different time scales, where slow parts are updated less frequently. By choosing appropriate partitionings we can obtain substantial computational speed-up for transfer learning tasks. We show for applications in vision and NLP that we can fine-tune deep neural networks in almost half the time, without reducing the generalization performance of the resulting models. We analyze the convergence properties of our multirate scheme and draw a comparison with vanilla SGD. We also discuss splitting choices for the neural network parameters which could enhance generalization performance when neural networks are trained from scratch. A multirate approach can be used to learn different features present in the data and as a form of regularization. Our paper unlocks the potential of using multirate techniques for neural network training and provides several starting points for future work in this area.

Provably Adversarially Robust Nearest Prototype Classifiers

- Václav Voráček, Matthias Hein
- abstract: Nearest prototype classifiers (NPCs) assign to each input point the label of the nearest prototype with respect to a chosen distance metric. A direct advantage of NPCs is that the decisions are interpretable. Previous work could provide lower bounds on the minimal adversarial perturbation in the ℓ_p-threat model when using the same ℓ_p-distance for the NPCs. In this paper we provide a complete discussion on the complexity when using ℓ_p-distances for decision and ℓ_q-threat models for certification for $p,q \in {1,2,\infty}$. In particular we provide scalable algorithms for the exact computation of the minimal adversarial perturbation when using ℓ_2-distance and improved lower bounds in other cases. Using efficient improved lower bounds we train our \textbf{P}rovably adversarially robust \textbf{NPC} (PNPC), for MNIST which have better ℓ_2-robustness guarantees than neural networks. Additionally, we show up to our knowledge the first certification results w.r.t. to the LPIPS perceptual metric which has been argued to be a more realistic threat model for image classification than ℓ_p-balls. Our PNPC has on CIFAR10 higher certified robust accuracy than the empirical robust accuracy reported in \cite{laidlaw2021perceptual}. The code is available in our \href{https://github.com/vvoracek/Provably-Adversarially-Robust-Nearest-Prototype-Classifiers}{repository}.

First-Order Regret in Reinforcement Learning with Linear Function Approximation: A Robust Estimation Approach

- Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, Kevin Jamieson
- abstract: Obtaining first-order regret bounds—regret bounds scaling not as the worst-case but with some measure of the performance of the optimal policy on a given instance—is a core question in sequential decision-making. While such bounds exist in many settings, they have proven elusive in

reinforcement learning with large state spaces. In this work we address this gap, and show that it is possible to obtain regret scaling as $\widetilde{\mathcal{O}}(\sqrt{d^3 H^3 \cdot V_1^\star \cdot K} + d^{3.5}H^3\log K)$ in reinforcement learning with large state spaces, namely the linear MDP setting. Here $V_1^\star$ is the value of the optimal policy and K is the number of episodes. We demonstrate that existing techniques based on least squares estimation are insufficient to obtain this result, and instead develop a novel robust self-normalized concentration bound based on the robust Catoni mean estimator, which may be of independent interest.

Reward-Free RL is No Harder Than Reward-Aware RL in Linear Markov Decision Processes

- Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, Kevin Jamieson
- abstract: Reward-free reinforcement learning (RL) considers the setting where the agent does not have access to a reward function during exploration, but must propose a near-optimal policy for an arbitrary reward function revealed only after exploring. In the the tabular setting, it is well known that this is a more difficult problem than reward-aware (PAC) RL—where the agent has access to the reward function during exploration—with optimal sample complexities in the two settings differing by a factor of $|\mathcal{S}|$, the size of the state space. We show that this separation does not exist in the setting of linear MDPs. We first develop a computationally efficient algorithm for reward-free RL in a d-dimensional linear MDP with sample complexity scaling as $\widetilde{\mathcal{O}}(d^2 H^5/\epsilon^2)$. We then show a lower bound with matching dimension-dependence of $\Omega(d^2 H^2/\epsilon^2)$, which holds for the reward-aware RL setting. To our knowledge, our approach is the first computationally efficient algorithm to achieve optimal d dependence in linear MDPs, even in the single-reward PAC setting. Our algorithm relies on a novel procedure which efficiently traverses a linear MDP, collecting samples in any given “feature direction”, and enjoys a sample complexity scaling optimally in the (linear MDP equivalent of the) maximal state visitation probability. We show that this exploration procedure can also be applied to solve the problem of obtaining “well-conditioned” covariates in linear MDPs.

Training Characteristic Functions with Reinforcement Learning: XAI-methods play Connect Four

- Stephan Wäldchen, Sebastian Pokutta, Felix Huber
- abstract: Characteristic functions (from cooperative game theory) are able to evaluate partial inputs and form the basis for attribution methods like Shapley values. These attribution methods allow us to measure how important each input component is for the function output—one of the goals of explainable AI (XAI). Given a standard classifier function, it is unclear how partial input should be realised. Instead, most XAI-methods for black-box classifiers like neural networks consider counterfactual inputs that generally lie off-manifold, which makes them hard to evaluate and easy to manipulate. We propose a setup to directly train characteristic functions in the form of neural networks to play simple two-player games. We apply this to the game of Connect Four by randomly hiding colour information from our agents during training. This has three advantages for comparing XAI-methods: It alleviates the ambiguity about how to realise partial input, makes off-manifold evaluation unnecessary and allows us to compare the methods by letting them play against each other.

Retroformer: Pushing the Limits of End-to-end Retrosynthesis Transformer

- Yue Wan, Chang-Yu Hsieh, Ben Liao, Shengyu Zhang
- abstract: Retrosynthesis prediction is one of the fundamental challenges in organic synthesis. The task is to predict the reactants given a core product. With the advancement of machine learning, computer-aided synthesis planning has gained increasing interest. Numerous methods were proposed to solve this problem with different levels of dependency on additional chemical knowledge. In this paper, we propose Retroformer, a novel Transformer-based architecture for retrosynthesis prediction without relying on any cheminformatics tools for molecule editing. Via the proposed local attention head, the model can jointly encode the molecular sequence and graph, and efficiently exchange information between the local reactive region and the global reaction context. Retroformer reaches the new state-of-the-art accuracy for the end-to-end template-free retrosynthesis, and improves over many strong baselines on better molecule and reaction validity. In addition, its generative procedure is highly interpretable and controllable. Overall, Retroformer pushes the limits of the reaction reasoning ability of deep generative models.

Safe Exploration for Efficient Policy Evaluation and Comparison

- Runzhe Wan, Branislav Kveton, Rui Song
- abstract: High-quality data plays a central role in ensuring the accuracy of policy evaluation. This paper initiates the study of efficient and safe data collection for bandit policy evaluation. We formulate the problem and investigate its several representative variants. For each variant, we analyze its statistical properties, derive the corresponding exploration policy, and design an efficient algorithm for computing it. Both theoretical analysis and experiments support the usefulness of the proposed methods.

Greedy based Value Representation for Optimal Coordination in Multi-agent Reinforcement Learning

- Lipeng Wan, Zeyang Liu, Xingyu Chen, Xuguang Lan, Nanning Zheng
- abstract: Due to the representation limitation of the joint Q value function, multi-agent reinforcement learning methods with linear value decomposition (LVD) or monotonic value decomposition (MVD) suffer from relative overgeneralization. As a result, they can not ensure optimal consistency (i.e., the correspondence between individual greedy actions and the best team performance). In this paper, we derive the expression of the joint Q value function of LVD and MVD. According to the expression, we draw a transition diagram, where each self-transition node (STN) is a possible convergence. To ensure the optimal consistency, the optimal node is required to be the unique STN. Therefore, we propose the greedy-based value representation (GVR), which turns the optimal node into an STN via inferior target shaping and eliminates the non-optimal STNs via superior experience replay. Theoretical proofs and empirical results demonstrate that given the true Q values, GVR ensures the optimal consistency under sufficient exploration. Besides, in tasks where the true Q values are unavailable, GVR achieves an adaptive trade-off between optimality and stability. Our method outperforms state-of-the-art baselines in experiments on various benchmarks.

Towards Evaluating Adaptivity of Model-Based Reinforcement Learning Methods

- Yi Wan, Ali Rahimi-Kalahroudi, Janarthanan Rajendran, Ida Momennejad, Sarath Chandar, Harm H Van Seijen
- abstract: In recent years, a growing number of deep model-based reinforcement learning (RL) methods have been introduced. The interest in deep model-based RL is not surprising, given its many potential benefits, such as higher sample efficiency and the potential for fast adaption to changes in the environment. However, we demonstrate, using an improved version of the recently introduced Local Change Adaptation (LoCA) setup, that well-known model-based methods such as PlaNet and DreamerV2 perform poorly in their ability to adapt to local environmental changes. Combined with prior work that made a similar observation about the other popular model-based method, MuZero, a trend appears to emerge, suggesting that current deep model-based methods have serious limitations. We dive deeper into the causes of this poor performance, by identifying elements that hurt adaptive behavior and linking these to underlying techniques frequently used in deep model-based RL. We empirically validate these insights in the case of linear function approximation by demonstrating that a modified version of linear Dyna achieves effective adaptation to local changes. Furthermore, we provide detailed insights into the challenges of building an adaptive nonlinear model-based method, by experimenting with a nonlinear version of Dyna.

Fast Lossless Neural Compression with Integer-Only Discrete Flows

- Siyu Wang, Jianfei Chen, Chongxuan Li, Jun Zhu, Bo Zhang
- abstract: By applying entropy codecs with learned data distributions, neural compressors have significantly outperformed traditional codecs in terms of compression ratio. However, the high inference latency of neural networks hinders the deployment of neural compressors in practical applications. In this work, we propose Integer-only Discrete Flows (IODF) an efficient neural compressor with integer-only arithmetic. Our work is built upon integer discrete flows, which consists of invertible transformations between discrete random variables. We propose efficient invertible transformations with integer-only arithmetic based on 8-bit quantization. Our invertible transformation is equipped with learnable binary gates to remove redundant filters during inference. We deploy IODF with TensorRT on GPUs, achieving $10\times$ inference speedup compared to the fastest existing neural compressors, while retaining the high compression rates on ImageNet32 and ImageNet64.

Accelerating Shapley Explanation via Contributive Cooperator Selection

- Guanchu Wang, Yu-Neng Chuang, Mengnan Du, Fan Yang, Quan Zhou, Pushkar Tripathi, Xuanting Cai, Xia Hu
- abstract: Even though Shapley value provides an effective explanation for a DNN model prediction, the computation relies on the enumeration of all possible input feature coalitions, which leads to the exponentially growing complexity. To address this problem, we propose a novel method SHEAR to significantly accelerate the Shapley explanation for DNN models, where only a few coalitions of input features are involved in the computation. The selection of the feature coalitions follows our proposed Shapley chain rule to minimize the absolute error from the ground-truth Shapley values, such that the computation can be both efficient and accurate. To demonstrate the effectiveness, we comprehensively evaluate SHEAR across multiple metrics including the absolute error from the ground-truth Shapley value, the faithfulness of the explanations, and running speed. The experimental results indicate SHEAR consistently outperforms state-of-the-art baseline methods across different evaluation metrics, which demonstrates its potentials in real-world applications where the computational resource is limited.

Denoised MDPs: Learning World Models Better Than the World Itself

- Tongzhou Wang, Simon Du, Antonio Torralba, Phillip Isola, Amy Zhang, Yuandong Tian
- abstract: The ability to separate signal from noise, and reason with clean abstractions, is critical to intelligence. With this ability, humans can efficiently perform real world tasks without considering all possible nuisance factors. How can artificial agents do the same? What kind of information can agents safely discard as noises? In this work, we categorize information out in the wild into four types based on controllability and relation with reward, and formulate useful information as that which is both controllable and reward-relevant. This framework clarifies the kinds information removed by various prior work on representation learning in reinforcement learning (RL), and leads to our proposed approach of learning a Denoised MDP that explicitly factors out certain noise distractors. Extensive experiments on variants of DeepMind Control Suite and RoboDesk demonstrate superior performance of our denoised world model over using raw observations alone, and over prior works, across policy optimization control tasks as well as the non-control task of joint position regression. Project Page: https://ssnl.github.io/denoised_mdp/ Code: https://github.com/facebookresearch/denoised_mdp/

Neural Implicit Dictionary Learning via Mixture-of-Expert Training

- Peihao Wang, Zhiwen Fan, Tianlong Chen, Zhangyang Wang
- abstract: Representing visual signals by coordinate-based deep fully-connected networks has been shown advantageous in fitting complex details and solving inverse problems than discrete grid-based representation. However, acquiring such a continuous Implicit Neural Representation (INR) requires tedious per-scene training on tons of signal measurements, which limits its practicality. In this paper, we present a generic INR framework that achieves both data and training efficiency by learning a Neural Implicit Dictionary (NID) from a data collection and representing INR as a functional combination of wavelets sampled from the dictionary. Our NID assembles a group of coordinate-based subnetworks which are tuned to span the desired function space. After training, one can instantly and robustly acquire an unseen scene representation by solving the coding coefficients. To parallelly optimize a large group of networks, we borrow the idea from Mixture-of-Expert (MoE) to design and train our network with a sparse gating mechanism. Our experiments show that, NID can improve reconstruction of 2D images or 3D scenes by 2 orders of magnitude faster with up to 98% less input data. We further demonstrate various applications of NID in image inpainting and occlusion removal, which are considered to be challenging with vanilla INR. Our codes are available in https://github.com/VITA-Group/Neural-Implicit-Dict.

Robust Models Are More Interpretable Because Attributions Look Normal

- Zifan Wang, Matt Fredrikson, Anupam Datta
- abstract: Recent work has found that adversarially-robust deep networks used for image classification are more interpretable: their feature attributions tend to be sharper, and are more concentrated on the objects associated with the image’s ground- truth class. We show that smooth decision boundaries play an important role in this enhanced interpretability, as the model’s input gradients around data points will more closely align with boundaries’ normal vectors when they are smooth. Thus, because robust models have smoother boundaries, the results of gradient- based attribution methods, like Integrated Gradients and DeepLift, will capture more accurate information about nearby decision boundaries. This understanding of robust interpretability leads to our second contribution: boundary attributions, which aggregate information about the normal vectors of local decision bound- aries to explain a classification outcome. We show that by leveraging the key fac- tors underpinning robust interpretability, boundary attributions produce sharper, more concentrated visual explanations{—}even on non-robust models.

Disentangling Disease-related Representation from Obscure for Disease Prediction

- Chu-Ran Wang, Fei Gao, Fandong Zhang, Fangwei Zhong, Yizhou Yu, Yizhou Wang
- abstract: Disease-related representations play a crucial role in image-based disease prediction such as cancer diagnosis, due to its considerable generalization capacity. However, it is still a challenge to identify lesion characteristics in obscured images, as many lesions are obscured by other tissues. In this paper, to learn the representations for identifying obscured lesions, we propose a disentanglement learning strategy under the guidance of alpha blending generation in an encoder-decoder framework (DAB-Net). Specifically, we take mammogram mass benign/malignant classification as an example. In our framework, composite obscured mass images are generated by alpha blending and then explicitly disentangled into disease-related mass features and interference glands features. To achieve disentanglement learning, features of these two parts are decoded to reconstruct the mass and the glands with corresponding reconstruction losses, and only disease-related mass features are fed into the classifier for disease prediction. Experimental results on one public dataset DDSM and three in-house datasets demonstrate that the proposed strategy can achieve state-of-the-art performance. DAB-Net achieves substantial improvements of 3.9%~4.4% AUC in obscured cases. Besides, the visualization analysis shows the model can better disentangle the mass and glands in the obscured image, suggesting the effectiveness of our solution in exploring the hidden characteristics in this challenging problem.

Solving Stackelberg Prediction Game with Least Squares Loss via Spherically Constrained Least Squares Reformulation

- Jiali Wang, Wen Huang, Rujun Jiang, Xudong Li, Alex L Wang
- abstract: The Stackelberg prediction game (SPG) is popular in characterizing strategic interactions between a learner and an attacker. As an important special case, the SPG with least squares loss (SPG-LS) has recently received much research attention. Although initially formulated as a difficult bi-level optimization problem, SPG-LS admits tractable reformulations which can be polynomially globally solved by semidefinite programming or second order cone programming. However, all the available approaches are not well-suited for handling large-scale datasets, especially those with huge numbers of features. In this paper, we explore an alternative reformulation of the SPG-LS. By a novel nonlinear change of variables, we rewrite the SPG-LS as a spherically constrained least squares (SCLS) problem. Theoretically, we show that an ϵ optimal solutions to the SCLS (and the SPG-LS) can be

achieved in $\tilde O(N/\sqrt{\epsilon})$ floating-point operations, where N is the number of nonzero entries in the data matrix. Practically, we apply two well-known methods for solving this new reformulation, i.e., the Krylov subspace method and the Riemannian trust region method. Both algorithms are factorization free so that they are suitable for solving large scale problems. Numerical results on both synthetic and real-world datasets indicate that the SPG-LS, equipped with the SCLS reformulation, can be solved orders of magnitude faster than the state of the art.

VLMixer: Unpaired Vision-Language Pre-training via Cross-Modal CutMix

- Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, Ping Luo
- abstract: Existing vision-language pre-training (VLP) methods primarily rely on paired image-text datasets, which are either annotated by enormous human labors or crawled from the internet followed by elaborate data cleaning techniques. To reduce the dependency on well-aligned image-text pairs, it is promising to directly leverage the large-scale text-only and image-only corpora. This paper proposes a data augmentation method, namely cross-modal CutMix (CMC), for implicit cross-modal alignment learning in unpaired VLP. Specifically, CMC transforms natural sentences in the textual view into a multi-modal view, where visually-grounded words in a sentence are randomly replaced by diverse image patches with similar semantics. There are several appealing proprieties of the proposed CMC. First, it enhances the data diversity while keeping the semantic meaning intact for tackling problems where the aligned data are scarce; Second, by attaching cross-modal noise on uni-modal data, it guides models to learn token-level interactions across modalities for better denoising. Furthermore, we present a new unpaired VLP method, dubbed as VLMixer, that integrates CMC with contrastive learning to pull together the uni-modal and multi-modal views for better instance-level alignments among different modalities. Extensive experiments on five downstream tasks show that VLMixer could surpass previous state-of-the-art unpaired VLP methods.

DynaMixer: A Vision MLP Architecture with Dynamic Mixing

- Ziyu Wang, Wenhao Jiang, Yiming M Zhu, Li Yuan, Yibing Song, Wei Liu
- abstract: Recently, MLP-like vision models have achieved promising performances on mainstream visual recognition tasks. In contrast with vision transformers and CNNs, the success of MLP-like models shows that simple information fusion operations among tokens and channels can yield a good representation power for deep recognition models. However, existing MLP-like models fuse tokens through static fusion operations, lacking adaptability to the contents of the tokens to be mixed. Thus, customary information fusion procedures are not effective enough. To this end, this paper presents an efficient MLP-like network architecture, dubbed DynaMixer, resorting to dynamic information fusion. Critically, we propose a procedure, on which the DynaMixer model relies, to dynamically generate mixing matrices by leveraging the contents of all the tokens to be mixed. To reduce the time complexity and improve the robustness, a dimensionality reduction technique and a multi-segment fusion mechanism are adopted. Our proposed DynaMixer model (97M parameters) achieves 84.3% top-1 accuracy on the ImageNet-1K dataset without extra training data, performing favorably against the state-of-the-art vision MLP models. When the number of parameters is reduced to 26M, it still achieves 82.7% top-1 accuracy, surpassing the existing MLP-like models with a similar capacity. The code is available at \url{https://github.com/ziyuwwang/DynaMixer}.

Improving Screening Processes via Calibrated Subset Selection

- Lequn Wang, Thorsten Joachims, Manuel Gomez Rodriguez
- abstract: Many selection processes such as finding patients qualifying for a medical trial or retrieval pipelines in search engines consist of multiple stages, where an initial screening stage focuses the resources on shortlisting the most promising candidates. In this paper, we investigate what guarantees a screening classifier can provide, independently of whether it is constructed manually or trained. We find that current solutions do not enjoy distribution-free theoretical guarantees and we show that, in general, even for a perfectly calibrated classifier, there always exist specific pools of candidates for which its shortlist is suboptimal. Then, we develop a distribution-free screening algorithm—called Calibrated Subsect Selection (CSS)—that, given any classifier and some amount of calibration data, finds near-optimal shortlists of candidates that contain a desired number of qualified candidates in expectation. Moreover, we show that a variant of CSS that calibrates a given classifier multiple times across specific groups can create shortlists with provable diversity guarantees. Experiments on US Census survey data validate our theoretical results and show that the shortlists provided by our algorithm are superior to those provided by several competitive baselines.

The Geometry of Robust Value Functions

- Kaixin Wang, Navdeep Kumar, Kuangqi Zhou, Bryan Hooi, Jiashi Feng, Shie Mannor
- abstract: The space of value functions is a fundamental concept in reinforcement learning. Characterizing its geometric properties may provide insights for optimization and representation. Existing works mainly focus on the value space for Markov Decision Processes (MDPs). In this paper, we study the geometry of the robust value space for the more general Robust MDPs (RMDPs) setting, where transition uncertainties are considered. Specifically, since we find it hard to directly adapt prior approaches to RMDPs, we start with revisiting the non-robust case, and introduce a new perspective that enables us to characterize both the non-robust and robust value space in a similar fashion. The key of this perspective is to decompose the value space, in a state-wise manner, into unions of hypersurfaces. Through our analysis, we show that the robust value space is determined by a set of conic hypersurfaces, each of which contains the robust values of all policies that agree on one state. Furthermore, we find that taking only extreme points in the uncertainty set is sufficient to determine the robust value space. Finally, we discuss some other aspects about the robust value space, including its non-convexity and policy agreement on multiple states.

What Dense Graph Do You Need for Self-Attention?

- Yuxin Wang, Chu-Tak Lee, Qipeng Guo, Zhangyue Yin, Yunhua Zhou, Xuanjing Huang, Xipeng Qiu
- abstract: Transformers have made progress in miscellaneous tasks, but suffer from quadratic computational and memory complexities. Recent works propose sparse transformers with attention on sparse graphs to reduce complexity and remain strong performance. While effective, the crucial parts of how dense a graph needs to be to perform well are not fully explored. In this paper, we propose Normalized Information Payload (NIP), a graph scoring function measuring information transfer on graph, which provides an analysis tool for trade-offs between performance and complexity. Guided by this theoretical analysis, we present Hypercube Transformer, a sparse transformer that models token interactions in a hypercube and shows comparable or even better results with vanilla transformer while yielding $O(N\log N)$ complexity with sequence length N. Experiments on tasks requiring various sequence lengths lay validation for our graph function well.

Improved Certified Defenses against Data Poisoning with (Deterministic) Finite Aggregation

- Wenxiao Wang, Alexander J Levine, Soheil Feizi
- abstract: Data poisoning attacks aim at manipulating model behaviors through distorting training data. Previously, an aggregation-based certified defense, Deep Partition Aggregation (DPA), was proposed to mitigate this threat. DPA predicts through an aggregation of base classifiers trained on disjoint subsets of data, thus restricting its sensitivity to dataset distortions. In this work, we propose an improved certified defense against general poisoning attacks, namely Finite Aggregation. In contrast to DPA, which directly splits the training set into disjoint subsets, our method first splits the training set into smaller disjoint subsets and then combines duplicates of them to build larger (but not disjoint) subsets for training base classifiers. This reduces the worst-case impacts of poison samples and thus improves certified robustness bounds. In addition, we offer an alternative view of our method, bridging the designs of deterministic and stochastic aggregation-based certified defenses. Empirically, our proposed Finite Aggregation consistently improves certificates on MNIST, CIFAR-10, and GTSRB, boosting certified fractions by up to 3.05%, 3.87% and 4.77%, respectively, while keeping the same clean accuracies as DPA's, effectively establishing a new state of the art in (pointwise) certified robustness against data poisoning.

Understanding Gradual Domain Adaptation: Improved Analysis, Optimal Path and Beyond

- Haoxiang Wang, Bo Li, Han Zhao
- abstract: The vast majority of existing algorithms for unsupervised domain adaptation (UDA) focus on adapting from a labeled source domain to an unlabeled target domain directly in a one-off way. Gradual domain adaptation (GDA), on the other hand, assumes a path of $(T-1)$ unlabeled intermediate domains bridging the source and target, and aims to provide better generalization in the target domain by leveraging the intermediate ones. Under certain assumptions, Kumar et al. (2020) proposed a simple algorithm, Gradual Self-Training, along with a generalization bound in the order of $e^{O(T)} \left(\varepsilon_0+O\left(\sqrt{log(T)/n}\right)\right)$ for the target domain error, where ε_0 is the source domain error and n is the data size of each domain. Due to the exponential factor, this upper bound becomes vacuous when T is only moderately large. In this work, we analyze gradual self-training under more general and relaxed assumptions, and prove a significantly improved generalization bound as $\widetilde{O}\left(\varepsilon_0 + T\Delta + T/\sqrt{n} + 1/\sqrt{nT}\right)$, where Δ is the average distributional distance between consecutive domains. Compared with the existing bound with an exponential dependency on T as a multiplicative factor, our bound only depends on T linearly and additively. Perhaps more interestingly, our result implies the existence of an optimal choice of T that minimizes the generalization error, and it also naturally suggests an optimal way to construct the path of intermediate domains so as to minimize the accumulative path length $T\Delta$ between the source and target. To corroborate the implications of our theory, we examine gradual self-training on multiple semi-synthetic and real datasets, which confirms our findings. We believe our insights provide a path forward toward the design of future GDA algorithms.

Communication-Efficient Adaptive Federated Learning

- Yujia Wang, Lu Lin, Jinghui Chen
- abstract: Federated learning is a machine learning training paradigm that enables clients to jointly train models without sharing their own localized data. However, the implementation of federated learning in practice still faces numerous challenges, such as the large communication overhead due to the repetitive server-client synchronization and the lack of adaptivity by SGD-based model updates. Despite that various methods have been proposed for reducing the communication cost by gradient compression or quantization, and the federated versions of adaptive optimizers such as FedAdam are proposed to add more adaptivity, the current federated learning framework still cannot solve the aforementioned challenges all at once. In this paper, we propose a novel communication-efficient adaptive federated learning method (FedCAMS) with theoretical convergence guarantees. We show that in the nonconvex stochastic optimization setting, our proposed FedCAMS achieves the same convergence rate of $O(\frac{1}{\sqrt{TKm}})$ as its non-compressed counterparts. Extensive experiments on various benchmarks verify our theoretical analysis.

Provable Acceleration of Heavy Ball beyond Quadratics for a Class of Polyak-Lojasiewicz Functions when the Non-Convexity is Averaged-Out

- Jun-Kun Wang, Chi-Heng Lin, Andre Wibisono, Bin Hu
- abstract: Heavy Ball (HB) nowadays is one of the most popular momentum methods in non-convex optimization. It has been widely observed that incorporating the Heavy Ball dynamic in gradient-based methods accelerates the training process of modern machine learning models. However, the progress on establishing its theoretical foundation of acceleration is apparently far behind its empirical success. Existing provable acceleration results are of the quadratic or close-to-quadratic functions, as the current techniques of showing HB's acceleration are limited to the case when the Hessian is fixed. In this work, we develop some new techniques that help show acceleration beyond quadratics, which is achieved by analyzing how the change of the Hessian at two consecutive time points affects the convergence speed. Based on our technical results, a class of Polyak-Lojasiewicz (PL) optimization problems for which provable acceleration can be achieved via HB is identified. Moreover, our analysis demonstrates a benefit of adaptively setting the momentum parameter.

Robustness Verification for Contrastive Learning

- Zekai Wang, Weiwei Liu
- abstract: Contrastive adversarial training has successfully improved the robustness of contrastive learning (CL). However, the robustness metric used in these methods is linked to attack algorithms, image labels and downstream tasks, all of which may affect the consistency and reliability of robustness metric for CL. To address these problems, this paper proposes a novel Robustness Verification framework for Contrastive Learning (RVCL). Furthermore, we use extreme value theory to reveal the relationship between the robust radius of the CL encoder and that of the supervised downstream task. Extensive experimental results on various benchmark models and datasets verify our theoretical findings, and further demonstrate that our proposed RVCL is able to evaluate the robustness of both models and images. Our code is available at https://github.com/wzekai99/RVCL.

Convergence and Recovery Guarantees of the K-Subspaces Method for Subspace Clustering

- Peng Wang, Huikang Liu, Anthony Man-Cho So, Laura Balzano
- abstract: The K-subspaces (KSS) method is a generalization of the K-means method for subspace clustering. In this work, we present local convergence analysis and a recovery guarantee for KSS, assuming data are generated by the semi-random union of subspaces model, where N points are randomly sampled from $K \ge 2$ overlapping subspaces. We show that if the initial assignment of the KSS method lies within a neighborhood of a true clustering, it converges at a superlinear rate and finds the correct clustering within $\Theta(\log\log N)$ iterations with high probability. Moreover, we propose a thresholding inner-product based spectral method for initialization and prove that it produces a point in this neighborhood. We also present numerical results of the studied method to support our theoretical developments.

NP-Match: When Neural Processes meet Semi-Supervised Learning

- Jianfeng Wang, Thomas Lukasiewicz, Daniela Massiceti, Xiaolin Hu, Vladimir Pavlovic, Alexandros Neophytou
- abstract: Semi-supervised learning (SSL) has been widely explored in recent years, and it is an effective way of leveraging unlabeled data to reduce the reliance on labeled data. In this work, we adjust neural processes (NPs) to the semi-supervised image classification task, resulting in a new method named NP-Match. NP-Match is suited to this task for two reasons. Firstly, NP-Match implicitly compares data points when making predictions, and as a result, the prediction of each unlabeled data point is affected by the labeled data points that are similar to it, which improves the quality of pseudolabels. Secondly, NP-Match is able to estimate uncertainty that can be used as a tool for selecting unlabeled samples with reliable pseudo-labels. Compared with uncertainty-based SSL methods implemented with Monte Carlo (MC) dropout, NP-Match estimates uncertainty with much less computational overhead, which can save time at both the training and the testing phases. We conducted extensive experiments on four public datasets, and NP-Match outperforms state-of-theart (SOTA) results or achieves competitive results on them, which shows the effectiveness of NPMatch and its potential for SSL.

Iterative Double Sketching for Faster Least-Squares Optimization

- Rui Wang, Yanyan Ouyang, Wangli Xu
- abstract: This work is concerned with the overdetermined linear least-squares problem for large scale data. We generalize the iterative Hessian sketching (IHS) algorithm and propose a new sketching framework named iterative double sketching (IDS) which uses approximations for both the gradient and the Hessian in each iteration. To understand the behavior of the IDS algorithm and choose the optimal hyperparameters, we derive the exact limit of the conditional prediction error of the IDS algorithm in the setting of Gaussian sketching. Guided by this theoretical result, we propose an efficient IDS

algorithm via a new class of sequentially related sketching matrices. We give a non-asymptotic analysis of this efficient IDS algorithm which shows that the proposed algorithm achieves the state-of-the-art trade-off between accuracy and efficiency.

What Language Model Architecture and Pretraining Objective Works Best for Zero-Shot Generalization?

- Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, Colin Raffel
- abstract: Large pretrained Transformer language models have been shown to exhibit zero-shot generalization, i.e. they can perform a wide variety of tasks that they were not explicitly trained on. However, the architectures and pretraining objectives used across state-of-the-art models differ significantly, and there has been limited systematic comparison of these factors. In this work, we present a large-scale evaluation of modeling choices and their impact on zero-shot generalization. In particular, we focus on text-to-text models and experiment with three model architectures (causal/non-causal decoder-only and encoder-decoder), trained with two different pretraining objectives (autoregressive and masked language modeling), and evaluated with and without multitask prompted finetuning. We train models with over 5 billion parameters for more than 168 billion tokens, thereby increasing the likelihood that our conclusions will transfer to even larger scales. Our experiments show that causal decoder-only models trained on an autoregressive language modeling objective exhibit the strongest zero-shot generalization after purely self-supervised pretraining. However, models with non-causal visibility on their input trained with a masked language modeling objective followed by multitask finetuning perform the best among our experiments. We therefore consider the adaptation of pretrained models across architectures and objectives. Code and checkpoints are available at https://github.com/bigscience-workshop/architecture-objective.

Improving Task-free Continual Learning by Distributionally Robust Memory Evolution

- Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, Mingchen Gao
- abstract: Task-free continual learning (CL) aims to learn a non-stationary data stream without explicit task definitions and not forget previous knowledge. The widely adopted memory replay approach could gradually become less effective for long data streams, as the model may memorize the stored examples and overfit the memory buffer. Second, existing methods overlook the high uncertainty in the memory data distribution since there is a big gap between the memory data distribution and the distribution of all the previous data examples. To address these problems, for the first time, we propose a principled memory evolution framework to dynamically evolve the memory data distribution by making the memory buffer gradually harder to be memorized with distributionally robust optimization (DRO). We then derive a family of methods to evolve the memory buffer data in the continuous probability measure space with Wasserstein gradient flow (WGF). The proposed DRO is w.r.t the worst-case evolved memory data distribution, thus guarantees the model performance and learns significantly more robust features than existing memory-replay-based methods. Extensive experiments on existing benchmarks demonstrate the effectiveness of the proposed methods for alleviating forgetting. As a by-product of the proposed framework, our method is more robust to adversarial examples than existing task-free CL methods.

Risk-Averse No-Regret Learning in Online Convex Games

- Zifan Wang, Yi Shen, Michael Zavlanos
- abstract: We consider an online stochastic game with risk-averse agents whose goal is to learn optimal decisions that minimize the risk of incurring significantly high costs. Specifically, we use the Conditional Value at Risk (CVaR) as a risk measure that the agents can estimate using bandit feedback in the form of the cost values of only their selected actions. Since the distributions of the cost functions depend on the actions of all agents that are generally unobservable, they are themselves unknown and, therefore, the CVaR values of the costs are difficult to compute. To address this challenge, we propose a new online risk-averse learning algorithm that relies on one-point zeroth-order estimation of the CVaR gradients computed using CVaR values that are estimated by appropriately sampling the cost functions. We show that this algorithm achieves sub-linear regret with high probability. We also propose two variants of this algorithm that improve performance. The first variant relies on a new sampling strategy that uses samples from the previous iteration to improve the estimation accuracy of the CVaR values. The second variant employs residual feedback that uses CVaR values from the previous iteration to reduce the variance of the CVaR gradient estimates. We theoretically analyze the convergence properties of these variants and illustrate their performance on an online market problem that we model as a Cournot game.

Provable Domain Generalization via Invariant-Feature Subspace Recovery

- Haoxiang Wang, Haozhe Si, Bo Li, Han Zhao
- abstract: Domain generalization asks for models trained over a set of training environments to perform well in unseen test environments. Recently, a series of algorithms such as Invariant Risk Minimization (IRM) has been proposed for domain generalization. However, Rosenfeld et al. (2021) shows that in a simple linear data model, even if non-convexity issues are ignored, IRM and its extensions cannot generalize to unseen environments with less than d_s+1 training environments, where d_s is the dimension of the spurious-feature subspace. In this paper, we propose to achieve domain generalization with Invariant-feature Subspace Recovery (ISR). Our first algorithm, ISR-Mean, can identify the subspace spanned by invariant features from the first-order moments of the class-conditional distributions, and achieve provable domain generalization with d_s+1 training environments under the data model of Rosenfeld et al. (2021). Our second algorithm, ISR-Cov, further reduces the required number of training environments to $O(1)$ using the information of second-order moments. Notably, unlike IRM, our algorithms bypass non-convexity issues and enjoy global convergence guarantees. Empirically, our ISRs can obtain superior performance compared with IRM on synthetic benchmarks. In addition, on three real-world image and text datasets, we show that both ISRs can be used as simple yet effective post-processing methods to improve the worst-case accuracy of (pre-)trained models against spurious correlations and group shifts.

ProgFed: Effective, Communication, and Computation Efficient Federated Learning by Progressive Training

- Hui-Po Wang, Sebastian Stich, Yang He, Mario Fritz
- abstract: Federated learning is a powerful distributed learning scheme that allows numerous edge devices to collaboratively train a model without sharing their data. However, training is resource-intensive for edge devices, and limited network bandwidth is often the main bottleneck. Prior work often overcomes the constraints by condensing the models or messages into compact formats, e.g., by gradient compression or distillation. In contrast, we propose ProgFed, the first progressive training framework for efficient and effective federated learning. It inherently reduces computation and two-way communication costs while maintaining the strong performance of the final models. We theoretically prove that ProgFed converges at the same asymptotic rate as standard training on full models. Extensive results on a broad range of architectures, including CNNs (VGG, ResNet, ConvNets) and U-nets, and diverse tasks from simple classification to medical image segmentation show that our highly effective training approach saves up to $20%$ computation and up to $63%$ communication costs for converged models. As our approach is also complimentary to prior work on compression, we can achieve a wide range of trade-offs by combining these techniques, showing reduced communication of up to $50\times$ at only $0.1%$ loss in utility. Code is available at https://github.com/a514514772/ProgFed.

Model-based Meta Reinforcement Learning using Graph Structured Surrogate Models and Amortized Policy Search

- Qi Wang, Herke Van Hoof
- abstract: Reinforcement learning is a promising paradigm for solving sequential decision-making problems, but low data efficiency and weak generalization across tasks are bottlenecks in real-world applications. Model-based meta reinforcement learning addresses these issues by learning dynamics and leveraging knowledge from prior experience. In this paper, we take a closer look at this framework and propose a new posterior sampling based approach that consists of a new model to identify task dynamics together with an amortized policy optimization step. We show that our model,

called a graph structured surrogate model (GSSM), achieves competitive dynamics prediction performance with lower model complexity. Moreover, our approach in policy search is able to obtain high returns and allows fast execution by avoiding test-time policy gradient updates.

Approximately Equivariant Networks for Imperfectly Symmetric Dynamics

- Rui Wang, Robin Walters, Rose Yu
- abstract: Incorporating symmetry as an inductive bias into neural network architecture has led to improvements in generalization, data efficiency, and physical consistency in dynamics modeling. Methods such as CNNs or equivariant neural networks use weight tying to enforce symmetries such as shift invariance or rotational equivariance. However, despite the fact that physical laws obey many symmetries, real-world dynamical data rarely conforms to strict mathematical symmetry either due to noisy or incomplete data or to symmetry breaking features in the underlying dynamical system. We explore approximately equivariant networks which are biased towards preserving symmetry but are not strictly constrained to do so. By relaxing equivariance constraints, we find that our models can outperform both baselines with no symmetry bias and baselines with overly strict symmetry in both simulated turbulence domains and real-world multi-stream jet flow.

Three-stage Evolution and Fast Equilibrium for SGD with Non-degerate Critical Points

- Yi Wang, Zhiren Wang
- abstract: We justify the fast equilibrium conjecture on stochastic gradient descent from (Li et al. 2020) under the assumptions that critical points are non-degenerate and the stochastic noise is a standard Gaussian. In this case, we prove an SGD with constant effective learning rate consists of three stages: descent, diffusion and tunneling, and explicitly identify temporary equilibrium states in the normalized parameter space that can be observed within practical training time. This interprets the gap between the mixing time in the fast equilibrium conjecture and the previously known upper bound. While our assumptions do not represent typical implementations of SGD of neural networks in practice, this is the first description of the three-stage mechanism in any case. The main finding in this mechanism is that a temporary equilibrium of local nature is quickly achieved after polynomial time (in term of the reciprocal of the intrinsic learning rate) and then stabilizes within observable time scales; and that the temporary equilibrium is in general different from the global Gibbs equilibrium, which will only appear after an exponentially long period beyond typical training limits. Our experiments support that this mechanism may extend to the general case.

Understanding Instance-Level Impact of Fairness Constraints

- Jialu Wang, Xin Eric Wang, Yang Liu
- abstract: A variety of fairness constraints have been proposed in the literature to mitigate group-level statistical bias. Their impacts have been largely evaluated for different groups of populations corresponding to a set of sensitive attributes, such as race or gender. Nonetheless, the community has not observed sufficient explorations for how imposing fairness constraints fare at an instance level. Building on the concept of influence function, a measure that characterizes the impact of a training example on the target model and its predictive performance, this work studies the influence of training examples when fairness constraints are imposed. We find out that under certain assumptions, the influence function with respect to fairness constraints can be decomposed into a kernelized combination of training examples. One promising application of the proposed fairness influence function is to identify suspicious training examples that may cause model discrimination by ranking their influence scores. We demonstrate with extensive experiments that training on a subset of weighty data examples leads to lower fairness violations with a trade-off of accuracy.

Tractable Uncertainty for Structure Learning

- Benjie Wang, Matthew R Wicker, Marta Kwiatkowska
- abstract: Bayesian structure learning allows one to capture uncertainty over the causal directed acyclic graph (DAG) responsible for generating given data. In this work, we present Tractable Uncertainty for STructure learning (TRUST), a framework for approximate posterior inference that relies on probabilistic circuits as a representation of our posterior belief. In contrast to sample-based posterior approximations, our representation can capture a much richer space of DAGs, while being able to tractably answer a range of useful inference queries. We empirically demonstrate how probabilistic circuits can be used to as an augmented representation for structure learning methods, leading to improvement in both the quality of inferred structures and posterior uncertainty. Experimental results also demonstrate the improved representational capacity of TRUST, outperforming competing methods on conditional query answering.

Causal Dynamics Learning for Task-Independent State Abstraction

- Zizhao Wang, Xuesu Xiao, Zifan Xu, Yuke Zhu, Peter Stone
- abstract: Learning dynamics models accurately is an important goal for Model-Based Reinforcement Learning (MBRL), but most MBRL methods learn a dense dynamics model which is vulnerable to spurious correlations and therefore generalizes poorly to unseen states. In this paper, we introduce Causal Dynamics Learning for Task-Independent State Abstraction (CDL), which first learns a theoretically proved causal dynamics model that removes unnecessary dependencies between state variables and the action, thus generalizing well to unseen states. A state abstraction can then be derived from the learned dynamics, which not only improves sample efficiency but also applies to a wider range of tasks than existing state abstraction methods. Evaluated on two simulated environments and downstream tasks, both the dynamics model and policies learned by the proposed method generalize well to unseen states and the derived state abstraction improves sample efficiency compared to learning without it.

Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms

- Xuchuang Wang, Hong Xie, John C. S. Lui
- abstract: We generalize the multiple-play multi-armed bandits (MP-MAB) problem with a shareable arms setting, in which several plays can share the same arm. Furthermore, each shareable arm has a finite reward capacity and a “per-load” reward distribution, both of which are unknown to the learner. The reward from a shareable arm is load-dependent, which is the “per-load” reward multiplying either the number of plays pulling the arm, or its reward capacity when the number of plays exceeds the capacity limit. When the “per-load” reward follows a Gaussian distribution, we prove a sample complexity lower bound of learning the capacity from load-dependent rewards and also a regret lower bound of this new MP-MAB problem. We devise a capacity estimator whose sample complexity upper bound matches the lower bound in terms of reward means and capacities. We also propose an online learning algorithm to address the problem and prove its regret upper bound. This regret upper bound’s first term is the same as regret lower bound’s, and its second and third terms also evidently correspond to lower bound’s. Extensive experiments validate our algorithm’s performance and also its gain in 5G & 4G base station selection.

Generative Coarse-Graining of Molecular Conformations

- Wujie Wang, Minkai Xu, Chen Cai, Benjamin K Miller, Tess Smidt, Yusu Wang, Jian Tang, Rafael Gomez-Bombarelli
- abstract: Coarse-graining (CG) of molecular simulations simplifies the particle representation by grouping selected atoms into pseudo-beads and therefore drastically accelerates simulation. However, such CG procedure induces information losses, which makes accurate backmapping, i.e., restoring fine-grained (FG) coordinates from CG coordinates, a long-standing challenge. Inspired by the recent progress in generative models and equivariant networks, we propose a novel model that rigorously embeds the vital probabilistic nature and geometrical consistency requirements of the backmapping

transformation. Our model encodes the FG uncertainties into an invariant latent space and decodes them back to FG geometries via equivariant convolutions. To standardize the evaluation of this domain, we further provide three comprehensive benchmarks based on molecular dynamics trajectories. Extensive experiments show that our approach always recovers more realistic structures and outperforms existing data-driven methods with a significant margin.

Nonparametric Embeddings of Sparse High-Order Interaction Events

- Zheng Wang, Yiming Xu, Conor Tillinghast, Shibo Li, Akil Narayan, Shandian Zhe
- abstract: High-order interaction events are common in real-world applications. Learning embeddings that encode the complex relationships of the participants from these events is of great importance in knowledge mining and predictive tasks. Despite the success of existing approaches, e.g. Poisson tensor factorization, they ignore the sparse structure underlying the data, namely the occurred interactions are far less than the possible interactions among all the participants. In this paper, we propose Nonparametric Embeddings of Sparse High-order interaction events (NESH). We hybridize a sparse hypergraph (tensor) process and a matrix Gaussian process to capture both the asymptotic structural sparsity within the interactions and nonlinear temporal relationships between the participants. We prove strong asymptotic bounds (including both a lower and an upper bound) of the sparse ratio, which reveals the asymptotic properties of the sampled structure. We use batch-normalization, stick-breaking construction and sparse variational GP approximations to develop an efficient, scalable model inference algorithm. We demonstrate the advantage of our approach in several real-world applications.

When Are Linear Stochastic Bandits Attackable?

- Huazheng Wang, Haifeng Xu, Hongning Wang
- abstract: We study adversarial attacks on linear stochastic bandits: by manipulating the rewards, an adversary aims to control the behaviour of the bandit algorithm. Perhaps surprisingly, we first show that some attack goals can never be achieved. This is in a sharp contrast to context-free stochastic bandits, and is intrinsically due to the correlation among arms in linear stochastic bandits. Motivated by this finding, this paper studies the attackability of a k-armed linear bandit environment. We first provide a complete necessity and sufficiency characterization of attackability based on the geometry of the arms' context vectors. We then propose a two-stage attack method against LinUCB and Robust Phase Elimination. The method first asserts whether the given environment is attackable; and if yes, it poisons the rewards to force the algorithm to pull a target arm linear times using only a sublinear cost. Numerical experiments further validate the effectiveness and cost-efficiency of the proposed attack method.

DRAGONN: Distributed Randomized Approximate Gradients of Neural Networks

- Zhuang Wang, Zhaozhuo Xu, Xinyu Wu, Anshumali Shrivastava, T. S. Eugene Ng
- abstract: Data-parallel distributed training (DDT) has become the de-facto standard for accelerating the training of most deep learning tasks on massively parallel hardware. In the DDT paradigm, the communication overhead of gradient synchronization is the major efficiency bottleneck. A widely adopted approach to tackle this issue is gradient sparsification (GS). However, the current GS methods introduce significant new overhead in compressing the gradients, outweighing the communication overhead and becoming the new efficiency bottleneck. In this paper, we propose DRAGONN, a randomized hashing algorithm for GS in DDT. DRAGONN can significantly reduce the compression time by up to 70% compared to state-of-the-art GS approaches, and achieve up to 3.52x speedup in total training throughput.

Finite-Sum Coupled Compositional Stochastic Optimization: Theory and Applications

- Bokun Wang, Tianbao Yang
- abstract: This paper studies stochastic optimization for a sum of compositional functions, where the inner-level function of each summand is coupled with the corresponding summation index. We refer to this family of problems as finite-sum coupled compositional optimization (FCCO). It has broad applications in machine learning for optimizing non-convex or convex compositional measures/objectives such as average precision (AP), p-norm push, listwise ranking losses, neighborhood component analysis (NCA), deep survival analysis, deep latent variable models, etc., which deserves finer analysis. Yet, existing algorithms and analyses are restricted in one or other aspects. The contribution of this paper is to provide a comprehensive convergence analysis of a simple stochastic algorithm for both non-convex and convex objectives. Our key result is the improved oracle complexity with the parallel speed-up by using the moving-average based estimator with mini-batching. Our theoretical analysis also exhibits new insights for improving the practical implementation by sampling the batches of equal size for the outer and inner levels. Numerical experiments on AP maximization, NCA, and p-norm push corroborate some aspects of the theory.

OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework

- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, Hongxia Yang
- abstract: In this work, we pursue a unified paradigm for multimodal pretraining to break the shackles of complex task/modality-specific customization. We propose OFA, a Task-Agnostic and Modality-Agnostic framework that supports Task Comprehensiveness. OFA unifies a diverse set of cross-modal and unimodal tasks, including image generation, visual grounding, image captioning, image classification, language modeling, etc., in a simple sequence-to-sequence learning framework. OFA follows the instruction-based learning in both pretraining and finetuning stages, requiring no extra task-specific layers for downstream tasks. In comparison with the recent state-of-the-art vision & language models that rely on extremely large cross-modal datasets, OFA is pretrained on only 20M publicly available image-text pairs. Despite its simplicity and relatively small-scale training data, OFA achieves new SOTAs in a series of cross-modal tasks while attaining highly competitive performances on uni-modal tasks. Our further analysis indicates that OFA can also effectively transfer to unseen tasks and unseen domains. Our code and models are publicly available at https://github.com/OFA-Sys/OFA.

How Powerful are Spectral Graph Neural Networks

- Xiyuan Wang, Muhan Zhang
- abstract: Spectral Graph Neural Network is a kind of Graph Neural Network (GNN) based on graph signal filters. Some models able to learn arbitrary spectral filters have emerged recently. However, few works analyze the expressive power of spectral GNNs. This paper studies spectral GNNs' expressive power theoretically. We first prove that even spectral GNNs without nonlinearity can produce arbitrary graph signals and give two conditions for reaching universality. They are: 1) no multiple eigenvalues of graph Laplacian, and 2) no missing frequency components in node features. We also establish a connection between the expressive power of spectral GNNs and Graph Isomorphism (GI) testing, the latter of which is often used to characterize spatial GNNs' expressive power. Moreover, we study the difference in empirical performance among different spectral GNNs with the same expressive power from an optimization perspective, and motivate the use of an orthogonal basis whose weight function corresponds to the graph signal density in the spectrum. Inspired by the analysis, we propose JacobiConv, which uses Jacobi basis due to its orthogonality and flexibility to adapt to a wide range of weight functions. JacobiConv deserts nonlinearity while outperforming all baselines on both synthetic and real-world datasets.

Thompson Sampling for Robust Transfer in Multi-Task Bandits

- Zhi Wang, Chicheng Zhang, Kamalika Chaudhuri

- abstract: We study the problem of online multi-task learning where the tasks are performed within similar but not necessarily identical multi-armed bandit environments. In particular, we study how a learner can improve its overall performance across multiple related tasks through robust transfer of knowledge. While an upper confidence bound (UCB)-based algorithm has recently been shown to achieve nearly-optimal performance guarantees in a setting where all tasks are solved concurrently, it remains unclear whether Thompson sampling (TS) algorithms, which have superior empirical performance in general, share similar theoretical properties. In this work, we present a TS-type algorithm for a more general online multi-task learning protocol, which extends the concurrent setting. We provide its frequentist analysis and prove that it is also nearly-optimal using a novel concentration inequality for multi-task data aggregation at random stopping times. Finally, we evaluate the algorithm on synthetic data and show that the TS-type algorithm enjoys superior empirical performance in comparison with the UCB-based algorithm and a baseline algorithm that performs TS for each individual task without transfer.

Individual Reward Assisted Multi-Agent Reinforcement Learning

- Li Wang, Yupeng Zhang, Yujing Hu, Weixun Wang, Chongjie Zhang, Yang Gao, Jianye Hao, Tangjie Lv, Changjie Fan
- abstract: In many real-world multi-agent systems, the sparsity of team rewards often makes it difficult for an algorithm to successfully learn a cooperative team policy. At present, the common way for solving this problem is to design some dense individual rewards for the agents to guide the cooperation. However, most existing works utilize individual rewards in ways that do not always promote teamwork and sometimes are even counterproductive. In this paper, we propose Individual Reward Assisted Team Policy Learning (IRAT), which learns two policies for each agent from the dense individual reward and the sparse team reward with discrepancy constraints for updating the two policies mutually. Experimental results in different scenarios, such as the Multi-Agent Particle Environment and the Google Research Football Environment, show that IRAT significantly outperforms the baseline methods and can greatly promote team policy learning without deviating from the original team objective, even when the individual rewards are misleading or conflict with the team rewards.

Removing Batch Normalization Boosts Adversarial Training

- Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, Zhangyang Wang
- abstract: Adversarial training (AT) defends deep neural networks against adversarial attacks. One challenge that limits its practical application is the performance degradation on clean samples. A major bottleneck identified by previous works is the widely used batch normalization (BN), which struggles to model the different statistics of clean and adversarial training samples in AT. Although the dominant approach is to extend BN to capture this mixture of distribution, we propose to completely eliminate this bottleneck by removing all BN layers in AT. Our normalizer-free robust training (NoFrost) method extends recent advances in normalizer-free networks to AT for its unexplored advantage on handling the mixture distribution challenge. We show that NoFrost achieves adversarial robustness with only a minor sacrifice on clean sample accuracy. On ImageNet with ResNet50, NoFrost achieves $74.06%$ clean accuracy, which drops merely $2.00%$ from standard training. In contrast, BN-based AT obtains $59.28%$ clean accuracy, suffering a significant $16.78%$ drop from standard training. In addition, NoFrost achieves a $23.56%$ adversarial robustness against PGD attack, which improves the $13.57%$ robustness in BN-based AT. We observe better model smoothness and larger decision margins from NoFrost, which make the models less sensitive to input perturbations and thus more robust. Moreover, when incorporating more data augmentations into NoFrost, it achieves comprehensive robustness against multiple distribution shifts. Code and pre-trained models are public at https://github.com/amazon-research/normalizer-free-robust-training.

Partial and Asymmetric Contrastive Learning for Out-of-Distribution Detection in Long-Tailed Recognition

- Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, Zhangyang Wang
- abstract: Existing out-of-distribution (OOD) detection methods are typically benchmarked on training sets with balanced class distributions. However, in real-world applications, it is common for the training sets to have long-tailed distributions. In this work, we first demonstrate that existing OOD detection methods commonly suffer from significant performance degradation when the training set is long-tail distributed. Through analysis, we posit that this is because the models struggle to distinguish the minority tail-class in-distribution samples, from the true OOD samples, making the tail classes more prone to be falsely detected as OOD. To solve this problem, we propose Partial and Asymmetric Supervised Contrastive Learning (PASCL), which explicitly encourages the model to distinguish between tail-class in-distribution samples and OOD samples. To further boost in-distribution classification accuracy, we propose Auxiliary Branch Finetuning, which uses two separate branches of BN and classification layers for anomaly detection and in-distribution classification, respectively. The intuition is that in-distribution and OOD anomaly data have different underlying distributions. Our method outperforms previous state-of-the-art method by $1.29%$, $1.45%$, $0.69%$ anomaly detection false positive rate (FPR) and $3.24%$, $4.06%$, $7.89%$ in-distribution classification accuracy on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT, respectively. Code and pre-trained models are available at https://github.com/amazon-research/long-tailed-ood-detection.

Nonparametric Factor Trajectory Learning for Dynamic Tensor Decomposition

- Zheng Wang, Shandian Zhe
- abstract: Tensor decomposition is a fundamental framework to analyze data that can be represented by multi-dimensional arrays. In practice, tensor data are often accompanied with temporal information, namely the time points when the entry values were generated. This information implies abundant, complex temporal variation patterns. However, current methods always assume the factor representations of the entities in each tensor mode are static, and never consider their temporal evolution. To fill this gap, we propose NONparametric FActor Trajectory learning for dynamic tensor decomposition (NONFAT). We place Gaussian process (GP) priors in the frequency domain and conduct inverse Fourier transform via Gauss-Laguerre quadrature to sample the trajectory functions. In this way, we can overcome data sparsity and obtain robust trajectory estimates across long time horizons. Given the trajectory values at specific time points, we use a second-level GP to sample the entry values and to capture the temporal relationship between the entities. For efficient and scalable inference, we leverage the matrix Gaussian structure in the model, introduce a matrix Gaussian posterior, and develop a nested sparse variational learning algorithm. We have shown the advantage of our method in several real-world applications.

Thompson Sampling for (Combinatorial) Pure Exploration

- Siwei Wang, Jun Zhu
- abstract: Existing methods of combinatorial pure exploration mainly focus on the UCB approach. To make the algorithm efficient, they usually use the sum of upper confidence bounds within arm set S to represent the upper confidence bound of S, which can be much larger than the tight upper confidence bound of S and leads to a much higher complexity than necessary, since the empirical means of different arms in S are independent. To deal with this challenge, we explore the idea of Thompson Sampling (TS) that uses independent random samples instead of the upper confidence bounds, and design the first TS-based algorithm TS-Explore for (combinatorial) pure exploration. In TS-Explore, the sum of independent random samples within arm set S will not exceed the tight upper confidence bound of S with high probability. Hence it solves the above challenge, and achieves a lower complexity upper bound than existing efficient UCB-based algorithms in general combinatorial pure exploration. As for pure exploration of classic multi-armed bandit, we show that TS-Explore achieves an asymptotically optimal complexity upper bound.

Policy Gradient Method For Robust Reinforcement Learning

- Yue Wang, Shaofeng Zou

- abstract: This paper develops the first policy gradient method with global optimality guarantee and complexity analysis for robust reinforcement learning under model mismatch. Robust reinforcement learning is to learn a policy robust to model mismatch between simulator and real environment. We first develop the robust policy (sub-)gradient, which is applicable for any differentiable parametric policy class. We show that the proposed robust policy gradient method converges to the global optimum asymptotically under direct policy parameterization. We further develop a smoothed robust policy gradient method, and show that to achieve an ϵ-global optimum, the complexity is $\mathcal O(\epsilon^{-3})$. We then extend our methodology to the general model-free setting, and design the robust actor-critic method with differentiable parametric policy class and value function. We further characterize its asymptotic convergence and sample complexity under the tabular setting. Finally, we provide simulation results to demonstrate the robustness of our methods.

Certifying Out-of-Domain Generalization for Blackbox Functions

- Maurice G Weber, Linyi Li, Boxin Wang, Zhikuan Zhao, Bo Li, Ce Zhang
- abstract: Certifying the robustness of model performance under bounded data distribution drifts has recently attracted intensive interest under the umbrella of distributional robustness. However, existing techniques either make strong assumptions on the model class and loss functions that can be certified, such as smoothness expressed via Lipschitz continuity of gradients, or require to solve complex optimization problems. As a result, the wider application of these techniques is currently limited by its scalability and flexibility — these techniques often do not scale to large-scale datasets with modern deep neural networks or cannot handle loss functions which may be non-smooth such as the 0-1 loss. In this paper, we focus on the problem of certifying distributional robustness for blackbox models and bounded loss functions, and propose a novel certification framework based on the Hellinger distance. Our certification technique scales to ImageNet-scale datasets, complex models, and a diverse set of loss functions. We then focus on one specific application enabled by such scalability and flexibility, i.e., certifying out-of-domain generalization for large neural networks and loss functions such as accuracy and AUC. We experimentally validate our certification method on a number of datasets, ranging from ImageNet, where we provide the first non-vacuous certified out-of-domain generalization, to smaller classification tasks where we are able to compare with the state-of-the-art and show that our method performs considerably better.

More Than a Toy: Random Matrix Models Predict How Real-World Neural Representations Generalize

- Alexander Wei, Wei Hu, Jacob Steinhardt
- abstract: Of theories for why large-scale machine learning models generalize despite being vastly overparameterized, which of their assumptions are needed to capture the qualitative phenomena of generalization in the real world? On one hand, we find that most theoretical analyses fall short of capturing these qualitative phenomena even for kernel regression, when applied to kernels derived from large-scale neural networks (e.g., ResNet-50) and real data (e.g., CIFAR-100). On the other hand, we find that the classical GCV estimator (Craven and Wahba, 1978) accurately predicts generalization risk even in such overparameterized settings. To bolster this empirical finding, we prove that the GCV estimator converges to the generalization risk whenever a local random matrix law holds. Finally, we apply this random matrix theory lens to explain why pretrained representations generalize better as well as what factors govern scaling laws for kernel regression. Our findings suggest that random matrix theory, rather than just being a toy model, may be central to understanding the properties of neural representations in practice.

To Smooth or Not? When Label Smoothing Meets Noisy Labels

- Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, Yang Liu
- abstract: Label smoothing (LS) is an arising learning paradigm that uses the positively weighted average of both the hard training labels and uniformly distributed soft labels. It was shown that LS serves as a regularizer for training data with hard labels and therefore improves the generalization of the model. Later it was reported LS even helps with improving robustness when learning with noisy labels. However, we observed that the advantage of LS vanishes when we operate in a high label noise regime. Intuitively speaking, this is due to the increased entropy of P(noisy label|X) when the noise rate is high, in which case, further applying LS tends to “over-smooth” the estimated posterior. We proceeded to discover that several learning-with-noisy-labels solutions in the literature instead relate more closely to negative/not label smoothing (NLS), which acts counter to LS and defines as using a negative weight to combine the hard and soft labels! We provide understandings for the properties of LS and NLS when learning with noisy labels. Among other established properties, we theoretically show NLS is considered more beneficial when the label noise rates are high. We provide extensive experimental results on multiple benchmarks to support our findings too. Code is publicly available at https://github.com/UCSC-REAL/negative-label-smoothing.

Open-Sampling: Exploring Out-of-Distribution data for Re-balancing Long-tailed datasets

- Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, Bo An
- abstract: Deep neural networks usually perform poorly when the training dataset suffers from extreme class imbalance. Recent studies found that directly training with out-of-distribution data (i.e., open-set samples) in a semi-supervised manner would harm the generalization performance. In this work, we theoretically show that out-of-distribution data can still be leveraged to augment the minority classes from a Bayesian perspective. Based on this motivation, we propose a novel method called Open-sampling, which utilizes open-set noisy labels to re-balance the class priors of the training dataset. For each open-set instance, the label is sampled from our pre-defined distribution that is complementary to the distribution of original class priors. We empirically show that Open-sampling not only re-balances the class priors but also encourages the neural network to learn separable representations. Extensive experiments demonstrate that our proposed method significantly outperforms existing data re-balancing methods and can boost the performance of existing state-of-the-art methods.

Mitigating Neural Network Overconfidence with Logit Normalization

- Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, Yixuan Li
- abstract: Detecting out-of-distribution inputs is critical for the safe deployment of machine learning models in the real world. However, neural networks are known to suffer from the overconfidence issue, where they produce abnormally high confidence for both in- and out-of-distribution inputs. In this work, we show that this issue can be mitigated through Logit Normalization (LogitNorm)—a simple fix to the cross-entropy loss—by enforcing a constant vector norm on the logits in training. Our method is motivated by the analysis that the norm of the logit keeps increasing during training, leading to overconfident output. Our key idea behind LogitNorm is thus to decouple the influence of output’s norm during network optimization. Trained with LogitNorm, neural networks produce highly distinguishable confidence scores between in- and out-of-distribution data. Extensive experiments demonstrate the superiority of LogitNorm, reducing the average FPR95 by up to 42.30% on common benchmarks.

Koopman Q-learning: Offline Reinforcement Learning via Symmetries of Dynamics

- Matthias Weissenbacher, Samarth Sinha, Animesh Garg, Kawahara Yoshinobu
- abstract: Offline reinforcement learning leverages large datasets to train policies without interactions with the environment. The learned policies may then be deployed in real-world settings where interactions are costly or dangerous. Current algorithms over-fit to the training dataset and as a consequence perform poorly when deployed to out-of-distribution generalizations of the environment. We aim to address these limitations by learning a Koopman latent representation which allows us to infer symmetries of the system’s underlying dynamic. The latter is then utilized to extend the otherwise static offline dataset during training; this constitutes a novel data augmentation framework which reflects the system’s dynamic and is thus to be interpreted as an exploration of the environments phase space. To obtain the symmetries we employ Koopman theory in which nonlinear dynamics are represented in terms of a linear operator acting on the space of measurement functions of the system. We provide novel theoretical results on the existence and nature of

symmetries relevant for control systems such as reinforcement learning settings. Moreover, we empirically evaluate our method on several benchmark offline reinforcement learning tasks and datasets including D4RL, Metaworld and Robosuite and find that by using our framework we consistently improve the state-of-the-art of model-free Q-learning methods.

Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification

- Yuxin Wen, Jonas A. Geiping, Liam Fowl, Micah Goldblum, Tom Goldstein
- abstract: Federated learning (FL) has rapidly risen in popularity due to its promise of privacy and efficiency. Previous works have exposed privacy vulnerabilities in the FL pipeline by recovering user data from gradient updates. However, existing attacks fail to address realistic settings because they either 1) require toy settings with very small batch sizes, or 2) require unrealistic and conspicuous architecture modifications. We introduce a new strategy that dramatically elevates existing attacks to operate on batches of arbitrarily large size, and without architectural modifications. Our model-agnostic strategy only requires modifications to the model parameters sent to the user, which is a realistic threat model in many scenarios. We demonstrate the strategy in challenging large-scale settings, obtaining high-fidelity data extraction in both cross-device and cross-silo federated learning. Code is available at https://github.com/JonasGeiping/breaching.

BabelTower: Learning to Auto-parallelized Program Translation

- Yuanbo Wen, Qi Guo, Qiang Fu, Xiaqing Li, Jianxing Xu, Yanlin Tang, Yongwei Zhao, Xing Hu, Zidong Du, Ling Li, Chao Wang, Xuehai Zhou, Yunji Chen
- abstract: GPUs have become the dominant computing platforms for many applications, while programming GPUs with the widely-used CUDA parallel programming model is difficult. As sequential C code is relatively easy to obtain either from legacy repositories or by manual implementation, automatically translating C to its parallel CUDA counterpart is promising to relieve the burden of GPU programming. However, because of huge differences between the sequential C and the parallel CUDA programming model, existing approaches fail to conduct the challenging auto-parallelized program translation. In this paper, we propose a learning-based framework, i.e., BabelTower, to address this problem. We first create a large-scale dataset consisting of compute-intensive function-level monolingual corpora. We further propose using back-translation with a discriminative reranker to cope with unpaired corpora and parallel semantic conversion. Experimental results show that BabelTower outperforms state-of-the-art by 1.79, 6.09, and 9.39 in terms of BLEU, CodeBLEU, and specifically designed ParaBLEU, respectively. The CUDA code generated by BabelTower attains a speedup of up to 347x over the sequential C code, and the developer productivity is improved by at most 3.8x.

Random Forest Density Estimation

- Hongwei Wen, Hanyuan Hang
- abstract: We propose a density estimation algorithm called random forest density estimation (RFDE) based on random trees where the split of cell is along the midpoint of the randomly chosen dimension. By combining the efficient random tree density estimation (RTDE) and the ensemble procedure, RFDE can alleviate the problems of boundary discontinuity suffered by partition-based density estimations. From the theoretical perspective, we first prove the fast convergence rates of RFDE if the density function lies in the Hölder space $C^{0,\alpha}$. Moreover, if the target function resides in the subspace $C^{1,\alpha}$, which contains smoother density functions, we for the first time manage to explain the benefits of the ensemble learning in density estimation. To be specific, we show that the upper bound of the ensemble estimator RFDE turns out to be strictly smaller than the lower bound of its base estimator RTDE in terms of convergence rates. In the experiments, we verify the theoretical results and show the promising performance of RFDE on both synthetic and real world datasets. Moreover, we evaluate our RFDE through the problem of anomaly detection as a possible application.

Fighting Fire with Fire: Avoiding DNN Shortcuts through Priming

- Chuan Wen, Jianing Qian, Jierui Lin, Jiaye Teng, Dinesh Jayaraman, Yang Gao
- abstract: Across applications spanning supervised classification and sequential control, deep learning has been reported to find “shortcut” solutions that fail catastrophically under minor changes in the data distribution. In this paper, we show empirically that DNNs can be coaxed to avoid poor shortcuts by providing an additional “priming” feature computed from key input features, usually a coarse output estimate. Priming relies on approximate domain knowledge of these task-relevant key input features, which is often easy to obtain in practical settings. For example, one might prioritize recent frames over past frames in a video input for visual imitation learning, or salient foreground over background pixels for image classification. On NICO image classification, MuJoCo continuous control, and CARLA autonomous driving, our priming strategy works significantly better than several popular state-of-the-art approaches for feature selection and data augmentation. We connect these empirical findings to recent theoretical results on DNN optimization, and argue theoretically that priming distracts the optimizer away from poor shortcuts by creating better, simpler shortcuts.

Preconditioning for Scalable Gaussian Process Hyperparameter Optimization

- Jonathan Wenger, Geoff Pleiss, Philipp Hennig, John Cunningham, Jacob Gardner
- abstract: Gaussian process hyperparameter optimization requires linear solves with, and log-determinants of, large kernel matrices. Iterative numerical techniques are becoming popular to scale to larger datasets, relying on the conjugate gradient method (CG) for the linear solves and stochastic trace estimation for the log-determinant. This work introduces new algorithmic and theoretical insights for preconditioning these computations. While preconditioning is well understood in the context of CG, we demonstrate that it can also accelerate convergence and reduce variance of the estimates for the log-determinant and its derivative. We prove general probabilistic error bounds for the preconditioned computation of the log-determinant, log-marginal likelihood and its derivatives. Additionally, we derive specific rates for a range of kernel-preconditioner combinations, showing that up to exponential convergence can be achieved. Our theoretical results enable provably efficient optimization of kernel hyperparameters, which we validate empirically on large-scale benchmark problems. There our approach accelerates training by up to an order of magnitude.

Measure Estimation in the Barycentric Coding Model

- Matthew Werenski, Ruijie Jiang, Abiy Tasissa, Shuchin Aeron, James M Murphy
- abstract: This paper considers the problem of measure estimation under the barycentric coding model (BCM), in which an unknown measure is assumed to belong to the set of Wasserstein-2 barycenters of a finite set of known measures. Estimating a measure under this model is equivalent to estimating the unknown barycentric coordinates. We provide novel geometrical, statistical, and computational insights for measure estimation under the BCM, consisting of three main results. Our first main result leverages the Riemannian geometry of Wasserstein-2 space to provide a procedure for recovering the barycentric coordinates as the solution to a quadratic optimization problem assuming access to the true reference measures. The essential geometric insight is that the parameters of this quadratic problem are determined by inner products between the optimal displacement maps from the given measure to the reference measures defining the BCM. Our second main result then establishes an algorithm for solving for the coordinates in the BCM when all the measures are observed empirically via i.i.d. samples. We prove precise rates of convergence for this algorithm—determined by the smoothness of the underlying measures and their dimensionality—thereby guaranteeing its statistical consistency. Finally, we demonstrate the utility of the BCM and associated estimation procedures in three application areas: (i) covariance estimation for Gaussian measures; (ii) image processing; and (iii) natural language processing.

COLA: Consistent Learning with Opponent-Learning Awareness

- Timon Willi, Alistair Hp Letcher, Johannes Treutlein, Jakob Foerster
- abstract: Learning in general-sum games is unstable and frequently leads to socially undesirable (Pareto-dominated) outcomes. To mitigate this, Learning with Opponent-Learning Awareness (LOLA) introduced opponent shaping to this setting, by accounting for each agent's influence on their opponents' anticipated learning steps. However, the original LOLA formulation (and follow-up work) is inconsistent because LOLA models other agents as naive learners rather than LOLA agents. In previous work, this inconsistency was suggested as a cause of LOLA's failure to preserve stable fixed points (SFPs). First, we formalize consistency and show that higher-order LOLA (HOLA) solves LOLA's inconsistency problem if it converges. Second, we correct a claim made in the literature by Sch{ä}fer and Anandkumar (2019), proving that Competitive Gradient Descent (CGD) does not recover HOLA as a series expansion (and fails to solve the consistency problem). Third, we propose a new method called Consistent LOLA (COLA), which learns update functions that are consistent under mutual opponent shaping. It requires no more than second-order derivatives and learns consistent update functions even when HOLA fails to converge. However, we also prove that even consistent update functions do not preserve SFPs, contradicting the hypothesis that this shortcoming is caused by LOLA's inconsistency. Finally, in an empirical evaluation on a set of general-sum games, we find that COLA finds prosocial solutions and that it converges under a wider range of learning rates than HOLA and LOLA. We support the latter finding with a theoretical result for a simple game.

Distributional Hamilton-Jacobi-Bellman Equations for Continuous-Time Reinforcement Learning

- Harley E Wiltzer, David Meger, Marc G. Bellemare
- abstract: Continuous-time reinforcement learning offers an appealing formalism for describing control problems in which the passage of time is not naturally divided into discrete increments. Here we consider the problem of predicting the distribution of returns obtained by an agent interacting in a continuous-time, stochastic environment. Accurate return predictions have proven useful for determining optimal policies for risk-sensitive control, learning state representations, multiagent coordination, and more. We begin by establishing the distributional analogue of the Hamilton-Jacobi-Bellman (HJB) equation for Ito diffusions and the broader class of Feller-Dynkin processes. We then specialize this equation to the setting in which the return distribution is approximated by N uniformly-weighted particles, a common design choice in distributional algorithms. Our derivation highlights additional terms due to statistical diffusivity which arise from the proper handling of distributions in the continuous-time setting. Based on this, we propose a tractable algorithm for approximately solving the distributional HJB based on a JKO scheme, which can be implemented in an online, control algorithm. We demonstrate the effectiveness of such an algorithm in a synthetic control problem.

Easy Variational Inference for Categorical Models via an Independent Binary Approximation

- Michael T Wojnowicz, Shuchin Aeron, Eric L Miller, Michael Hughes
- abstract: We pursue tractable Bayesian analysis of generalized linear models (GLMs) for categorical data. GLMs have been difficult to scale to more than a few dozen categories due to non-conjugacy or strong posterior dependencies when using conjugate auxiliary variable methods. We define a new class of GLMs for categorical data called categorical-from-binary (CB) models. Each CB model has a likelihood that is bounded by the product of binary likelihoods, suggesting a natural posterior approximation. This approximation makes inference straightforward and fast; using well-known auxiliary variables for probit or logistic regression, the product of binary models admits conjugate closed-form variational inference that is embarrassingly parallel across categories and invariant to category ordering. Moreover, an independent binary model simultaneously approximates multiple CB models. Bayesian model averaging over these can improve the quality of the approximation for any given dataset. We show that our approach scales to thousands of categories, outperforming posterior estimation competitors like Automatic Differentiation Variational Inference (ADVI) and No U-Turn Sampling (NUTS) in the time required to achieve fixed prediction quality.

Continual Learning with Guarantees via Weight Interval Constraints

- Maciej Wołczyk, Karol Piczak, Bartosz Wójcik, Lukasz Pustelnik, Paweł Morawiecki, Jacek Tabor, Tomasz Trzcinski, Przemysław Spurek
- abstract: We introduce a new training paradigm that enforces interval constraints on neural network parameter space to control forgetting. Contemporary Continual Learning (CL) methods focus on training neural networks efficiently from a stream of data, while reducing the negative impact of catastrophic forgetting, yet they do not provide any firm guarantees that network performance will not deteriorate uncontrollably over time. In this work, we show how to put bounds on forgetting by reformulating continual learning of a model as a continual contraction of its parameter space. To that end, we propose Hyperrectangle Training, a new training methodology where each task is represented by a hyperrectangle in the parameter space, fully contained in the hyperrectangles of the previous tasks. This formulation reduces the NP-hard CL problem back to polynomial time while providing full resilience against forgetting. We validate our claim by developing InterContiNet (Interval Continual Learning) algorithm which leverages interval arithmetic to effectively model parameter regions as hyperrectangles. Through experimental results, we show that our approach performs well in a continual learning setup without storing data from previous tasks.

A Deep Learning Approach for the Segmentation of Electroencephalography Data in Eye Tracking Applications

- Lukas Wolf, Ard Kastrati, Martyna B Plomecka, Jie-Ming Li, Dustin Klebe, Alexander Veicht, Roger Wattenhofer, Nicolas Langer
- abstract: The collection of eye gaze information provides a window into many critical aspects of human cognition, health and behaviour. Additionally, many neuroscientific studies complement the behavioural information gained from eye tracking with the high temporal resolution and neurophysiological markers provided by electroencephalography (EEG). One of the essential eye-tracking software processing steps is the segmentation of the continuous data stream into events relevant to eye-tracking applications, such as saccades, fixations, and blinks. Here, we introduce DETRtime, a novel framework for time-series segmentation that creates ocular event detectors that do not require additionally recorded eye-tracking modality and rely solely on EEG data. Our end-to-end deep-learning-based framework brings recent advances in Computer Vision to the forefront of the times series segmentation of EEG data. DETRtime achieves state-of-the-art performance in ocular event detection across diverse eye-tracking experiment paradigms. In addition to that, we provide evidence that our model generalizes well in the task of EEG sleep stage segmentation.

Leverage Score Sampling for Tensor Product Matrices in Input Sparsity Time

- David Woodruff, Amir Zandieh
- abstract: We propose an input sparsity time sampling algorithm that can spectrally approximate the Gram matrix corresponding to the q-fold column-wise tensor product of q matrices using a nearly optimal number of samples, improving upon all previously known methods by poly(q) factors. Furthermore, for the important special case of the q-fold self-tensoring of a dataset, which is the feature matrix of the degree-q polynomial kernel, the leading term of our method's runtime is proportional to the size of the dataset and has no dependence on q. Previous techniques either incur a poly(q) factor slowdown in their runtime or remove the dependence on q at the expense of having sub-optimal target dimension, and depend quadratically on the number of data-points in their runtime. Our sampling technique relies on a collection of q partially correlated random projections which can be simultaneously applied to a dataset X in total time that only depends on the size of X, and at the same time their q-fold Kronecker product acts as a near-isometry for any fixed vector in the column span of $X^{\otimes q}$. We also show that our sampling methods generalize to other classes of kernels beyond polynomial, such as Gaussian and Neural Tangent kernels.

Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, Ludwig Schmidt

- abstract: The conventional recipe for maximizing model accuracy is to (1) train multiple models with various hyperparameters and (2) pick the individual model which performs best on a held-out validation set, discarding the remainder. In this paper, we revisit the second step of this procedure in the context of fine-tuning large pre-trained models, where fine-tuned models often appear to lie in a single low error basin. We show that averaging the weights of multiple models fine-tuned with different hyperparameter configurations often improves accuracy and robustness. Unlike a conventional ensemble, we may average many models without incurring any additional inference or memory costs—we call the results "model soups." When fine-tuning large pre-trained models such as CLIP, ALIGN, and a ViT-G pre-trained on JFT, our soup recipe provides significant improvements over the best model in a hyperparameter sweep on ImageNet. The resulting ViT-G model, which attains 90.94% top-1 accuracy on ImageNet, achieved a new state of the art. Furthermore, we show that the model soup approach extends to multiple image classification and natural language processing tasks, improves out-of-distribution performance, and improves zero-shot performance on new downstream tasks. Finally, we analytically relate the performance similarity of weight-averaging and logit-ensembling to flatness of the loss and confidence of the predictions, and validate this relation empirically. Code is available at https://github.com/mlfoundations/model-soups.

Metric-Fair Classifier Derandomization

- Jimmy Wu, Yatong Chen, Yang Liu
- abstract: We study the problem of classifier derandomization in machine learning: given a stochastic binary classifier $f: X \to [0,1]$, sample a deterministic classifier $\hat{f}: X \to {0,1}$ that approximates the output of f in aggregate over any data distribution. Recent work revealed how to efficiently derandomize a stochastic classifier with strong output approximation guarantees, but at the cost of individual fairness — that is, if f treated similar inputs similarly, $\hat{f}$ did not. In this paper, we initiate a systematic study of classifier derandomization with metric fairness guarantees. We show that the prior derandomization approach is almost maximally metric-unfair, and that a simple "random threshold" derandomization achieves optimal fairness preservation but with weaker output approximation. We then devise a derandomization procedure that provides an appealing tradeoff between these two: if f is α-metric fair according to a metric d with a locality-sensitive hash (LSH) family, then our derandomized $\hat{f}$ is, with high probability, $O(\alpha)$-metric fair and a close approximation of f. We also prove generic results applicable to all (fair and unfair) classifier derandomization procedures, including a bias-variance decomposition and reductions between various notions of metric fairness.

Structural Entropy Guided Graph Hierarchical Pooling

- Junran Wu, Xueyuan Chen, Ke Xu, Shangzhe Li
- abstract: Following the success of convolution on non-Euclidean space, the corresponding pooling approaches have also been validated on various tasks regarding graphs. However, because of the fixed compression ratio and stepwise pooling design, these hierarchical pooling methods still suffer from local structure damage and suboptimal problem. In this work, inspired by structural entropy, we propose a hierarchical pooling approach, SEP, to tackle the two issues. Specifically, without assigning the layer-specific compression ratio, a global optimization algorithm is designed to generate the cluster assignment matrices for pooling at once. Then, we present an illustration of the local structure damage from previous methods in reconstruction of ring and grid synthetic graphs. In addition to SEP, we further design two classification models, SEP-G and SEP-N for graph classification and node classification, respectively. The results show that SEP outperforms state-of-the-art graph pooling methods on graph classification benchmarks and obtains superior performance on node classifications.

Self-supervised Models are Good Teaching Assistants for Vision Transformers

- Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, Ke Li
- abstract: Transformers have shown remarkable progress on computer vision tasks in the past year. Compared to their CNN counterparts, transformers usually need the help of distillation to achieve comparable results on middle or small sized datasets. Meanwhile, recent researches discover that when transformers are trained with supervised and self-supervised manner respectively, the captured patterns are quite different both qualitatively and quantitatively. These findings motivate us to introduce an self-supervised teaching assistant (SSTA) besides the commonly used supervised teacher to improve the performance of transformers. Specifically, we propose a head-level knowledge distillation method that selects the most important head of the supervised teacher and self-supervised teaching assistant, and let the student mimic the attention distribution of these two heads, so as to make the student focus on the relationship between tokens deemed by the teacher and the teacher assistant. Extensive experiments verify the effectiveness of SSTA and demonstrate that the proposed SSTA is a good compensation to the supervised teacher. Meanwhile, some analytical experiments towards multiple perspectives (e.g. prediction, shape bias, robustness, and transferability to downstream tasks) with supervised teachers, self-supervised teaching assistants and students are inductive and may inspire future researches.

Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks

- Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, Krzysztof J Geras
- abstract: We hypothesize that due to the greedy nature of learning in multi-modal deep neural networks, these models tend to rely on just one modality while under-fitting the other modalities. Such behavior is counter-intuitive and hurts the models' generalization, as we observe empirically. To estimate the model's dependence on each modality, we compute the gain on the accuracy when the model has access to it in addition to another modality. We refer to this gain as the conditional utilization rate. In the experiments, we consistently observe an imbalance in conditional utilization rates between modalities, across multiple tasks and architectures. Since conditional utilization rate cannot be computed efficiently during training, we introduce a proxy for it based on the pace at which the model learns from each modality, which we refer to as the conditional learning speed. We propose an algorithm to balance the conditional learning speeds between modalities during training and demonstrate that it indeed addresses the issue of greedy learning. The proposed algorithm improves the model's generalization on three datasets: Colored MNIST, ModelNet40, and NVIDIA Dynamic Hand Gesture.

Instrumental Variable Regression with Confounder Balancing

- Anpeng Wu, Kun Kuang, Bo Li, Fei Wu
- abstract: This paper considers the challenge of estimating treatment effects from observational data in the presence of unmeasured confounders. A popular way to address this challenge is to utilize an instrumental variable (IV) for two-stage regression, i.e., 2SLS and variants, but limited to the linear setting. Recently, many nonlinear IV regression variants were proposed to overcome it by regressing the treatment with IVs and observed confounders in stage 1, leading to the imbalance of the observed confounders in stage 2. In this paper, we propose a Confounder Balanced IV Regression (CB-IV) algorithm to jointly remove the bias from the unmeasured confounders and balance the observed confounders. To the best of our knowledge, this is the first work to combine confounder balancing in IV regression for treatment effect estimation. Theoretically, we re-define and solve the inverse problems for the response-outcome function. Experiments show that our algorithm outperforms the existing approaches.

MemSR: Training Memory-efficient Lightweight Model for Image Super-Resolution

- Kailu Wu, Chung-Kuei Lee, Kaisheng Ma
- abstract: Methods based on deep neural networks with a massive number of layers and skip-connections have made impressive improvements on single image super-resolution (SISR). The skip-connections in these complex models boost the performance at the cost of a large amount of memory. With the increase of camera resolution from 1 million pixels to 100 million pixels on mobile phones, the memory footprint of these algorithms also increases hundreds of times, which restricts the applicability of these models on memory-limited devices. A plain model consisting of a stack of 3{\texttimes}3 convolutions with ReLU, in contrast, has the highest memory efficiency but poorly performs on super-resolution. This paper aims at calculating a winning

initialization from a complex teacher network for a plain student network, which can provide performance comparable to complex models. To this end, we convert the teacher model to an equivalent large plain model and derive the plain student's initialization. We further improve the student's performance through initialization-aware feature distillation. Extensive experiments suggest that the proposed method results in a model with a competitive trade-off between accuracy and speed at a much lower memory footprint than other state-of-the-art lightweight approaches.

Delay-Adaptive Step-sizes for Asynchronous Learning

- Xuyang Wu, Sindri Magnusson, Hamid Reza Feyzmahdavian, Mikael Johansson
- abstract: In scalable machine learning systems, model training is often parallelized over multiple nodes that run without tight synchronization. Most analysis results for the related asynchronous algorithms use an upper bound on the information delays in the system to determine learning rates. Not only are such bounds hard to obtain in advance, but they also result in unnecessarily slow convergence. In this paper, we show that it is possible to use learning rates that depend on the actual time-varying delays in the system. We develop general convergence results for delay-adaptive asynchronous iterations and specialize these to proximal incremental gradient descent and block coordinate descent algorithms. For each of these methods, we demonstrate how delays can be measured on-line, present delay-adaptive step-size policies, and illustrate their theoretical and practical advantages over the state-of-the-art.

Variational nearest neighbor Gaussian process

- Luhuan Wu, Geoff Pleiss, John P Cunningham
- abstract: Variational approximations to Gaussian processes (GPs) typically use a small set of inducing points to form a low-rank approximation to the covariance matrix. In this work, we instead exploit a sparse approximation of the precision matrix. We propose variational nearest neighbor Gaussian process (VNNGP), which introduces a prior that only retains correlations within K nearest-neighboring observations, thereby inducing sparse precision structure. Using the variational framework, VNNGP's objective can be factorized over both observations and inducing points, enabling stochastic optimization with a time complexity of $O(K^3)$. Hence, we can arbitrarily scale the inducing point size, even to the point of putting inducing points at every observed location. We compare VNNGP to other scalable GPs through various experiments, and demonstrate that VNNGP (1) can dramatically outperform low-rank methods, and (2) is less prone to overfitting than other nearest neighbor methods.

Understanding Policy Gradient Algorithms: A Sensitivity-Based Approach

- Shuang Wu, Ling Shi, Jun Wang, Guangjian Tian
- abstract: The REINFORCE algorithm \cite{williams1992simple} is popular in policy gradient (PG) for solving reinforcement learning (RL) problems. Meanwhile, the theoretical form of PG is from \cite{sutton1999policy}. Although both formulae prescribe PG, their precise connections are not yet illustrated. Recently, \citeauthor{nota2020policy} (\citeyear{nota2020policy}) have found that the ambiguity causes implementation errors. Motivated by the ambiguity and implementation incorrectness, we study PG from a perturbation perspective. In particular, we derive PG in a unified framework, precisely clarify the relation between PG implementation and theory, and echos back the findings by \citeauthor{nota2020policy}. Diving into factors contributing to empirical successes of the existing erroneous implementations, we find that small approximation error and the experience replay mechanism play critical roles.

DAVINZ: Data Valuation using Deep Neural Networks at Initialization

- Zhaoxuan Wu, Yao Shu, Bryan Kian Hsiang Low
- abstract: Recent years have witnessed a surge of interest in developing trustworthy methods to evaluate the value of data in many real-world applications (e.g., collaborative machine learning, data marketplaces). Existing data valuation methods typically valuate data using the generalization performance of converged machine learning models after their long-term model training, hence making data valuation on large complex deep neural networks (DNNs) unaffordable. To this end, we theoretically derive a domain-aware generalization bound to estimate the generalization performance of DNNs without model training. We then exploit this theoretically derived generalization bound to develop a novel training-free data valuation method named data valuation at initialization (DAVINZ) on DNNs, which consistently achieves remarkable effectiveness and efficiency in practice. Moreover, our training-free DAVINZ, surprisingly, can even theoretically and empirically enjoy the desirable properties that training-based data valuation methods usually attain, thus making it more trustworthy in practice.

Robust Deep Reinforcement Learning through Bootstrapped Opportunistic Curriculum

- Junlin Wu, Yevgeniy Vorobeychik
- abstract: Despite considerable advances in deep reinforcement learning, it has been shown to be highly vulnerable to adversarial perturbations to state observations. Recent efforts that have attempted to improve adversarial robustness of reinforcement learning can nevertheless tolerate only very small perturbations, and remain fragile as perturbation size increases. We propose Bootstrapped Opportunistic Adversarial Curriculum Learning (BCL), a novel flexible adversarial curriculum learning framework for robust reinforcement learning. Our framework combines two ideas: conservatively bootstrapping each curriculum phase with highest quality solutions obtained from multiple runs of the previous phase, and opportunistically skipping forward in the curriculum. In our experiments we show that the proposed BCL framework enables dramatic improvements in robustness of learned policies to adversarial perturbations. The greatest improvement is for Pong, where our framework yields robustness to perturbations of up to 25/255; in contrast, the best existing approach can only tolerate adversarial noise up to 5/255. Our code is available at: https://github.com/jlwu002/BCL.

Revisiting Consistency Regularization for Deep Partial Label Learning

- Dong-Dong Wu, Deng-Bao Wang, Min-Ling Zhang
- abstract: Partial label learning (PLL), which refers to the classification task where each training instance is ambiguously annotated with a set of candidate labels, has been recently studied in deep learning paradigm. Despite advances in recent deep PLL literature, existing methods (e.g., methods based on self-training or contrastive learning) are confronted with either ineffectiveness or inefficiency. In this paper, we revisit a simple idea namely consistency regularization, which has been shown effective in traditional PLL literature, to guide the training of deep models. Towards this goal, a new regularized training framework, which performs supervised learning on non-candidate labels and employs consistency regularization on candidate labels, is proposed for PLL. We instantiate the regularization term by matching the outputs of multiple augmentations of an instance to a conformal label distribution, which can be adaptively inferred by the closed-form solution. Experiments on benchmark datasets demonstrate the superiority of the proposed method compared with other state-of-the-art methods.

Flowformer: Linearizing Transformers with Conservation Flows

- Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, Mingsheng Long
- abstract: Transformers based on the attention mechanism have achieved impressive success in various areas. However, the attention mechanism has a quadratic complexity, significantly impeding Transformers from dealing with numerous tokens and scaling up to bigger models. Previous methods mainly utilize the similarity decomposition and the associativity of matrix multiplication to devise linear-time attention mechanisms. They avoid degeneration of attention to a trivial distribution by reintroducing inductive biases such as the locality, thereby at the expense of model generality and expressiveness. In this paper, we linearize Transformers free from specific inductive biases based on the flow network theory. We cast attention as the information flow

aggregated from the sources (values) to the sinks (results) through the learned flow capacities (attentions). Within this framework, we apply the property of flow conservation into attention and propose the Flow-Attention mechanism of linear complexity. By respectively conserving the incoming flow of sinks for source competition and the outgoing flow of sources for sink allocation, Flow-Attention inherently generates informative attentions without using specific inductive biases. Empowered by the Flow-Attention, Flowformer yields strong performance in linear time for wide areas, including long sequence, time series, vision, natural language, and reinforcement learning. The code and settings are available at this repository: https://github.com/thuml/Flowformer.

Nearly Optimal Policy Optimization with Stable at Any Time Guarantee

- Tianhao Wu, Yunchang Yang, Han Zhong, Liwei Wang, Simon Du, Jiantao Jiao
- abstract: Policy optimization methods are one of the most widely used classes of Reinforcement Learning (RL) algorithms. However, theoretical understanding of these methods remains insufficient. Even in the episodic (time-inhomogeneous) tabular setting, the state-of-the-art theoretical result of policy-based method in Shani et al. (2020) is only $\tilde{O}(\sqrt{S^2AH^4K})$ where S is the number of states, A is the number of actions, H is the horizon, and K is the number of episodes, and there is a $\sqrt{SH}$ gap compared with the information theoretic lower bound $\tilde{\Omega}(\sqrt{SAH^3K})$ (Jin et al., 2018). To bridge such a gap, we propose a novel algorithm Reference-based Policy Optimization with Stable at Any Time guarantee (RPO-SAT), which features the property “Stable at Any Time”. We prove that our algorithm achieves $\tilde{O}(\sqrt{SAH^3K} + \sqrt{AH^4K})$ regret. When $S > H$, our algorithm is minimax optimal when ignoring logarithmic factors. To our best knowledge, RPO-SAT is the first computationally efficient, nearly minimax optimal policy-based algorithm for tabular RL.

RetrievalGuard: Provably Robust 1-Nearest Neighbor Image Retrieval

- Yihan Wu, Hongyang Zhang, Heng Huang
- abstract: Recent research works have shown that image retrieval models are vulnerable to adversarial attacks, where slightly modified test inputs could lead to problematic retrieval results. In this paper, we aim to design a provably robust image retrieval model which keeps the most important evaluation metric Recall@1 invariant to adversarial perturbation. We propose the first 1-nearest neighbor (NN) image retrieval algorithm, RetrievalGuard, which is provably robust against adversarial perturbations within an ℓ_2 ball of calculable radius. The challenge is to design a provably robust algorithm that takes into consideration the 1-NN search and the high-dimensional nature of the embedding space. Algorithmically, given a base retrieval model and a query sample, we build a smoothed retrieval model by carefully analyzing the 1-NN search procedure in the high-dimensional embedding space. We show that the smoothed retrieval model has bounded Lipschitz constant and thus the retrieval score is invariant to ℓ_2 adversarial perturbations. Experiments on on image retrieval tasks validate the robustness of our RetrievalGuard method.

Last Iterate Risk Bounds of SGD with Decaying Stepsize for Overparameterized Linear Regression

- Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, Sham Kakade
- abstract: Stochastic gradient descent (SGD) has been shown to generalize well in many deep learning applications. In practice, one often runs SGD with a geometrically decaying stepsize, i.e., a constant initial stepsize followed by multiple geometric stepsize decay, and uses the last iterate as the output. This kind of SGD is known to be nearly minimax optimal for classical finite-dimensional linear regression problems (Ge et al., 2019). However, a sharp analysis for the last iterate of SGD in the overparameterized setting is still open. In this paper, we provide a problem-dependent analysis on the last iterate risk bounds of SGD with decaying stepsize, for (overparameterized) linear regression problems. In particular, for last iterate SGD with (tail) geometrically decaying stepsize, we prove nearly matching upper and lower bounds on the excess risk. Moreover, we provide an excess risk lower bound for last iterate SGD with polynomially decaying stepsize and demonstrate the advantage of geometrically decaying stepsize in an instance-wise manner, which complements the minimax rate comparison made in prior work.

Optimal Clustering with Noisy Queries via Multi-Armed Bandit

- Jinghui Xia, Zengfeng Huang
- abstract: Motivated by many applications, we study clustering with a faulty oracle. In this problem, there are n items belonging to k unknown clusters, and the algorithm is allowed to ask the oracle whether two items belong to the same cluster or not. However, the answer from the oracle is correct only with probability $\frac{1}{2}+\frac{\delta}{2}$. The goal is to recover the hidden clusters with minimum number of noisy queries. Previous works have shown that the problem can be solved with $O(\frac{nk\log n}{\delta^2} + \text{poly}(k,\frac{1}{\delta}, \log n))$ queries, while $\Omega(\frac{nk}{\delta^2})$ queries is known to be necessary. So, for any values of k and δ, there is still a non-trivial gap between upper and lower bounds. In this work, we obtain the first matching upper and lower bounds for a wide range of parameters. In particular, a new polynomial time algorithm with $O(\frac{n(k+\log n)}{\delta^2} + \text{poly}(k,\frac{1}{\delta}, \log n))$ queries is proposed. Moreover, we prove a new lower bound of $\Omega(\frac{n\log n}{\delta^2})$, which, combined with the existing $\Omega(\frac{nk}{\delta^2})$ bound, matches our upper bound up to an additive $\text{poly}(k,\frac{1}{\delta},\log n)$ term. To obtain the new results, our main ingredient is an interesting connection between our problem and multi-armed bandit, which might provide useful insights for other similar problems.

ProGCL: Rethinking Hard Negative Mining in Graph Contrastive Learning

- Jun Xia, Lirong Wu, Ge Wang, Jintao Chen, Stan Z. Li
- abstract: Contrastive Learning (CL) has emerged as a dominant technique for unsupervised representation learning which embeds augmented versions of the anchor close to each other (positive samples) and pushes the embeddings of other samples (negatives) apart. As revealed in recent studies, CL can benefit from hard negatives (negatives that are most similar to the anchor). However, we observe limited benefits when we adopt existing hard negative mining techniques of other domains in Graph Contrastive Learning (GCL). We perform both experimental and theoretical analysis on this phenomenon and find it can be attributed to the message passing of Graph Neural Networks (GNNs). Unlike CL in other domains, most hard negatives are potentially false negatives (negatives that share the same class with the anchor) if they are selected merely according to the similarities between anchor and themselves, which will undesirably push away the samples of the same class. To remedy this deficiency, we propose an effective method, dubbed \textbf{ProGCL}, to estimate the probability of a negative being true one, which constitutes a more suitable measure for negatives’ hardness together with similarity. Additionally, we devise two schemes (i.e., \textbf{ProGCL-weight} and \textbf{ProGCL-mix}) to boost the performance of GCL. Extensive experiments demonstrate that ProGCL brings notable and consistent improvements over base GCL methods and yields multiple state-of-the-art results on several unsupervised benchmarks or even exceeds the performance of supervised ones. Also, ProGCL is readily pluggable into various negatives-based GCL methods for performance improvement. We release the code at \textcolor{magenta}\url{https://github.com/junxia97/ProGCL}.

Synergy and Symmetry in Deep Learning: Interactions between the Data, Model, and Inference Algorithm

- Lechao Xiao, Jeffrey Pennington
- abstract: Although learning in high dimensions is commonly believed to suffer from the curse of dimensionality, modern machine learning methods often exhibit an astonishing power to tackle a wide range of challenging real-world learning problems without using abundant amounts of data. How exactly these methods break this curse remains a fundamental open question in the theory of deep learning. While previous efforts have investigated this question by studying the data ($\mathcal D$), model ($\mathcal M$), and inference algorithm ($\mathcal I$) as independent modules, in this paper we analyzes the triplet $(\mathcal D, \mathcal M, \mathcal I)$ as an integrated system and identify important synergies that help mitigate the curse of dimensionality. We first study the basic symmetries associated with various learning algorithms ($\mathcal M, \mathcal I$), focusing on four prototypical architectures in deep

learning: fully-connected networks, locally-connected networks, and convolutional networks with and without pooling. We find that learning is most efficient when these symmetries are compatible with those of the data distribution and that performance significantly deteriorates when any member of the \dmi triplet is inconsistent or suboptimal.

Identification of Linear Non-Gaussian Latent Hierarchical Structure

- Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, Kun Zhang
- abstract: Traditional causal discovery methods mainly focus on estimating causal relations among measured variables, but in many real-world problems, such as questionnaire-based psychometric studies, measured variables are generated by latent variables that are causally related. Accordingly, this paper investigates the problem of discovering the hidden causal variables and estimating the causal structure, including both the causal relations among latent variables and those between latent and measured variables. We relax the frequently-used measurement assumption and allow the children of latent variables to be latent as well, and hence deal with a specific type of latent hierarchical causal structure. In particular, we define a minimal latent hierarchical structure and show that for linear non-Gaussian models with the minimal latent hierarchical structure, the whole structure is identifiable from only the measured variables. Moreover, we develop a principled method to identify the structure by testing for Generalized Independent Noise (GIN) conditions in specific ways. Experimental results on both synthetic and real-world data show the effectiveness of the proposed approach.

COAT: Measuring Object Compositionality in Emergent Representations

- Sirui Xie, Ari S Morcos, Song-Chun Zhu, Ramakrishna Vedantam
- abstract: Learning representations that can decompose a multi-object scene into its constituent objects and recompose them flexibly is desirable for object-oriented reasoning and planning. Built upon object masks in the pixel space, existing metrics for objectness can only evaluate generative models with an object-specific “slot” structure. We propose to directly measure compositionality in the representation space as a form of objections, making such evaluations tractable for a wider class of models. Our metric, COAT (Compositional Object Algebra Test), evaluates if a generic representation exhibits certain geometric properties that underpin object compositionality beyond what is already captured by the raw pixel space. Our experiments on the popular CLEVR (Johnson et.al., 2018) domain reveal that existing disentanglement-based generative models are not as compositional as one might expect, suggesting room for further modeling improvements. We hope our work allows for a unified evaluation of object-centric representations, spanning generative as well as discriminative, self-supervised models.

Robust Policy Learning over Multiple Uncertainty Sets

- Annie Xie, Shagun Sodhani, Chelsea Finn, Joelle Pineau, Amy Zhang
- abstract: Reinforcement learning (RL) agents need to be robust to variations in safety-critical environments. While system identification methods provide a way to infer the variation from online experience, they can fail in settings where fast identification is not possible. Another dominant approach is robust RL which produces a policy that can handle worst-case scenarios, but these methods are generally designed to achieve robustness to a single uncertainty set that must be specified at train time. Towards a more general solution, we formulate the multi-set robustness problem to learn a policy robust to different perturbation sets. We then design an algorithm that enjoys the benefits of both system identification and robust RL: it reduces uncertainty where possible given a few interactions, but can still act robustly with respect to the remaining uncertainty. On a diverse set of control tasks, our approach demonstrates improved worst-case performance on new environments compared to prior methods based on system identification and on robust RL alone.

Adaptive Inertia: Disentangling the Effects of Adaptive Learning Rate and Momentum

- Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, Masashi Sugiyama
- abstract: Adaptive Moment Estimation (Adam), which combines Adaptive Learning Rate and Momentum, would be the most popular stochastic optimizer for accelerating the training of deep neural networks. However, it is empirically known that Adam often generalizes worse than Stochastic Gradient Descent (SGD). The purpose of this paper is to unveil the mystery of this behavior in the diffusion theoretical framework. Specifically, we disentangle the effects of Adaptive Learning Rate and Momentum of the Adam dynamics on saddle-point escaping and flat minima selection. We prove that Adaptive Learning Rate can escape saddle points efficiently, but cannot select flat minima as SGD does. In contrast, Momentum provides a drift effect to help the training process pass through saddle points, and almost does not affect flat minima selection. This partly explains why SGD (with Momentum) generalizes better, while Adam generalizes worse but converges faster. Furthermore, motivated by the analysis, we design a novel adaptive optimization framework named Adaptive Inertia, which uses parameter-wise adaptive inertia to accelerate the training and provably favors flat minima as well as SGD. Our extensive experiments demonstrate that the proposed adaptive inertia method can generalize significantly better than SGD and conventional adaptive gradient methods.

Self-Supervised Representation Learning via Latent Graph Prediction

- Yaochen Xie, Zhao Xu, Shuiwang Ji
- abstract: Self-supervised learning (SSL) of graph neural networks is emerging as a promising way of leveraging unlabeled data. Currently, most methods are based on contrastive learning adapted from the image domain, which requires view generation and a sufficient number of negative samples. In contrast, existing predictive models do not require negative sampling, but lack theoretical guidance on the design of pretext training tasks. In this work, we propose the LaGraph, a theoretically grounded predictive SSL framework based on latent graph prediction. Learning objectives of LaGraph are derived as self-supervised upper bounds to objectives for predicting unobserved latent graphs. In addition to its improved performance, LaGraph provides explanations for recent successes of predictive models that include invariance-based objectives. We provide theoretical analysis comparing LaGraph to related methods in different domains. Our experimental results demonstrate the superiority of LaGraph in performance and the robustness to decreasing of training sample size on both graph-level and node-level tasks.

Efficient Computation of Higher-Order Subgraph Attribution via Message Passing

- Ping Xiong, Thomas Schnake, Grégoire Montavon, Klaus-Robert Müller, Shinichi Nakajima
- abstract: Explaining graph neural networks (GNNs) has become more and more important recently. Higher-order interpretation schemes, such as GNN-LRP (layer-wise relevance propagation for GNN), emerged as powerful tools for unraveling how different features interact thereby contributing to explaining GNNs. GNN-LRP gives a relevance attribution of walks between nodes at each layer, and the subgraph attribution is expressed as a sum over exponentially many such walks. In this work, we demonstrate that such exponential complexity can be avoided. In particular, we propose novel algorithms that enable to attribute subgraphs with GNN-LRP in linear-time (w.r.t. the network depth). Our algorithms are derived via message passing techniques that make use of the distributive property, thereby directly computing quantities for higher-order explanations. We further adapt our efficient algorithms to compute a generalization of subgraph attributions that also takes into account the neighboring graph features. Experimental results show the significant acceleration of the proposed algorithms and demonstrate the high usefulness and scalability of our novel generalized subgraph attribution method.

A Self-Play Posterior Sampling Algorithm for Zero-Sum Markov Games

- Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Tong Zhang

- abstract: Existing studies on provably efficient algorithms for Markov games (MGs) almost exclusively build on the “optimism in the face of uncertainty” (OFU) principle. This work focuses on a distinct approach of posterior sampling, which is celebrated in many bandits and reinforcement learning settings but remains under-explored for MGs. Specifically, for episodic two-player zero-sum MGs, a novel posterior sampling algorithm is developed with general function approximation. Theoretical analysis demonstrates that the posterior sampling algorithm admits a $\sqrt{T}$-regret bound for problems with a low multi-agent decoupling coefficient, which is a new complexity measure for MGs, where T denotes the number of episodes. When specializing to linear MGs, the obtained regret bound matches the state-of-the-art results. To the best of our knowledge, this is the first provably efficient posterior sampling algorithm for MGs with frequentist regret guarantees, which extends the toolbox for MGs and promotes the broad applicability of posterior sampling.

Importance Weighted Kernel Bayes’ Rule

- Liyuan Xu, Yutian Chen, Arnaud Doucet, Arthur Gretton
- abstract: We study a nonparametric approach to Bayesian computation via feature means, where the expectation of prior features is updated to yield expected posterior features, based on regression from kernel or neural net features of the observations. All quantities involved in the Bayesian update are learned from observed data, making the method entirely model-free. The resulting algorithm is a novel instance of a kernel Bayes’ rule (KBR). Our approach is based on importance weighting, which results in superior numerical stability to the existing approach to KBR, which requires operator inversion. We show the convergence of the estimator using a novel consistency analysis on the importance weighting estimator in the infinity norm. We evaluate our KBR on challenging synthetic benchmarks, including a filtering problem with a state-space model involving high dimensional image observations. The proposed method yields uniformly better empirical performance than the existing KBR, and competitive performance with other competing methods. We evaluate our KBR on challenging synthetic benchmarks, including a filtering problem with a state-space model involving high dimensional image observations. The proposed method yields uniformly better empirical performance than the existing KBR, and competitive performance with other competing methods.

Learning to Separate Voices by Spatial Regions

- Alan Xu, Romit Roy Choudhury
- abstract: We consider the problem of audio voice separation for binaural applications, such as earphones and hearing aids. While today’s neural networks perform remarkably well (separating 4+ sources with 2 microphones) they assume a known or fixed maximum number of sources, K. Moreover, today’s models are trained in a supervised manner, using training data synthesized from generic sources, environments, and human head shapes. This paper intends to relax both these constraints at the expense of a slight alteration in the problem definition. We observe that, when a received mixture contains too many sources, it is still helpful to separate them by region, i.e., isolating signal mixtures from each conical sector around the user’s head. This requires learning the fine-grained spatial properties of each region, including the signal distortions imposed by a person’s head. We propose a two-stage self-supervised framework in which overheard voices from earphones are pre-processed to extract relatively clean personalized signals, which are then used to train a region-wise separation model. Results show promising performance, underscoring the importance of personalization over a generic supervised approach. (audio samples available at our project website: https://uiuc-earable-computing.github.io/binaural). We believe this result could help real-world applications in selective hearing, noise cancellation, and audio augmented reality.

Detached Error Feedback for Distributed SGD with Random Sparsification

- An Xu, Heng Huang
- abstract: The communication bottleneck has been a critical problem in large-scale distributed deep learning. In this work, we study distributed SGD with random block-wise sparsification as the gradient compressor, which is ring-allreduce compatible and highly computation-efficient but leads to inferior performance. To tackle this important issue, we improve the communication-efficient distributed SGD from a novel aspect, that is, the trade-off between the variance and second moment of the gradient. With this motivation, we propose a new detached error feedback (DEF) algorithm, which shows better convergence bound than error feedback for non-convex problems. We also propose DEF-A to accelerate the generalization of DEF at the early stages of the training, which shows better generalization bounds than DEF. Furthermore, we establish the connection between communication-efficient distributed SGD and SGD with iterate averaging (SGD-IA) for the first time. Extensive deep learning experiments show significant empirical improvement of the proposed methods under various settings. Our reproducible codes and scripts for all experiments in this work will be made publicly available.

Accurate Quantization of Measures via Interacting Particle-based Optimization

- Lantian Xu, Anna Korba, Dejan Slepcev
- abstract: Approximating a target probability distribution can be cast as an optimization problem where the objective functional measures the dissimilarity to the target. This optimization can be addressed by approximating Wasserstein and related gradient flows. In practice, these are simulated by interacting particle systems, whose stationary states define an empirical measure approximating the target distribution. This approach has been popularized recently to design sampling algorithms, e.g. Stein Variational Gradient Descent, or by minimizing the Maximum Mean or Kernel Stein Discrepancy. However, little is known about quantization properties of these approaches, i.e. how well is the target approximated by a finite number particles. We investigate this question theoretically and numerically. In particular, we prove general upper bounds on the quantization error of MMD and KSD at rates which significantly outperform quantization by i.i.d. samples. We conduct experiments which show that the particle systems at study achieve fast rates in practice, and notably outperform greedy algorithms, such as kernel herding. We compare different gradient flows and highlight their quantization rates. Furthermore we introduce a Normalized Stein Variational Gradient Descent and argue in favor of adaptive kernels, which exhibit faster convergence. Finally we compare the Gaussian and Laplace kernels and argue that the Laplace kernel provides a more robust quantization.

Unified Fourier-based Kernel and Nonlinearity Design for Equivariant Networks on Homogeneous Spaces

- Yinshuang Xu, Jiahui Lei, Edgar Dobriban, Kostas Daniilidis
- abstract: We introduce a unified framework for group equivariant networks on homogeneous spaces derived from a Fourier perspective. We consider tensor-valued feature fields, before and after a convolutional layer. We present a unified derivation of kernels via the Fourier domain by leveraging the sparsity of Fourier coefficients of the lifted feature fields. The sparsity emerges when the stabilizer subgroup of the homogeneous space is a compact Lie group. We further introduce a nonlinear activation, via an elementwise nonlinearity on the regular representation after lifting and projecting back to the field through an equivariant convolution. We show that other methods treating features as the Fourier coefficients in the stabilizer subgroup are special cases of our activation. Experiments on $SO(3)$ and $SE(3)$ show state-of-the-art performance in spherical vector field regression, point cloud classification, and molecular completion.

Inferring Cause and Effect in the Presence of Heteroscedastic Noise

- Sascha Xu, Osman A Mian, Alexander Marx, Jilles Vreeken
- abstract: We study the problem of identifying cause and effect over two univariate continuous variables X and Y from a sample of their joint distribution. Our focus lies on the setting when the variance of the noise may be dependent on the cause. We propose to partition the domain of the cause into multiple segments where the noise indeed is dependent. To this end, we minimize a scale-invariant, penalized regression score, finding the optimal partitioning using dynamic programming. We show under which conditions this allows us to identify the causal direction for the linear setting with heteroscedastic noise, for the non-linear setting with homoscedastic noise, as well as empirically confirm that these results generalize to the non-linear and

heteroscedastic case. Altogether, the ability to model heteroscedasticity translates into an improved performance in telling cause from effect on a wide range of synthetic and real-world datasets.

Prompting Decision Transformer for Few-Shot Policy Generalization

- Mengdi Xu, Yikang Shen, Shun Zhang, Yuchen Lu, Ding Zhao, Joshua Tenenbaum, Chuang Gan
- abstract: Human can leverage prior experience and learn novel tasks from a handful of demonstrations. In contrast to offline meta-reinforcement learning, which aims to achieve quick adaptation through better algorithm design, we investigate the effect of architecture inductive bias on the few-shot learning capability. We propose a Prompt-based Decision Transformer (Prompt-DT), which leverages the sequential modeling ability of the Transformer architecture and the prompt framework to achieve few-shot adaptation in offline RL. We design the trajectory prompt, which contains segments of the few-shot demonstrations, and encodes task-specific information to guide policy generation. Our experiments in five MuJoCo control benchmarks show that Prompt-DT is a strong few-shot learner without any extra finetuning on unseen target tasks. Prompt-DT outperforms its variants and strong meta offline RL baselines by a large margin with a trajectory prompt containing only a few timesteps. Prompt-DT is also robust to prompt length changes and can generalize to out-of-distribution (OOD) environments. Project page: \href{https://mxu34.github.io/PromptDT/}{https://mxu34.github.io/PromptDT/}.

Analyzing and Mitigating Interference in Neural Architecture Search

- Jin Xu, Xu Tan, Kaitao Song, Renqian Luo, Yichong Leng, Tao Qin, Tie-Yan Liu, Jian Li
- abstract: Weight sharing is a popular approach to reduce the training cost of neural architecture search (NAS) by reusing the weights of shared operators from previously trained child models. However, the rank correlation between the estimated accuracy and ground truth accuracy of those child models is low due to the interference among different child models caused by weight sharing. In this paper, we investigate the interference issue by sampling different child models and calculating the gradient similarity of shared operators, and observe that: 1) the interference on a shared operator between two child models is positively correlated with the number of different operators between them; 2) the interference is smaller when the inputs and outputs of the shared operator are more similar. Inspired by these two observations, we propose two approaches to mitigate the interference: 1) rather than randomly sampling child models for optimization, we propose a gradual modification scheme by modifying one operator between adjacent optimization steps to minimize the interference on the shared operators; 2) forcing the inputs and outputs of the operator across all child models to be similar to reduce the interference. Experiments on a BERT search space verify that mitigating interference via each of our proposed methods improves the rank correlation of super-pet and combining both methods can achieve better results. Our discovered architecture outperforms RoBERTa${\rm base}$ by 1.1 and 0.6 points *and ELECTRA*${\rm base}$ by 1.6 and 1.1 points on the dev and test set of GLUE benchmark. Extensive results on the BERT compression, reading comprehension and large-scale image classification tasks also demonstrate the effectiveness and generality of our proposed methods.

On the Statistical Benefits of Curriculum Learning

- Ziping Xu, Ambuj Tewari
- abstract: Curriculum learning (CL) is a commonly used machine learning training strategy. However, we still lack a clear theoretical understanding of CL's benefits. In this paper, we study the benefits of CL in the multitask linear regression problem under both structured and unstructured settings. For both settings, we derive the minimax rates for CL with the oracle that provides the optimal curriculum and without the oracle, where the agent has to adaptively learn a good curriculum. Our results reveal that adaptive learning can be fundamentally harder than the oracle learning in the unstructured setting, but it merely introduces a small extra term in the structured setting. To connect theory with practice, we provide justification for a popular empirical method that selects tasks with highest local prediction gain by comparing its guarantees with the minimax rates mentioned above.

A Difference Standardization Method for Mutual Transfer Learning

- Haoqing Xu, Meng Wang, Beilun Wang
- abstract: In many real-world applications, mutual transfer learning is the paradigm that each data domain can potentially be a source or target domain. This is quite different from transfer learning tasks where the source and target are known a priori. However, previous studies about mutual transfer learning either suffer from high computational complexity or oversimplified hypothesis. To overcome these challenges, in this paper, we propose the \underline{Diff}erence \underline{S}tandardization method ({\bf DiffS}) for mutual transfer learning. Specifically, we put forward a novel distance metric between domains, the standardized domain difference, to obtain fast structure recovery and accurate parameter estimation simultaneously. We validate the method's performance using both synthetic and real-world data. Compared to previous methods, DiffS demonstrates a speed-up of approximately 3000 times that of similar methods and achieves the same accurate learnability structure estimation.

SkexGen: Autoregressive Generation of CAD Construction Sequences with Disentangled Codebooks

- Xiang Xu, Karl D.D. Willis, Joseph G Lambourne, Chin-Yi Cheng, Pradeep Kumar Jayaraman, Yasutaka Furukawa
- abstract: We present SkexGen, a novel autoregressive generative model for computer-aided design (CAD) construction sequences containing sketch-and-extrude modeling operations. Our model utilizes distinct Transformer architectures to encode topological, geometric, and extrusion variations of construction sequences into disentangled codebooks. Autoregressive Transformer decoders generate CAD construction sequences sharing certain properties specified by the codebook vectors. Extensive experiments demonstrate that our disentangled codebook representation generates diverse and high-quality CAD models, enhances user control, and enables efficient exploration of the design space. The code is available at https://samxuxiang.github.io/skexgen.

Discriminator-Weighted Offline Imitation Learning from Suboptimal Demonstrations

- Haoran Xu, Xianyuan Zhan, Honglei Yin, Huiling Qin
- abstract: We study the problem of offline Imitation Learning (IL) where an agent aims to learn an optimal expert behavior policy without additional online environment interactions. Instead, the agent is provided with a supplementary offline dataset from suboptimal behaviors. Prior works that address this problem either require that expert data occupies the majority proportion of the offline dataset, or need to learn a reward function and perform offline reinforcement learning (RL) afterwards. In this paper, we aim to address the problem without additional steps of reward learning and offline RL training for the case when demonstrations contain a large proportion of suboptimal data. Built upon behavioral cloning (BC), we introduce an additional discriminator to distinguish expert and non-expert data. We propose a cooperation framework to boost the learning of both tasks, Based on this framework, we design a new IL algorithm, where the outputs of the discriminator serve as the weights of the BC loss. Experimental results show that our proposed algorithm achieves higher returns and faster training speed compared to baseline algorithms.

Adversarial Attack and Defense for Non-Parametric Two-Sample Tests

- Xilie Xu, Jingfeng Zhang, Feng Liu, Masashi Sugiyama, Mohan Kankanhalli
- abstract: Non-parametric two-sample tests (TSTs) that judge whether two sets of samples are drawn from the same distribution, have been widely used in the analysis of critical data. People tend to employ TSTs as trusted basic tools and rarely have any doubt about their reliability. This paper systematically uncovers the failure mode of non-parametric TSTs through adversarial attacks and then proposes corresponding defense strategies. First, we theoretically show that an adversary can upper-bound the distributional shift which guarantees the attack's invisibility. Furthermore, we theoretically find that the

adversary can also degrade the lower bound of a TST's test power, which enables us to iteratively minimize the test criterion in order to search for adversarial pairs. To enable TST-agnostic attacks, we propose an ensemble attack (EA) framework that jointly minimizes the different types of test criteria. Second, to robustify TSTs, we propose a max-min optimization that iteratively generates adversarial pairs to train the deep kernels. Extensive experiments on both simulated and real-world datasets validate the adversarial vulnerabilities of non-parametric TSTs and the effectiveness of our proposed defense. Source code is available at https://github.com/GodXuxilie/Robust-TST.git.

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

- Xiaojun Xu, Jacky Y Zhang, Evelyn Ma, Hyun Ho Son, Sanmi Koyejo, Bo Li
- abstract: Machine learning (ML) robustness and domain generalization are fundamentally correlated: they essentially concern data distribution shifts under adversarial and natural settings, respectively. On one hand, recent studies show that more robust (adversarially trained) models are more generalizable. On the other hand, there is a lack of theoretical understanding of their fundamental connections. In this paper, we explore the relationship between regularization and domain transferability considering different factors such as norm regularization and data augmentations (DA). We propose a general theoretical framework proving that factors involving the model function class regularization are sufficient conditions for relative domain transferability. Our analysis implies that "robustness" is neither necessary nor sufficient for transferability; rather, regularization is a more fundamental perspective for understanding domain transferability. We then discuss popular DA protocols (including adversarial training) and show when they can be viewed as the function class regularization under certain conditions and therefore improve generalization. We conduct extensive experiments to verify our theoretical findings and show several counterexamples where robustness and generalization are negatively correlated on different datasets.

A Theoretical Analysis on Independence-driven Importance Weighting for Covariate-shift Generalization

- Renzhe Xu, Xingxuan Zhang, Zheyan Shen, Tong Zhang, Peng Cui
- abstract: Covariate-shift generalization, a typical case in out-of-distribution (OOD) generalization, requires a good performance on the unknown test distribution, which varies from the accessible training distribution in the form of covariate shift. Recently, independence-driven importance weighting algorithms in stable learning literature have shown empirical effectiveness to deal with covariate-shift generalization on several learning models, including regression algorithms and deep neural networks, while their theoretical analyses are missing. In this paper, we theoretically prove the effectiveness of such algorithms by explaining them as feature selection processes. We first specify a set of variables, named minimal stable variable set, that is the minimal and optimal set of variables to deal with covariate-shift generalization for common loss functions, such as the mean squared loss and binary cross-entropy loss. Afterward, we prove that under ideal conditions, independence-driven importance weighting algorithms could identify the variables in this set. Analysis of asymptotic properties is also provided. These theories are further validated in several synthetic experiments.

Langevin Monte Carlo for Contextual Bandits

- Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, Animashree Anandkumar
- abstract: We study the efficiency of Thompson sampling for contextual bandits. Existing Thompson sampling-based algorithms need to construct a Laplace approximation (i.e., a Gaussian distribution) of the posterior distribution, which is inefficient to sample in high dimensional applications for general covariance matrices. Moreover, the Gaussian approximation may not be a good surrogate for the posterior distribution for general reward generating functions. We propose an efficient posterior sampling algorithm, viz., Langevin Monte Carlo Thompson Sampling (LMC-TS), that uses Markov Chain Monte Carlo (MCMC) methods to directly sample from the posterior distribution in contextual bandits. Our method is computationally efficient since it only needs to perform noisy gradient descent updates without constructing the Laplace approximation of the posterior distribution. We prove that the proposed algorithm achieves the same sublinear regret bound as the best Thompson sampling algorithms for a special case of contextual bandits, viz., linear contextual bandits. We conduct experiments on both synthetic data and real-world datasets on different contextual bandit models, which demonstrates that directly sampling from the posterior is both computationally efficient and competitive in performance.

Investigating Why Contrastive Learning Benefits Robustness against Label Noise

- Yihao Xue, Kyle Whitecross, Baharan Mirzasoleiman
- abstract: Self-supervised Contrastive Learning (CL) has been recently shown to be very effective in preventing deep networks from overfitting noisy labels. Despite its empirical success, the theoretical understanding of the effect of contrastive learning on boosting robustness is very limited. In this work, we rigorously prove that the representation matrix learned by contrastive learning boosts robustness, by having: (i) one prominent singular value corresponding to each sub-class in the data, and significantly smaller remaining singular values; and (ii) a large alignment between the prominent singular vectors and the clean labels of each sub-class. The above properties enable a linear layer trained on such representations to effectively learn the clean labels without overfitting the noise. We further show that the low-rank structure of the Jacobian of deep networks pre-trained with contrastive learning allows them to achieve a superior performance initially, when fine-tuned on noisy labels. Finally, we demonstrate that the initial robustness provided by contrastive learning enables robust training methods to achieve state-of-the-art performance under extreme noise levels, e.g., an average of 27.18% and 15.58% increase in accuracy on CIFAR-10 and CIFAR-100 with 80% symmetric noisy labels, and 4.11% increase in accuracy on WebVision.

Diversified Adversarial Attacks based on Conjugate Gradient Method

- Keiichiro Yamamura, Haruki Sato, Nariaki Tateiwa, Nozomi Hata, Toru Mitsutake, Issa Oe, Hiroki Ishikura, Katsuki Fujisawa
- abstract: Deep learning models are vulnerable to adversarial examples, and adversarial attacks used to generate such examples have attracted considerable research interest. Although existing methods based on the steepest descent have achieved high attack success rates, ill-conditioned problems occasionally reduce their performance. To address this limitation, we utilize the conjugate gradient (CG) method, which is effective for this type of problem, and propose a novel attack algorithm inspired by the CG method, named the Auto Conjugate Gradient (ACG) attack. The results of large-scale evaluation experiments conducted on the latest robust models show that, for most models, ACG was able to find more adversarial examples with fewer iterations than the existing SOTA algorithm Auto-PGD (APGD). We investigated the difference in search performance between ACG and APGD in terms of diversification and intensification, and define a measure called Diversity Index (DI) to quantify the degree of diversity. From the analysis of the diversity using this index, we show that the more diverse search of the proposed method remarkably improves its attack success rate.

Cycle Representation Learning for Inductive Relation Prediction

- Zuoyu Yan, Tengfei Ma, Liangcai Gao, Zhi Tang, Chao Chen
- abstract: In recent years, algebraic topology and its modern development, the theory of persistent homology, has shown great potential in graph representation learning. In this paper, based on the mathematics of algebraic topology, we propose a novel solution for inductive relation prediction, an important learning task for knowledge graph completion. To predict the relation between two entities, one can use the existence of rules, namely a sequence of relations. Previous works view rules as paths and primarily focus on the searching of paths between entities. The space of rules is huge, and one has to sacrifice either efficiency or accuracy. In this paper, we consider rules as cycles and show that the space of cycles has a unique structure based on the mathematics of algebraic topology. By exploring the linear structure of the cycle space, we can improve the searching efficiency of rules. We propose to collect cycle bases that span the space of cycles. We build a novel GNN framework on the collected cycles to learn the representations of cycles, and to predict the existence/non-existence of a relation. Our method achieves state-of-the-art performance on benchmarks.

Optimally Controllable Perceptual Lossy Compression

- Zeyu Yan, Fei Wen, Peilin Liu
- abstract: Recent studies in lossy compression show that distortion and perceptual quality are at odds with each other, which put forward the tradeoff between distortion and perception (D-P). Intuitively, to attain different perceptual quality, different decoders have to be trained. In this paper, we present a nontrivial finding that only two decoders are sufficient for optimally achieving arbitrary (an infinite number of different) D-P tradeoff. We prove that arbitrary points of the D-P tradeoff bound can be achieved by a simple linear interpolation between the outputs of a minimum MSE decoder and a specifically constructed perfect perceptual decoder. Meanwhile, the perceptual quality (in terms of the squared Wasserstein-2 distance metric) can be quantitatively controlled by the interpolation factor. Furthermore, to construct a perfect perceptual decoder, we propose two theoretically optimal training frameworks. The new frameworks are different from the distortion-plus-adversarial loss based heuristic framework widely used in existing methods, which are not only theoretically optimal but also can yield state-of-the-art performance in practical perceptual decoding. Finally, we validate our theoretical finding and demonstrate the superiority of our frameworks via experiments. Code is available at: https://github.com/ZeyuYan/ControllablePerceptual-Compression

Active fairness auditing

- Tom Yan, Chicheng Zhang
- abstract: The fast spreading adoption of machine learning (ML) by companies across industries poses significant regulatory challenges. One such challenge is scalability: how can regulatory bodies efficiently audit these ML models, ensuring that they are fair? In this paper, we initiate the study of query-based auditing algorithms that can estimate the demographic parity of ML models in a query-efficient manner. We propose an optimal deterministic algorithm, as well as a practical randomized, oracle-efficient algorithm with comparable guarantees. Furthermore, we make inroads into understanding the optimal query complexity of randomized active fairness estimation algorithms. Our first exploration of active fairness estimation aims to put AI governance on firmer theoretical foundations.

Self-Organized Polynomial-Time Coordination Graphs

- Qianlan Yang, Weijun Dong, Zhizhou Ren, Jianhao Wang, Tonghan Wang, Chongjie Zhang
- abstract: Coordination graph is a promising approach to model agent collaboration in multi-agent reinforcement learning. It conducts a graph-based value factorization and induces explicit coordination among agents to complete complicated tasks. However, one critical challenge in this paradigm is the complexity of greedy action selection with respect to the factorized values. It refers to the decentralized constraint optimization problem (DCOP), which and whose constant-ratio approximation are NP-hard problems. To bypass this systematic hardness, this paper proposes a novel method, named Self-Organized Polynomial-time Coordination Graphs (SOP-CG), which uses structured graph classes to guarantee the accuracy and the computational efficiency of collaborated action selection. SOP-CG employs dynamic graph topology to ensure sufficient value function expressiveness. The graph selection is unified into an end-to-end learning paradigm. In experiments, we show that our approach learns succinct and well-adapted graph topologies, induces effective coordination, and improves performance across a variety of cooperative multi-agent tasks.

Regularizing a Model-based Policy Stationary Distribution to Stabilize Offline Reinforcement Learning

- Shentao Yang, Yihao Feng, Shujian Zhang, Mingyuan Zhou
- abstract: Offline reinforcement learning (RL) extends the paradigm of classical RL algorithms to purely learning from static datasets, without interacting with the underlying environment during the learning process. A key challenge of offline RL is the instability of policy training, caused by the mismatch between the distribution of the offline data and the undiscounted stationary state-action distribution of the learned policy. To avoid the detrimental impact of distribution mismatch, we regularize the undiscounted stationary distribution of the current policy towards the offline data during the policy optimization process. Further, we train a dynamics model to both implement this regularization and better estimate the stationary distribution of the current policy, reducing the error induced by distribution mismatch. On a wide range of continuous-control offline RL datasets, our method indicates competitive performance, which validates our algorithm. The code is publicly available.

A Psychological Theory of Explainability

- Scott Cheng-Hsin Yang, Nils Erik Tomas Folke, Patrick Shafto
- abstract: The goal of explainable Artificial Intelligence (XAI) is to generate human-interpretable explanations, but there are no computationally precise theories of how humans interpret AI generated explanations. The lack of theory means that validation of XAI must be done empirically, on a case-by-case basis, which prevents systematic theory-building in XAI. We propose a psychological theory of how humans draw conclusions from saliency maps, the most common form of XAI explanation, which for the first time allows for precise prediction of explainee inference conditioned on explanation. Our theory posits that absent explanation humans expect the AI to make similar decisions to themselves, and that they interpret an explanation by comparison to the explanations they themselves would give. Comparison is formalized via Shepard's universal law of generalization in a similarity space, a classic theory from cognitive science. A pre-registered user study on AI image classifications with saliency map explanations demonstrate that our theory quantitatively matches participants' predictions of the AI.

Omni-Granular Ego-Semantic Propagation for Self-Supervised Graph Representation Learning

- Ling Yang, Shenda Hong
- abstract: Unsupervised/self-supervised graph representation learning is critical for downstream node- and graph-level classification tasks. Global structure of graphs helps discriminating representations and existing methods mainly utilize the global structure by imposing additional supervisions. However, their global semantics are usually invariant for all nodes/graphs and they fail to explicitly embed the global semantics to enrich the representations. In this paper, we propose Omni-Granular Ego-Semantic Propagation for Self-Supervised Graph Representation Learning (OEPG). Specifically, we introduce instance-adaptive global-aware ego-semantic descriptors, leveraging the first- and second-order feature differences between each node/graph and hierarchical global clusters of the entire graph dataset. The descriptors can be explicitly integrated into local graph convolution as new neighbor nodes. Besides, we design an omni-granular normalization on the whole scales and hierarchies of the ego-semantic to assign attentional weight to each descriptor from an omni-granular perspective. Specialized pretext tasks and cross-iteration momentum update are further developed for local-global mutual adaptation. In downstream tasks, OEPG consistently achieves the best performance with a 2%~6% accuracy gain on multiple datasets cross scales and domains. Notably, OEPG also generalizes to quantity- and topology-imbalance scenarios.

Unsupervised Time-Series Representation Learning with Iterative Bilinear Temporal-Spectral Fusion

- Ling Yang, Shenda Hong
- abstract: Unsupervised/self-supervised time series representation learning is a challenging problem because of its complex dynamics and sparse annotations. Existing works mainly adopt the framework of contrastive learning with the time-based augmentation techniques to sample positives and negatives for contrastive training. Nevertheless, they mostly use segment-level augmentation derived from time slicing, which may bring about sampling bias and incorrect optimization with false negatives due to the loss of global context. Besides, they all pay no attention to incorporate the spectral information in feature representation. In this paper, we propose a unified framework, namely Bilinear Temporal-Spectral Fusion (BTSF). Specifically, we

firstly utilize the instance-level augmentation with a simple dropout on the entire time series for maximally capturing long-term dependencies. We devise a novel iterative bilinear temporal-spectral fusion to explicitly encode the affinities of abundant time-frequency pairs, and iteratively refines representations in a fusion-and-squeeze manner with Spectrum-to-Time (S2T) and Time-to-Spectrum (T2S) Aggregation modules. We firstly conducts downstream evaluations on three major tasks for time series including classification, forecasting and anomaly detection. Experimental results shows that our BTSF consistently significantly outperforms the state-of-the-art methods.

Searching for BurgerFormer with Micro-Meso-Macro Space Design

- Longxing Yang, Yu Hu, Shun Lu, Zihao Sun, Jilin Mei, Yinhe Han, Xiaowei Li
- abstract: With the success of Transformers in the computer vision field, the automated design of vision Transformers has attracted significant attention. Recently, MetaFormer found that simple average pooling can achieve impressive performance, which naturally raises the question of how to design a search space to search diverse and high-performance Transformer-like architectures. By revisiting typical search spaces, we design micro-meso-macro space to search for Transformer-like architectures, namely BurgerFormer. Micro, meso, and macro correspond to the granularity levels of operation, block and stage, respectively. At the microscopic level, we enrich the atomic operations to include various normalizations, activation functions, and basic operations (e.g., multi-head self attention, average pooling). At the mesoscopic level, a hamburger structure is searched out as the basic BurgerFormer block. At the macroscopic level, we search for the depth, width, and expansion ratio of the network based on the multi-stage architecture. Meanwhile, we propose a hybrid sampling method for effectively training the supernet. Experimental results demonstrate that the searched BurgerFormer architectures achieve comparable even superior performance compared with current state-of-the-art Transformers on the ImageNet and COCO datasets. The codes can be available at https://github.com/xingxing-123/BurgerFormer.

Efficient Variance Reduction for Meta-learning

- Hansi Yang, James Kwok
- abstract: Meta-learning tries to learn meta-knowledge from a large number of tasks. However, the stochastic meta-gradient can have large variance due to data sampling (from each task) and task sampling (from the whole task distribution), leading to slow convergence. In this paper, we propose a novel approach that integrates variance reduction with first-order meta-learning algorithms such as Reptile. It retains the bilevel formulation which better captures the structure of meta-learning, but does not require storing the vast number of task-specific parameters in general bilevel variance reduction methods. Theoretical results show that it has fast convergence rate due to variance reduction. Experiments on benchmark few-shot classification data sets demonstrate its effectiveness over state-of-the-art meta-learning algorithms with and without variance reduction.

Injecting Logical Constraints into Neural Networks via Straight-Through Estimators

- Zhun Yang, Joohyung Lee, Chiyoun Park
- abstract: Injecting discrete logical constraints into neural network learning is one of the main challenges in neuro-symbolic AI. We find that a straight-through-estimator, a method introduced to train binary neural networks, could effectively be applied to incorporate logical constraints into neural network learning. More specifically, we design a systematic way to represent discrete logical constraints as a loss function; minimizing this loss using gradient descent via a straight-through-estimator updates the neural network's weights in the direction that the binarized outputs satisfy the logical constraints. The experimental results show that by leveraging GPUs and batch training, this method scales significantly better than existing neuro-symbolic methods that require heavy symbolic computation for computing gradients. Also, we demonstrate that our method applies to different types of neural networks, such as MLP, CNN, and GNN, making them learn with no or fewer labeled data by learning directly from known constraints.

Locally Sparse Neural Networks for Tabular Biomedical Data

- Junchen Yang, Ofir Lindenbaum, Yuval Kluger
- abstract: Tabular datasets with low-sample-size or many variables are prevalent in biomedicine. Practitioners in this domain prefer linear or tree-based models over neural networks since the latter are harder to interpret and tend to overfit when applied to tabular datasets. To address these neural networks' shortcomings, we propose an intrinsically interpretable network for heterogeneous biomedical data. We design a locally sparse neural network where the local sparsity is learned to identify the subset of most relevant features for each sample. This sample-specific sparsity is predicted via a gating network, which is trained in tandem with the prediction network. By forcing the model to select a subset of the most informative features for each sample, we reduce model overfitting in low-sample-size data and obtain an interpretable model. We demonstrate that our method outperforms state-of-the-art models when applied to synthetic or real-world biomedical datasets using extensive experiments. Furthermore, the proposed framework dramatically outperforms existing schemes when evaluating its interpretability capabilities. Finally, we demonstrate the applicability of our model to two important biomedical tasks: survival analysis and marker gene identification.

Not All Poisons are Created Equal: Robust Training against Data Poisoning

- Yu Yang, Tian Yu Liu, Baharan Mirzasoleiman
- abstract: Data poisoning causes misclassification of test time target examples, by injecting maliciously crafted samples in the training data. Existing defenses are often effective only against a specific type of targeted attack, significantly degrade the generalization performance, or are prohibitive for standard deep learning pipelines. In this work, we propose an efficient defense mechanism that significantly reduces the success rate of various data poisoning attacks, and provides theoretical guarantees for the performance of the model. Targeted attacks work by adding bounded perturbations to a randomly selected subset of training data to match the targets' gradient or representation. We show that: (i) under bounded perturbations, only a number of poisons can be optimized to have a gradient that is close enough to that of the target and make the attack successful; (ii) such effective poisons move away from their original class and get isolated in the gradient space; (iii) dropping examples in low-density gradient regions during training can successfully eliminate the effective poisons, and guarantees similar training dynamics to that of training on full data. Our extensive experiments show that our method significantly decreases the success rate of state-of-the-art targeted attacks, including Gradient Matching and Bullseye Polytope, and easily scales to large datasets.

Does the Data Induce Capacity Control in Deep Learning?

- Rubing Yang, Jialin Mao, Pratik Chaudhari
- abstract: We show that the input correlation matrix of typical classification datasets has an eigenspectrum where, after a sharp initial drop, a large number of small eigenvalues are distributed uniformly over an exponentially large range. This structure is mirrored in a network trained on this data: we show that the Hessian and the Fisher Information Matrix (FIM) have eigenvalues that are spread uniformly over exponentially large ranges. We call such eigenspectra "sloppy" because sets of weights corresponding to small eigenvalues can be changed by large magnitudes without affecting the loss. Networks trained on atypical datasets with non-sloppy inputs do not share these traits and deep networks trained on such datasets generalize poorly. Inspired by this, we study the hypothesis that sloppiness of inputs aids generalization in deep networks. We show that if the Hessian is sloppy, we can compute non-vacuous PAC-Bayes generalization bounds analytically. By exploiting our empirical observation that training predominantly takes place in the non-sloppy subspace of the FIM, we develop data-distribution dependent PAC-Bayes priors that lead to accurate generalization bounds using numerical optimization.

Informed Learning by Wide Neural Networks: Convergence, Generalization and Sampling Complexity

- Jianyi Yang, Shaolei Ren
- abstract: By integrating domain knowledge with labeled samples, informed machine learning has been emerging to improve the learning performance for a wide range of applications. Nonetheless, rigorous understanding of the role of injected domain knowledge has been under-explored. In this paper, we consider an informed deep neural network (DNN) with over-parameterization and domain knowledge integrated into its training objective function, and study how and why domain knowledge benefits the performance. Concretely, we quantitatively demonstrate the two benefits of domain knowledge in informed learning {—} regularizing the label-based supervision and supplementing the labeled samples {—} and reveal the trade-off between label and knowledge imperfectness in the bound of the population risk. Based on the theoretical analysis, we propose a generalized informed training objective to better exploit the benefits of knowledge and balance the label and knowledge imperfectness, which is validated by the population risk bound. Our analysis on sampling complexity sheds lights on how to choose the hyper-parameters for informed learning, and further justifies the advantages of knowledge informed learning.

Linear Bandit Algorithms with Sublinear Time Complexity

- Shuo Yang, Tongzheng Ren, Sanjay Shakkottai, Eric Price, Inderjit S. Dhillon, Sujay Sanghavi
- abstract: We propose two linear bandits algorithms with per-step complexity sublinear in the number of arms K. The algorithms are designed for applications where the arm set is extremely large and slowly changing. Our key realization is that choosing an arm reduces to a maximum inner product search (MIPS) problem, which can be solved approximately without breaking regret guarantees. Existing approximate MIPS solvers run in sublinear time. We extend those solvers and present theoretical guarantees for online learning problems, where adaptivity (i.e., a later step depends on the feedback in previous steps) becomes a unique challenge. We then explicitly characterize the tradeoff between the per-step complexity and regret. For sufficiently large K, our algorithms have sublinear per-step complexity and $\widetilde O(\sqrt{T})$ regret. Empirically, we evaluate our proposed algorithms in a synthetic environment and a real-world online movie recommendation problem. Our proposed algorithms can deliver a more than 72 times speedup compared to the linear time baselines while retaining similar regret.

A New Perspective on the Effects of Spectrum in Graph Neural Networks

- Mingqi Yang, Yanming Shen, Rui Li, Heng Qi, Qiang Zhang, Baocai Yin
- abstract: Many improvements on GNNs can be deemed as operations on the spectrum of the underlying graph matrix, which motivates us to directly study the characteristics of the spectrum and their effects on GNN performance. By generalizing most existing GNN architectures, we show that the correlation issue caused by the unsmooth spectrum becomes the obstacle to leveraging more powerful graph filters as well as developing deep architectures, which therefore restricts GNNs' performance. Inspired by this, we propose the correlation-free architecture which naturally removes the correlation issue among different channels, making it possible to utilize more sophisticated filters within each channel. The final correlation-free architecture with more powerful filters consistently boosts the performance of learning graph representations. Code is available at https://github.com/qslim/gnn-spectrum.

Fourier Learning with Cyclical Data

- Yingxiang Yang, Zhihan Xiong, Tianyi Liu, Taiqing Wang, Chong Wang
- abstract: Many machine learning models for online applications, such as recommender systems, are often trained on data with cyclical properties. These data sequentially arrive from a time-varying distribution that is periodic in time. Existing algorithms either use streaming learning to track a time-varying set of optimal model parameters, yielding a dynamic regret that scales linearly in time; or partition the data of each cycle into multiple segments and train a separate model for each—a pluralistic approach that is computationally and storage-wise expensive. In this paper, we have designed a novel approach to overcome the aforementioned shortcomings. Our method, named "Fourier learning", encodes the periodicity into the model representation using a partial Fourier sequence, and trains the coefficient functions modeled by neural networks. Particularly, we design a Fourier multi-layer perceptron (F-MLP) that can be trained on streaming data with stochastic gradient descent (streaming-SGD), and we derive its convergence guarantees. We demonstrate Fourier learning's better performance with extensive experiments on synthetic and public datasets, as well as on a large-scale recommender system that is updated in real-time, and trained with tens of millions of samples per day.

Estimating Instance-dependent Bayes-label Transition Matrix using a Deep Neural Network

- Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, Tongliang Liu
- abstract: In label-noise learning, estimating the transition matrix is a hot topic as the matrix plays an important role in building statistically consistent classifiers. Traditionally, the transition from clean labels to noisy labels (i.e., clean-label transition matrix (CLTM)) has been widely exploited to learn a clean label classifier by employing the noisy data. Motivated by that classifiers mostly output Bayes optimal labels for prediction, in this paper, we study to directly model the transition from Bayes optimal labels to noisy labels (i.e., Bayes-label transition matrix (BLTM)) and learn a classifier to predict Bayes optimal labels. Note that given only noisy data, it is ill-posed to estimate either the CLTM or the BLTM. But favorably, Bayes optimal labels have less uncertainty compared with the clean labels, i.e., the class posteriors of Bayes optimal labels are one-hot vectors while those of clean labels are not. This enables two advantages to estimate the BLTM, i.e., (a) a set of examples with theoretically guaranteed Bayes optimal labels can be collected out of noisy data; (b) the feasible solution space is much smaller. By exploiting the advantages, we estimate the BLTM parametrically by employing a deep neural network, leading to better generalization and superior classification performance.

A Study of Face Obfuscation in ImageNet

- Kaiyu Yang, Jacqueline H. Yau, Li Fei-Fei, Jia Deng, Olga Russakovsky
- abstract: Face obfuscation (blurring, mosaicing, etc.) has been shown to be effective for privacy protection; nevertheless, object recognition research typically assumes access to complete, unobfuscated images. In this paper, we explore the effects of face obfuscation on the popular ImageNet challenge visual recognition benchmark. Most categories in the ImageNet challenge are not people categories; however, many incidental people appear in the images, and their privacy is a concern. We first annotate faces in the dataset. Then we demonstrate that face obfuscation has minimal impact on the accuracy of recognition models. Concretely, we benchmark multiple deep neural networks on obfuscated images and observe that the overall recognition accuracy drops only slightly (<= 1.0%). Further, we experiment with transfer learning to 4 downstream tasks (object recognition, scene recognition, face attribute classification, and object detection) and show that features learned on obfuscated images are equally transferable. Our work demonstrates the feasibility of privacy-aware visual recognition, improves the highly-used ImageNet challenge benchmark, and suggests an important path for future visual datasets. Data and code are available at https://github.com/princetonvisualai/imagenet-face-obfuscation.

Anarchic Federated Learning

- Haibo Yang, Xin Zhang, Prashant Khanduri, Jia Liu
- abstract: Present-day federated learning (FL) systems deployed over edge networks consists of a large number of workers with high degrees of heterogeneity in data and/or computing capabilities, which call for flexible worker participation in terms of timing, effort, data heterogeneity, etc. To satisfy the need for flexible worker participation, we consider a new FL paradigm called "Anarchic Federated Learning" (AFL) in this paper. In stark contrast to conventional FL models, each worker in AFL has the freedom to choose i) when to participate in FL, and ii) the number of local steps to

perform in each round based on its current situation (e.g., battery level, communication channels, privacy concerns). However, such chaotic worker behaviors in AFL impose many new open questions in algorithm design. In particular, it remains unclear whether one could develop convergent AFL training algorithms, and if yes, under what conditions and how fast the achievable convergence speed is. Toward this end, we propose two Anarchic Federated Averaging (AFA) algorithms with two-sided learning rates for both cross-device and cross-silo settings, which are named AFA-CD and AFA-CS, respectively. Somewhat surprisingly, we show that, under mild anarchic assumptions, both AFL algorithms achieve the best known convergence rate as the state-of-the-art algorithms for conventional FL. Moreover, they retain the highly desirable linear speedup effect with respect of both the number of workers and local steps in the new AFL paradigm. We validate the proposed algorithms with extensive experiments on real-world datasets.

Identity-Disentangled Adversarial Augmentation for Self-supervised Learning

- Kaiwen Yang, Tianyi Zhou, Xinmei Tian, Dacheng Tao
- abstract: Data augmentation is critical to contrastive self-supervised learning, whose goal is to distinguish a sample's augmentations (positives) from other samples (negatives). However, strong augmentations may change the sample-identity of the positives, while weak augmentation produces easy positives/negatives leading to nearly-zero loss and ineffective learning. In this paper, we study a simple adversarial augmentation method that can modify training data to be hard positives/negatives without distorting the key information about their original identities. In particular, we decompose a sample x to be its variational auto-encoder (VAE) reconstruction $G(x)$ plus the residual $R(x)=x-G(x)$, where $R(x)$ retains most identity-distinctive information due to an information-theoretic interpretation of the VAE objective. We then adversarially perturb $G(x)$ in the VAE's bottleneck space and adds it back to the original $R(x)$ as an augmentation, which is therefore sufficiently challenging for contrastive learning and meanwhile preserves the sample identity intact. We apply this "identity-disentangled adversarial augmentation (IDAA)" to different self-supervised learning methods. On multiple benchmark datasets, IDAA consistently improves both their efficiency and generalization performance. We further show that IDAA learned on a dataset can be transferred to other datasets. Code is available at \href{https://github.com/kai-wen-yang/IDAA}{https://github.com/kai-wen-yang/IDAA}.

Learning from a Learning User for Optimal Recommendations

- Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, Haifeng Xu
- abstract: In real-world recommendation problems, especially those with a formidably large item space, users have to gradually learn to estimate the utility of any fresh recommendations from their experience about previously consumed items. This in turn affects their interaction dynamics with the system and can invalidate previous algorithms built on the omniscient user assumption. In this paper, we formalize a model to capture such "learning users" and design an efficient system-side learning solution, coined Noise-Robust Active Ellipsoid Search (RAES), to confront the challenges brought by the non-stationary feedback from such a learning user. Interestingly, we prove that the regret of RAES deteriorates gracefully as the convergence rate of user learning becomes worse, until reaching linear regret when the user's learning fails to converge. Experiments on synthetic datasets demonstrate the strength of RAES for such a contemporaneous system-user learning problem. Our study provides a novel perspective on modeling the feedback loop in recommendation problems.

Improving Out-of-Distribution Robustness via Selective Augmentation

- Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, Chelsea Finn
- abstract: Machine learning algorithms typically assume that training and test examples are drawn from the same distribution. However, distribution shift is a common problem in real-world applications and can cause models to perform dramatically worse at test time. In this paper, we specifically consider the problems of subpopulation shifts (e.g., imbalanced data) and domain shifts. While prior works often seek to explicitly regularize internal representations or predictors of the model to be domain invariant, we instead aim to learn invariant predictors without restricting the model's internal representations or predictors. This leads to a simple mixup-based technique which learns invariant predictors via selective augmentation called LISA. LISA selectively interpolates samples either with the same labels but different domains or with the same domain but different labels. Empirically, we study the effectiveness of LISA on nine benchmarks ranging from subpopulation shifts to domain shifts, and we find that LISA consistently outperforms other state-of-the-art methods and leads to more invariant predictors. We further analyze a linear setting and theoretically show how LISA leads to a smaller worst-group error.

NLP From Scratch Without Large-Scale Pretraining: A Simple and Efficient Framework

- Xingcheng Yao, Yanan Zheng, Xiaocong Yang, Zhilin Yang
- abstract: Pretrained language models have become the standard approach for many NLP tasks due to strong performance, but they are very expensive to train. We propose a simple and efficient learning framework, TLM, that does not rely on large-scale pretraining. Given some labeled task data and a large general corpus, TLM uses task data as queries to retrieve a tiny subset of the general corpus and jointly optimizes the task objective and the language modeling objective from scratch. On eight classification datasets in four domains, TLM achieves results better than or similar to pretrained language models (e.g., RoBERTa-Large) while reducing the training FLOPs by two orders of magnitude. With high accuracy and efficiency, we hope TLM will contribute to democratizing NLP and expediting its development.

Feature Space Particle Inference for Neural Network Ensembles

- Shingo Yashima, Teppei Suzuki, Kohta Ishikawa, Ikuro Sato, Rei Kawakami
- abstract: Ensembles of deep neural networks demonstrate improved performance over single models. For enhancing the diversity of ensemble members while keeping their performance, particle-based inference methods offer a promising approach from a Bayesian perspective. However, the best way to apply these methods to neural networks is still unclear: seeking samples from the weight-space posterior suffers from inefficiency due to the over-parameterization issues, while seeking samples directly from the function-space posterior often leads to serious underfitting. In this study, we propose to optimize particles in the feature space where activations of a specific intermediate layer lie to alleviate the abovementioned difficulties. Our method encourages each member to capture distinct features, which are expected to increase the robustness of the ensemble prediction. Extensive evaluation on real-world datasets exhibits that our model significantly outperforms the gold-standard Deep Ensembles on various metrics, including accuracy, calibration, and robustness.

Centroid Approximation for Bootstrap: Improving Particle Quality at Inference

- Mao Ye, Qiang Liu
- abstract: Bootstrap is a principled and powerful frequentist statistical tool for uncertainty quantification. Unfortunately, standard bootstrap methods are computationally intensive due to the need of drawing a large i.i.d. bootstrap sample to approximate the ideal bootstrap distribution; this largely hinders their application in large-scale machine learning, especially deep learning problems. In this work, we propose an efficient method to explicitly optimize a small set of high quality "centroid" points to better approximate the ideal bootstrap distribution. We achieve this by minimizing a simple objective function that is asymptotically equivalent to the Wasserstein distance to the ideal bootstrap distribution. This allows us to provide an accurate estimation of uncertainty with a small number of bootstrap centroids, outperforming the naive i.i.d. sampling approach. Empirically, we show that our method can boost the performance of bootstrap in a variety of applications.

Be Like Water: Adaptive Floating Point for Machine Learning

- Thomas Yeh, Max Sterner, Zerlina Lai, Brandon Chuang, Alexander Ihler
- abstract: In the pursuit of optimizing memory and compute density to accelerate machine learning applications, reduced precision training and inference has been an active area of research. While some approaches selectively apply low precision computations, this may require costly off-chip data transfers or mixed precision support. In this paper, we propose a novel numerical representation, Adaptive Floating Point (AFP), that dynamically adjusts to the characteristics of deep learning data. AFP requires no changes to the model topology, requires no additional training, and applies to all layers of DNN models. We evaluate AFP on a spectrum of representative models in computer vision and NLP, and show that our technique enables ultra-low precision inference of deep learning models while providing accuracy comparable to full precision inference. By dynamically adjusting to ML data, AFP increases memory density by 1.6x, 1.6x, and 3.2x and compute density by 4x, 1.3x, and 12x when compared to BFP, BFloat16, and FP32.

QSFL: A Two-Level Uplink Communication Optimization Framework for Federated Learning

- Liping Yi, Wang Gang, Liu Xiaoguang
- abstract: In cross-device Federated Learning (FL), the communication cost of transmitting full-precision models between edge devices and a central server is a significant bottleneck, due to expensive, unreliable, and low-bandwidth wireless connections. As a solution, we propose a novel FL framework named QSFL, towards optimizing FL uplink (client-to-server) communication at both client and model levels. At the client level, we design a Qualification Judgment (QJ) algorithm to sample high-qualification clients to upload models. At the model level, we explore a Sparse Cyclic Sliding Segment (SCSS) algorithm to further compress transmitted models. We prove that QSFL can converge over wall-to-wall time, and develop an optimal hyperparameter searching algorithm based on theoretical analysis to enable QSFL to make the best trade-off between model accuracy and communication cost. Experimental results show that QSFL achieves state-of-the-art compression ratios with marginal model accuracy degradation.

De novo mass spectrometry peptide sequencing with a transformer model

- Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, William S Noble
- abstract: Tandem mass spectrometry is the only high-throughput method for analyzing the protein content of complex biological samples and is thus the primary technology driving the growth of the field of proteomics. A key outstanding challenge in this field involves identifying the sequence of amino acids -the peptide- responsible for generating each observed spectrum, without making use of prior knowledge in the form of a peptide sequence database. Although various machine learning methods have been developed to address this de novo sequencing problem, challenges that arise when modeling tandem mass spectra have led to complex models that combine multiple neural networks and post-processing steps. We propose a simple yet powerful method for de novo peptide sequencing, Casanovo, that uses a transformer framework to map directly from a sequence of observed peaks (a mass spectrum) to a sequence of amino acids (a peptide). Our experiments show that Casanovo achieves state-of-the-art performance on a benchmark dataset using a standard cross-species evaluation framework which involves testing with spectra with never-before-seen peptide labels. Casanovo not only achieves superior performance but does so at a fraction of the model complexity and inference time required by other methods.

Bayesian Nonparametric Learning for Point Processes with Spatial Homogeneity: A Spatial Analysis of NBA Shot Locations

- Fan Yin, Jieying Jiao, Jun Yan, Guanyu Hu
- abstract: Basketball shot location data provide valuable summary information regarding players to coaches, sports analysts, fans, statisticians, as well as players themselves. Represented by spatial points, such data are naturally analyzed with spatial point process models. We present a novel nonparametric Bayesian method for learning the underlying intensity surface built upon a combination of Dirichlet process and Markov random field. Our method has the advantage of effectively encouraging local spatial homogeneity when estimating a globally heterogeneous intensity surface. Posterior inferences are performed with an efficient Markov chain Monte Carlo (MCMC) algorithm. Simulation studies show that the inferences are accurate and the method is superior compared to a wide range of competing methods. Application to the shot location data of 20 representative NBA players in the 2017-2018 regular season offers interesting insights about the shooting patterns of these players. A comparison against the competing method shows that the proposed method can effectively incorporate spatial contiguity into the estimation of intensity surfaces.

Bitwidth Heterogeneous Federated Learning with Progressive Weight Dequantization

- Jaehong Yoon, Geon Park, Wonyong Jeong, Sung Ju Hwang
- abstract: In practical federated learning scenarios, the participating devices may have different bitwidths for computation and memory storage by design. However, despite the progress made in device-heterogeneous federated learning scenarios, the heterogeneity in the bitwidth specifications in the hardware has been mostly overlooked. We introduce a pragmatic FL scenario with bitwidth heterogeneity across the participating devices, dubbed as Bitwidth Heterogeneous Federated Learning (BHFL). BHFL brings in a new challenge, that the aggregation of model parameters with different bitwidths could result in severe performance degeneration, especially for high-bitwidth models. To tackle this problem, we propose ProWD framework, which has a trainable weight dequantizer at the central server that progressively reconstructs the low-bitwidth weights into higher bitwidth weights, and finally into full-precision weights. ProWD further selectively aggregates the model parameters to maximize the compatibility across bit-heterogeneous weights. We validate ProWD against relevant FL baselines on the benchmark datasets, using clients with varying bitwidths. Our ProWD largely outperforms the baseline FL algorithms as well as naive approaches (e.g. grouped averaging) under the proposed BHFL scenario.

ShiftAddNAS: Hardware-Inspired Search for More Accurate and Efficient Neural Networks

- Haoran You, Baopu Li, Shi Huihong, Yonggan Fu, Yingyan Lin
- abstract: Neural networks (NNs) with intensive multiplications (e.g., convolutions and transformers) are powerful yet power hungry, impeding their more extensive deployment into resource-constrained edge devices. As such, multiplication-free networks, which follow a common practice in energy-efficient hardware implementation to parameterize NNs with more efficient operators (e.g., bitwise shifts and additions), have gained growing attention. However, multiplication-free networks in general under-perform their vanilla counterparts in terms of the achieved accuracy. To this end, this work advocates hybrid NNs that consist of both powerful yet costly multiplications and efficient yet less powerful operators for marrying the best of both worlds, and proposes ShiftAddNAS, which can automatically search for more accurate and more efficient NNs. Our ShiftAddNAS highlights two enablers. Specifically, it integrates (1) the first hybrid search space that incorporates both multiplication-based and multiplication-free operators for facilitating the development of both accurate and efficient hybrid NNs; and (2) a novel weight sharing strategy that enables effective weight sharing among different operators that follow heterogeneous distributions (e.g., Gaussian for convolutions vs. Laplacian for add operators) and simultaneously leads to a largely reduced supernet size and much better searched networks. Extensive experiments and ablation studies on various models, datasets, and tasks consistently validate the effectiveness of ShiftAddNAS, e.g., achieving up to a +7.7% higher accuracy or a +4.9 better BLEU score as compared to state-of-the-art expert-designed and neural architecture searched NNs, while leading to up to 93% or 69% energy and latency savings, respectively. Codes and pretrained models are available at https://github.com/RICE-EIC/ShiftAddNAS.

Molecular Representation Learning via Heterogeneous Motif Graph Neural Networks

- Zhaoning Yu, Hongyang Gao
- abstract: We consider feature representation learning problem of molecular graphs. Graph Neural Networks have been widely used in feature representation learning of molecular graphs. However, most existing methods deal with molecular graphs individually while neglecting their connections, such as motif-level relationships. We propose a novel molecular graph representation learning method by constructing a heterogeneous motif graph to address this issue. In particular, we build a heterogeneous motif graph that contains motif nodes and molecular nodes. Each motif node corresponds to a

motif extracted from molecules. Then, we propose a Heterogeneous Motif Graph Neural Network (HM-GNN) to learn feature representations for each node in the heterogeneous motif graph. Our heterogeneous motif graph also enables effective multi-task learning, especially for small molecular datasets. To address the potential efficiency issue, we propose to use an edge sampler, which can significantly reduce computational resources usage. The experimental results show that our model consistently outperforms previous state-of-the-art models. Under multi-task settings, the promising performances of our methods on combined datasets shed light on a new learning paradigm for small molecular datasets. Finally, we show that our model achieves similar performances with significantly less computational resources by using our edge sampler.

Understanding Robust Overfitting of Adversarial Training and Beyond

- Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, Tongliang Liu
- abstract: Robust overfitting widely exists in adversarial training of deep networks. The exact underlying reasons for this are still not completely understood. Here, we explore the causes of robust overfitting by comparing the data distribution of non-overfit (weak adversary) and overfitted (strong adversary) adversarial training, and observe that the distribution of the adversarial data generated by weak adversary mainly contain small-loss data. However, the adversarial data generated by strong adversary is more diversely distributed on the large-loss data and the small-loss data. Given these observations, we further designed data ablation adversarial training and identify that some small-loss data which are not worthy of the adversary strength cause robust overfitting in the strong adversary mode. To relieve this issue, we propose minimum loss constrained adversarial training (MLCAT): in a minibatch, we learn large-loss data as usual, and adopt additional measures to increase the loss of the small-loss data. Technically, MLCAT hinders data fitting when they become easy to learn to prevent robust overfitting; philosophically, MLCAT reflects the spirit of turning waste into treasure and making the best use of each adversarial data; algorithmically, we designed two realizations of MLCAT, and extensive experiments demonstrate that MLCAT can eliminate robust overfitting and further boost adversarial robustness.

How to Leverage Unlabeled Data in Offline Reinforcement Learning

- Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Chelsea Finn, Sergey Levine
- abstract: Offline reinforcement learning (RL) can learn control policies from static datasets but, like standard RL methods, it requires reward annotations for every transition. In many cases, labeling large datasets with rewards may be costly, especially if those rewards must be provided by human labelers, while collecting diverse unlabeled data might be comparatively inexpensive. How can we best leverage such unlabeled data in offline RL? One natural solution is to learn a reward function from the labeled data and use it to label the unlabeled data. In this paper, we find that, perhaps surprisingly, a much simpler method that simply applies zero rewards to unlabeled data leads to effective data sharing both in theory and in practice, without learning any reward model at all. While this approach might seem strange (and incorrect) at first, we provide extensive theoretical and empirical analysis that illustrates how it trades off reward bias, sample complexity and distributional shift, often leading to good results. We characterize conditions under which this simple strategy is effective, and further show that extending it with a simple reweighting approach can further alleviate the bias introduced by using incorrect reward labels. Our empirical evaluation confirms these findings in simulated robotic locomotion, navigation, and manipulation settings.

Reachability Constrained Reinforcement Learning

- Dongjie Yu, Haitong Ma, Shengbo Li, Jianyu Chen
- abstract: Constrained reinforcement learning (CRL) has gained significant interest recently, since safety constraints satisfaction is critical for real-world problems. However, existing CRL methods constraining discounted cumulative costs generally lack rigorous definition and guarantee of safety. In contrast, in the safe control research, safety is defined as persistently satisfying certain state constraints. Such persistent safety is possible only on a subset of the state space, called feasible set, where an optimal largest feasible set exists for a given environment. Recent studies incorporate feasible sets into CRL with energy-based methods such as control barrier function (CBF), safety index (SI), and leverage prior conservative estimations of feasible sets, which harms the performance of the learned policy. To deal with this problem, this paper proposes the reachability CRL (RCRL) method by using reachability analysis to establish the novel self-consistency condition and characterize the feasible sets. The feasible sets are represented by the safety value function, which is used as the constraint in CRL. We use the multi-time scale stochastic approximation theory to prove that the proposed algorithm converges to a local optimum, where the largest feasible set can be guaranteed. Empirical results on different benchmarks validate the learned feasible set, the policy performance, and constraint satisfaction of RCRL, compared to CRL and safe control baselines.

Topology-Aware Network Pruning using Multi-stage Graph Embedding and Reinforcement Learning

- Sixing Yu, Arya Mazaheri, Ali Jannesari
- abstract: Model compression is an essential technique for deploying deep neural networks (DNNs) on power and memory-constrained resources. However, existing model-compression methods often rely on human expertise and focus on parameters' local importance, ignoring the rich topology information within DNNs. In this paper, we propose a novel multi-stage graph embedding technique based on graph neural networks (GNNs) to identify DNN topologies and use reinforcement learning (RL) to find a suitable compression policy. We performed resource-constrained (i.e., FLOPs) channel pruning and compared our approach with state-of-the-art model compression methods. We evaluated our method on various models from typical to mobile-friendly networks, such as ResNet family, VGG-16, MobileNet-v1/v2, and ShuffleNet. Results show that our method can achieve higher compression ratios with a minimal fine-tuning cost yet yields outstanding and competitive performance.

The Combinatorial Brain Surgeon: Pruning Weights That Cancel One Another in Neural Networks

- Xin Yu, Thiago Serra, Srikumar Ramalingam, Shandian Zhe
- abstract: Neural networks tend to achieve better accuracy with training if they are larger {—} even if the resulting models are overparameterized. Nevertheless, carefully removing such excess of parameters before, during, or after training may also produce models with similar or even improved accuracy. In many cases, that can be curiously achieved by heuristics as simple as removing a percentage of the weights with the smallest absolute value {—} even though absolute value is not a perfect proxy for weight relevance. With the premise that obtaining significantly better performance from pruning depends on accounting for the combined effect of removing multiple weights, we revisit one of the classic approaches for impact-based pruning: the Optimal Brain Surgeon (OBS). We propose a tractable heuristic for solving the combinatorial extension of OBS, in which we select weights for simultaneous removal, and we combine it with a single-pass systematic update of unpruned weights. Our selection method outperforms other methods for high sparsity, and the single-pass weight update is also advantageous if applied after those methods.

GraphFM: Improving Large-Scale GNN Training via Feature Momentum

- Haiyang Yu, Limei Wang, Bokun Wang, Meng Liu, Tianbao Yang, Shuiwang Ji
- abstract: Training of graph neural networks (GNNs) for large-scale node classification is challenging. A key difficulty lies in obtaining accurate hidden node representations while avoiding the neighborhood explosion problem. Here, we propose a new technique, named feature momentum (FM), that uses a momentum step to incorporate historical embeddings when updating feature representations. We develop two specific algorithms, known as GraphFM-IB and GraphFM-OB, that consider in-batch and out-of-batch data, respectively. GraphFM-IB applies FM to in-batch sampled data, while GraphFM-OB applies FM to out-of-batch data that are 1-hop neighborhood of in-batch data. We provide a convergence analysis for GraphFM-IB and some theoretical insight for GraphFM-OB. Empirically, we observe that GraphFM-IB can effectively alleviate the neighborhood explosion problem of existing methods. In addition, GraphFM-OB achieves promising performance on multiple large-scale graph datasets.

Latent Diffusion Energy-Based Model for Interpretable Text Modelling

- Peiyu Yu, Sirui Xie, Xiaojian Ma, Baoxiong Jia, Bo Pang, Ruiqi Gao, Yixin Zhu, Song-Chun Zhu, Ying Nian Wu
- abstract: Latent space Energy-Based Models (EBMs), also known as energy-based priors, have drawn growing interests in generative modeling. Fueled by its flexibility in the formulation and strong modeling power of the latent space, recent works built upon it have made interesting attempts aiming at the interpretability of text modeling. However, latent space EBMs also inherit some flaws from EBMs in data space; the degenerate MCMC sampling quality in practice can lead to poor generation quality and instability in training, especially on data with complex latent structures. Inspired by the recent efforts that leverage diffusion recovery likelihood learning as a cure for the sampling issue, we introduce a novel symbiosis between the diffusion models and latent space EBMs in a variational learning framework, coined as the latent diffusion energy-based model. We develop a geometric clustering-based regularization jointly with the information bottleneck to further improve the quality of the learned latent space. Experiments on several challenging tasks demonstrate the superior performance of our model on interpretable text modeling over strong counterparts.

Predicting Out-of-Distribution Error with the Projection Norm

- Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, Jacob Steinhardt
- abstract: We propose a metric—Projection Norm—to predict a model's performance on out-of-distribution (OOD) data without access to ground truth labels. Projection Norm first uses model predictions to pseudo-label test samples and then trains a new model on the pseudo-labels. The more the new model's parameters differ from an in-distribution model, the greater the predicted OOD error. Empirically, our approach outperforms existing methods on both image and text classification tasks and across different network architectures. Theoretically, we connect our approach to a bound on the test error for overparameterized linear models. Furthermore, we find that Projection Norm is the only approach that achieves non-trivial detection performance on adversarial examples. Our code is available at \url{https://github.com/yaodongyu/ProjNorm}.

Robust Task Representations for Offline Meta-Reinforcement Learning via Contrastive Learning

- Haoqi Yuan, Zongqing Lu
- abstract: We study offline meta-reinforcement learning, a practical reinforcement learning paradigm that learns from offline data to adapt to new tasks. The distribution of offline data is determined jointly by the behavior policy and the task. Existing offline meta-reinforcement learning algorithms cannot distinguish these factors, making task representations unstable to the change of behavior policies. To address this problem, we propose a contrastive learning framework for task representations that are robust to the distribution mismatch of behavior policies in training and test. We design a bi-level encoder structure, use mutual information maximization to formalize task representation learning, derive a contrastive learning objective, and introduce several approaches to approximate the true distribution of negative pairs. Experiments on a variety of offline meta-reinforcement learning benchmarks demonstrate the advantages of our method over prior methods, especially on the generalization to out-of-distribution behavior policies.

Provable Stochastic Optimization for Global Contrastive Learning: Small Batch Does Not Harm Performance

- Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, Tianbao Yang
- abstract: In this paper, we study contrastive learning from an optimization perspective, aiming to analyze and address a fundamental issue of existing contrastive learning methods that either rely on a large batch size or a large dictionary of feature vectors. We consider a global objective for contrastive learning, which contrasts each positive pair with all negative pairs for an anchor point. From the optimization perspective, we explain why existing methods such as SimCLR require a large batch size in order to achieve a satisfactory result. In order to remove such requirement, we propose a memory-efficient Stochastic Optimization algorithm for solving the Global objective of Contrastive Learning of Representations, named SogCLR. We show that its optimization error is negligible under a reasonable condition after a sufficient number of iterations or is diminishing for a slightly different global contrastive objective. Empirically, we demonstrate that SogCLR with small batch size (e.g., 256) can achieve similar performance as SimCLR with large batch size (e.g., 8192) on self-supervised learning task on ImageNet-1K. We also attempt to show that the proposed optimization technique is generic and can be applied to solving other contrastive losses, e.g., two-way contrastive losses for bimodal contrastive learning. The proposed method is implemented in our open-sourced library LibAUC (www.libauc.org).

Neural Tangent Kernel Empowered Federated Learning

- Kai Yue, Richeng Jin, Ryan Pilgrim, Chau-Wai Wong, Dror Baron, Huaiyu Dai
- abstract: Federated learning (FL) is a privacy-preserving paradigm where multiple participants jointly solve a machine learning problem without sharing raw data. Unlike traditional distributed learning, a unique characteristic of FL is statistical heterogeneity, namely, data distributions across participants are different from each other. Meanwhile, recent advances in the interpretation of neural networks have seen a wide use of neural tangent kernels (NTKs) for convergence analyses. In this paper, we propose a novel FL paradigm empowered by the NTK framework. The paradigm addresses the challenge of statistical heterogeneity by transmitting update data that are more expressive than those of the conventional FL paradigms. Specifically, sample-wise Jacobian matrices, rather than model weights/gradients, are uploaded by participants. The server then constructs an empirical kernel matrix to update a global model without explicitly performing gradient descent. We further develop a variant with improved communication efficiency and enhanced privacy. Numerical results show that the proposed paradigm can achieve the same accuracy while reducing the number of communication rounds by an order of magnitude compared to federated averaging.

Time Is MattEr: Temporal Self-supervision for Video Transformers

- Sukmin Yun, Jaehyung Kim, Dongyoon Han, Hwanjun Song, Jung-Woo Ha, Jinwoo Shin
- abstract: Understanding temporal dynamics of video is an essential aspect of learning better video representations. Recently, transformer-based architectural designs have been extensively explored for video tasks due to their capability to capture long-term dependency of input sequences. However, we found that these Video Transformers are still biased to learn spatial dynamics rather than temporal ones, and debiasing the spurious correlation is critical for their performance. Based on the observations, we design simple yet effective self-supervised tasks for video models to learn temporal dynamics better. Specifically, for debiasing the spatial bias, our method learns the temporal order of video frames as extra self-supervision and enforces the randomly shuffled frames to have low-confidence outputs. Also, our method learns the temporal flow direction of video tokens among consecutive frames for enhancing the correlation toward temporal dynamics. Under various video action recognition tasks, we demonstrate the effectiveness of our method and its compatibility with state-of-the-art Video Transformers.

Pure Noise to the Rescue of Insufficient Data: Improving Imbalanced Classification by Training on Random Noise Images

- Shiran Zada, Itay Benou, Michal Irani
- abstract: Despite remarkable progress on visual recognition tasks, deep neural-nets still struggle to generalize well when training data is scarce or highly imbalanced, rendering them extremely vulnerable to real-world examples. In this paper, we present a surprisingly simple yet highly effective method to mitigate this limitation: using pure noise images as additional training data. Unlike the common use of additive noise or adversarial noise for data augmentation, we propose an entirely different perspective by directly training on pure random noise images. We present a new Distribution-Aware Routing Batch Normalization layer (DAR-BN), which enables training on pure noise images in addition to natural images within the same network. This encourages generalization and suppresses overfitting. Our proposed method significantly improves imbalanced classification performance, obtaining state-

of-the-art results on a large variety of long-tailed image classification datasets (CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, Places-LT, and CelebA-5). Furthermore, our method is extremely simple and easy to use as a general new augmentation tool (on top of existing augmentations), and can be incorporated in any training scheme. It does not require any specialized data generation or training procedures, thus keeping training fast and efficient.

Adaptive Conformal Predictions for Time Series

- Margaux Zaffran, Olivier Feron, Yannig Goude, Julie Josse, Aymeric Dieuleveut
- abstract: Uncertainty quantification of predictive models is crucial in decision-making problems. Conformal prediction is a general and theoretically sound answer. However, it requires exchangeable data, excluding time series. While recent works tackled this issue, we argue that Adaptive Conformal Inference (ACI, Gibbs & Cand{è}s, 2021), developed for distribution-shift time series, is a good procedure for time series with general dependency. We theoretically analyse the impact of the learning rate on its efficiency in the exchangeable and auto-regressive case. We propose a parameter-free method, AgACI, that adaptively builds upon ACI based on online expert aggregation. We lead extensive fair simulations against competing methods that advocate for ACI's use in time series. We conduct a real case study: electricity price forecasting. The proposed aggregation algorithm provides efficient prediction intervals for day-ahead forecasting. All the code and data to reproduce the experiments are made available on GitHub.

Actor-Critic based Improper Reinforcement Learning

- Mohammadi Zaki, Avi Mohan, Aditya Gopalan, Shie Mannor
- abstract: We consider an improper reinforcement learning setting where a learner is given M base controllers for an unknown Markov decision process, and wishes to combine them optimally to produce a potentially new controller that can outperform each of the base ones. This can be useful in tuning across controllers, learnt possibly in mismatched or simulated environments, to obtain a good controller for a given target environment with relatively few trials. Towards this, we propose two algorithms: (1) a Policy Gradient-based approach; and (2) an algorithm that can switch between a simple Actor-Critic (AC) based scheme and a Natural Actor-Critic (NAC) scheme depending on the available information. Both algorithms operate over a class of improper mixtures of the given controllers. For the first case, we derive convergence rate guarantees assuming access to a gradient oracle. For the AC-based approach we provide convergence rate guarantees to a stationary point in the basic AC case and to a global optimum in the NAC case. Numerical results on (i) the standard control theoretic benchmark of stabilizing an inverted pendulum; and (ii) a constrained queueing task show that our improper policy optimization algorithm can stabilize the system even when the base policies at its disposal are unstable.

Stabilizing Q-learning with Linear Architectures for Provable Efficient Learning

- Andrea Zanette, Martin Wainwright
- abstract: The Q-learning algorithm is a simple, fundamental and practically very effective reinforcement learning algorithm. However, the basic protocol can exhibit an unstable behavior when implemented even with simple linear function approximation. While tools like target networks and experience replay are often implemented to stabilize the learning process, the individual contribution of each of these mechanisms is not well understood theoretically. This work proposes an exploration variant of the basic Q-learning protocol with linear function approximation. Our modular analysis illustrates the role played by each algorithmic tool that we adopt: a second order update rule, a set of target networks, and a mechanism akin to experience replay. Together, they enable state of the art regret bounds on linear MDPs while preserving the most prominent feature of the algorithm, namely a space complexity independent of the number of steps elapsed. Furthermore, we show that the performance of the algorithm degrades very gracefully under a new, more permissive notion of approximation error. Finally, the algorithm partially inherits problem dependent regret bounds, function of the number of 'effective' feature dimension.

Multi Resolution Analysis (MRA) for Approximate Self-Attention

- Zhanpeng Zeng, Sourav Pal, Jeffery Kline, Glenn M Fung, Vikas Singh
- abstract: Transformers have emerged as a preferred model for many tasks in natural langugage processing and vision. Recent efforts on training and deploying Transformers more efficiently have identified many strategies to approximate the self-attention matrix, a key module in a Transformer architecture. Effective ideas include various prespecified sparsity patterns, low-rank basis expansions and combinations thereof. In this paper, we revisit classical Multiresolution Analysis (MRA) concepts such as Wavelets, whose potential value in this setting remains underexplored thus far. We show that simple approximations based on empirical feedback and design choices informed by modern hardware and implementation challenges, eventually yield a MRA-based approach for self-attention with an excellent performance profile across most criteria of interest. We undertake an extensive set of experiments and demonstrate that this multi-resolution scheme outperforms most efficient self-attention proposals and is favorable for both short and long sequences. Code is available at \url{https://github.com/mlpen/mra-attention}.

Efficient PAC Learning from the Crowd with Pairwise Comparisons

- Shiwei Zeng, Jie Shen
- abstract: We study crowdsourced PAC learning of threshold function, where the labels are gathered from a pool of annotators some of whom may behave adversarially. This is yet a challenging problem and until recently has computationally and query efficient PAC learning algorithm been established by Awasthi et al. (2017). In this paper, we show that by leveraging the more easily acquired pairwise comparison queries, it is possible to exponentially reduce the label complexity while retaining the overall query complexity and runtime. Our main algorithmic contributions are a comparison-equipped labeling scheme that can faithfully recover the true labels of a small set of instances, and a label-efficient filtering process that in conjunction with the small labeled set can reliably infer the true labels of a large instance set.

Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts

- Yan Zeng, Xinsong Zhang, Hang Li
- abstract: Most existing methods in vision language pre-training rely on object-centric features extracted through object detection and make fine-grained alignments between the extracted features and texts. It is challenging for these methods to learn relations among multiple objects. To this end, we propose a new method called X-VLM to perform 'multi-grained vision language pre-training.' The key to learning multi-grained alignments is to locate visual concepts in the image given the associated texts, and in the meantime align the texts with the visual concepts, where the alignments are in multi-granularity. Experimental results show that X-VLM effectively leverages the learned multi-grained alignments to many downstream vision language tasks and consistently outperforms state-of-the-art methods.

Position Prediction as an Effective Pretraining Strategy

- Shuangfei Zhai, Navdeep Jaitly, Jason Ramapuram, Dan Busbridge, Tatiana Likhomanenko, Joseph Y Cheng, Walter Talbott, Chen Huang, Hanlin Goh, Joshua M Susskind
- abstract: Transformers \cite{transformer} have gained increasing popularity in a wide range of applications, including Natural Language Processing (NLP), Computer Vision and Speech Recognition, because of their powerful representational capacity. However, harnessing this representational capacity effectively requires a large amount of data, strong regularization, or both, to mitigate overfitting. Recently, the power of the Transformer has been unlocked by self-supervised pretraining strategies based on masked autoencoders which rely on reconstructing masked inputs, directly, or contrastively

from unmasked content. This pretraining strategy which has been used in BERT models in NLP \cite{bert}, Wav2Vec models in Speech \cite{wv2v2} and, recently, in MAE models in Vision \cite{beit, mae}, forces the model to learn about relationships between the content in different parts of the input using autoencoding related objectives. In this paper, we propose a novel, but surprisingly simple alternative to content reconstruction – that of predicting locations from content, without providing positional information for it. Doing so requires the Transformer to understand the positional relationships between different parts of the input, from their content alone. This amounts to an efficient implementation where the pretext task is a classification problem among all possible positions for each input token. We experiment on both Vision and Speech benchmarks, where our approach brings improvements over strong supervised training baselines and is comparable to modern unsupervised/self-supervised pretraining methods. Our method also enables Transformers trained without position embeddings to outperform ones trained with full position information.

Anytime Information Cascade Popularity Prediction via Self-Exciting Processes

- Xi Zhang, Akshay Aravamudan, Georgios C Anagnostopoulos
- abstract: One important aspect of understanding behaviors of information cascades is to be able to accurately predict their popularity, that is, their message counts at any future time. Self-exciting Hawkes processes have been widely adopted for such tasks due to their success in describing cascading behaviors. In this paper, for general, marked Hawkes point processes, we present closed-form expressions for the mean and variance of future event counts, conditioned on observed events. Furthermore, these expressions allow us to develop a predictive approach, namely, Cascade Anytime Size Prediction via self-Exciting Regression model (CASPER), which is specifically tailored to popularity prediction, unlike existing generative approaches {–} based on point processes {–} for the same task. We showcase CASPER's merits via experiments entailing both synthetic and real-world data, and demonstrate that it considerably improves upon prior works in terms of accuracy, especially for early-stage prediction.

Understanding Clipping for Federated Learning: Convergence and Client-Level Differential Privacy

- Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Steven Wu, Jinfeng Yi
- abstract: Providing privacy protection has been one of the primary motivations of Federated Learning (FL). Recently, there has been a line of work on incorporating the formal privacy notion of differential privacy with FL. To guarantee the client-level differential privacy in FL algorithms, the clients' transmitted model updates have to be clipped before adding privacy noise. Such clipping operation is substantially different from its counterpart of gradient clipping in the centralized differentially private SGD and has not been well-understood. In this paper, we first empirically demonstrate that the clipped FedAvg can perform surprisingly well even with substantial data heterogeneity when training neural networks, which is partly because the clients' updates become similar for several popular deep architectures. Based on this key observation, we provide the convergence analysis of a differential private (DP) FedAvg algorithm and highlight the relationship between clipping bias and the distribution of the clients' updates. To the best of our knowledge, this is the first work that rigorously investigates theoretical and empirical issues regarding the clipping operation in FL algorithms.

Collaboration of Experts: Achieving 80% Top-1 Accuracy on ImageNet with 100M FLOPs

- Yikang Zhang, Zhuo Chen, Zhao Zhong
- abstract: In this paper, we propose a Collaboration of Experts (CoE) framework to assemble the expertise of multiple networks towards a common goal. Each expert is an individual network with expertise on a unique portion of the dataset, contributing to the collective capacity. Given a sample, delegator selects an expert and simultaneously outputs a rough prediction to trigger potential early termination. For each model in CoE, we propose a novel training algorithm with two major components: weight generation module (WGM) and label generation module (LGM). It fulfills the co-adaptation of experts and delegator. WGM partitions the training data into portions based on delegator via solving a balanced transportation problem, then impels each expert to focus on one portion by reweighting the losses. LGM generates the label to constitute the loss of delegator for expert selection. CoE achieves the state-of-the-art performance on ImageNet, 80.7% top-1 accuracy with 194M FLOPs. Combined with PWLU and CondConv, CoE further boosts the accuracy to 80.0% with only 100M FLOPs for the first time. Furthermore, experiment results on the translation task also demonstrate the strong generalizability of CoE. CoE is hardware-friendly, yielding a 3 6x acceleration compared with existing conditional computation approaches.

PDE-Based Optimal Strategy for Unconstrained Online Learning

- Zhiyu Zhang, Ashok Cutkosky, Ioannis Paschalidis
- abstract: Unconstrained Online Linear Optimization (OLO) is a practical problem setting to study the training of machine learning models. Existing works proposed a number of potential-based algorithms, but in general the design of these potential functions relies heavily on guessing. To streamline this workflow, we present a framework that generates new potential functions by solving a Partial Differential Equation (PDE). Specifically, when losses are 1-Lipschitz, our framework produces a novel algorithm with anytime regret bound $C\sqrt{T}+\|u\|\sqrt{2T}[\sqrt{\log(1+\|u\|/C)}+2]$, where C is a user-specified constant and u is any comparator unknown and unbounded a priori. Such a bound attains an optimal loss-regret trade-off without the impractical doubling trick. Moreover, a matching lower bound shows that the leading order term, including the constant multiplier $\sqrt{2}$, is tight. To our knowledge, the proposed algorithm is the first to achieve such optimalities.

Stochastic Continuous Submodular Maximization: Boosting via Non-oblivious Function

- Qixin Zhang, Zengde Deng, Zaiyi Chen, Haoyuan Hu, Yu Yang
- abstract: In this paper, we revisit Stochastic Continuous Submodular Maximization in both offline and online settings, which can benefit wide applications in machine learning and operations research areas. We present a boosting framework covering gradient ascent and online gradient ascent. The fundamental ingredient of our methods is a novel non-oblivious function F derived from a factor-revealing optimization problem, whose any stationary point provides a $(1-e^{-\gamma})$-approximation to the global maximum of the γ-weakly DR-submodular objective function $f\in C^{1,1}_L(\mathcal{X})$. Under the offline scenario, we propose a boosting gradient ascent method achieving $(1-e^{-\gamma}-\epsilon^{2})$-approximation after $O(1/\epsilon^2)$ iterations, which improves the $(\frac{\gamma^2}{1+\gamma^2})$ approximation ratio of the classical gradient ascent algorithm. In the online setting, for the first time we consider the adversarial delays for stochastic gradient feedback, under which we propose a boosting online gradient algorithm with the same non-oblivious function F. Meanwhile, we verify that this boosting online algorithm achieves a regret of $O(\sqrt{D})$ against a $(1-e^{-\gamma})$-approximation to the best feasible solution in hindsight, where D is the sum of delays of gradient feedback. To the best of our knowledge, this is the first result to obtain $O(\sqrt{T})$ regret against a $(1-e^{-\gamma})$-approximation with $O(1)$ gradient inquiry at each time step, when no delay exists, i.e., $D=T$. Finally, numerical experiments demonstrate the effectiveness of our boosting methods.

When and How Mixup Improves Calibration

- Linjun Zhang, Zhun Deng, Kenji Kawaguchi, James Zou
- abstract: In many machine learning applications, it is important for the model to provide confidence scores that accurately capture its prediction uncertainty. Although modern learning methods have achieved great success in predictive accuracy, generating calibrated confidence scores remains a major challenge. Mixup, a popular yet simple data augmentation technique based on taking convex combinations of pairs of training examples, has been empirically found to significantly improve confidence calibration across diverse applications. However, when and how Mixup helps calibration is still a mystery. In this paper, we theoretically prove that Mixup improves calibration in high-dimensional settings by investigating natural statistical models. Interestingly, the calibration benefit of Mixup increases as the model capacity increases. We support our theories with experiments on common architectures and datasets. In addition, we study how Mixup improves calibration in semi-supervised learning. While incorporating unlabeled data can

sometimes make the model less calibrated, adding Mixup training mitigates this issue and provably improves calibration. Our analysis provides new insights and a framework to understand Mixup and calibration.

UAST: Uncertainty-Aware Siamese Tracking

- Dawei Zhang, Yanwei Fu, Zhonglong Zheng
- abstract: Visual object tracking is basically formulated as target classification and bounding box estimation. Recent anchor-free Siamese trackers rely on predicting the distances to four sides for efficient regression but fail to estimate accurate bounding box in complex scenes. We argue that these approaches lack a clear probabilistic explanation, so it is desirable to model the uncertainty and ambiguity representation of target estimation. To address this issue, this paper presents an Uncertainty-Aware Siamese Tracker (UAST) by developing a novel distribution-based regression formulation with localization uncertainty. We exploit regression vectors to directly represent the discretized probability distribution for four offsets of boxes, which is general, flexible and informative. Based on the resulting distributed representation, our method is able to provide a probabilistic value of uncertainty. Furthermore, considering the high correlation between the uncertainty and regression accuracy, we propose to learn a joint representation head of classification and localization quality for reliable tracking, which also avoids the inconsistency of classification and quality estimation between training and inference. Extensive experiments on several challenging tracking benchmarks demonstrate the effectiveness of UAST and its superiority over other Siamese trackers.

Examining Scaling and Transfer of Language Model Architectures for Machine Translation

- Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, Orhan Firat
- abstract: Natural language understanding and generation models follow one of the two dominant architectural paradigms: language models (LMs) that process concatenated sequences in a single stack of layers, and encoder-decoder models (EncDec) that utilize separate layer stacks for input and output processing. In machine translation, EncDec has long been the favoured approach, but with few studies investigating the performance of LMs. In this work, we thoroughly examine the role of several architectural design choices on the performance of LMs on bilingual, (massively) multilingual and zero-shot translation tasks, under systematic variations of data conditions and model sizes. Our results show that: (i) Different LMs have different scaling properties, where architectural differences often have a significant impact on model performance at small scales, but the performance gap narrows as the number of parameters increases, (ii) Several design choices, including causal masking and language-modeling objectives for the source sequence, have detrimental effects on translation quality, and (iii) When paired with full-visible masking for source sequences, LMs could perform on par with EncDec on supervised bilingual and multilingual translation tasks, and improve greatly on zero-shot directions by facilitating the reduction of off-target translations.

Revisiting End-to-End Speech-to-Text Translation From Scratch

- Biao Zhang, Barry Haddow, Rico Sennrich
- abstract: End-to-end (E2E) speech-to-text translation (ST) often depends on pretraining its encoder and/or decoder using source transcripts via speech recognition or text translation tasks, without which translation performance drops substantially. However, transcripts are not always available, and how significant such pretraining is for E2E ST has rarely been studied in the literature. In this paper, we revisit this question and explore the extent to which the quality of E2E ST trained on speech-translation pairs alone can be improved. We reexamine several techniques proven beneficial to ST previously, and offer a set of best practices that biases a Transformer-based E2E ST system toward training from scratch. Besides, we propose parameterized distance penalty to facilitate the modeling of locality in the self-attention model for speech. On four benchmarks covering 23 languages, our experiments show that, without using any transcripts or pretraining, the proposed system reaches and even outperforms previous studies adopting pretraining, although the gap remains in (extremely) low-resource settings. Finally, we discuss neural acoustic feature modeling, where a neural model is designed to extract acoustic features from raw speech signals directly, with the goal to simplify inductive biases and add freedom to the model in describing speech. For the first time, we demonstrate its feasibility and show encouraging results on ST tasks.

A Stochastic Multi-Rate Control Framework For Modeling Distributed Optimization Algorithms

- Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Nicola Elia
- abstract: In modern machine learning systems, distributed algorithms are deployed across applications to ensure data privacy and optimal utilization of computational resources. This work offers a fresh perspective to model, analyze, and design distributed optimization algorithms through the lens of stochastic multi-rate feedback control. We show that a substantial class of distributed algorithms—including popular Gradient Tracking for decentralized learning, and FedPD and Scaffold for federated learning—can be modeled as a certain discrete-time stochastic feedback-control system, possibly with multiple sampling rates. This key observation allows us to develop a generic framework to analyze the convergence of the entire algorithm class. It also enables one to easily add desirable features such as differential privacy guarantees, or to deal with practical settings such as partial agent participation, communication compression, and imperfect communication in algorithm design and analysis.

GALAXY: Graph-based Active Learning at the Extreme

- Jifan Zhang, Julian Katz-Samuels, Robert Nowak
- abstract: Active learning is a label-efficient approach to train highly effective models while interactively selecting only small subsets of unlabelled data for labelling and training. In “open world" settings, the classes of interest can make up a small fraction of the overall dataset – most of the data may be viewed as an out-of-distribution or irrelevant class. This leads to extreme class-imbalance, and our theory and methods focus on this core issue. We propose a new strategy for active learning called GALAXY (Graph-based Active Learning At the eXtrEme), which blends ideas from graph-based active learning and deep learning. GALAXY automatically and adaptively selects more class-balanced examples for labeling than most other methods for active learning. Our theory shows that GALAXY performs a refined form of uncertainty sampling that gathers a much more class-balanced dataset than vanilla uncertainty sampling. Experimentally, we demonstrate GALAXY’s superiority over existing state-of-art deep active learning algorithms in unbalanced vision classification settings generated from popular datasets.

Fairness Interventions as (Dis)Incentives for Strategic Manipulation

- Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, Mingyan Liu
- abstract: Although machine learning (ML) algorithms are widely used to make decisions about individuals in various domains, concerns have arisen that (1) these algorithms are vulnerable to strategic manipulation and "gaming the algorithm"; and (2) ML decisions may exhibit bias against certain social groups. Existing works have largely examined these as two separate issues, e.g., by focusing on building ML algorithms robust to strategic manipulation, or on training a fair ML algorithm. In this study, we set out to understand the impact they each have on the other, and examine how to characterize fair policies in the presence of strategic behavior. The strategic interaction between a decision maker and individuals (as decision takers) is modeled as a two-stage (Stackelberg) game; when designing an algorithm, the former anticipates the latter may manipulate their features in order to receive more favorable decisions. We analytically characterize the equilibrium strategies of both, and examine how the algorithms and their resulting fairness properties are affected when the decision maker is strategic (anticipates manipulation), as well as the impact of fairness interventions on equilibrium strategies. In particular, we identify conditions under which anticipation of strategic behavior may mitigate/exacerbate unfairness, and conditions under which fairness interventions can serve as (dis)incentives for strategic manipulation.

Role-based Multiplex Network Embedding

- Hegui Zhang, Gang Kou
- abstract: In recent years, multiplex network embedding has received great attention from researchers. However, existing multiplex network embedding methods neglect structural role information, which can be used to determine the structural similarity between nodes. To overcome this shortcoming, this work proposes a simple, effective, role-based embedding method for multiplex networks, called RMNE. The RMNE uses the structural role information of nodes to preserve the structural similarity between nodes in the entire multiplex network. Specifically, a role-modified random walk is designed to generate node sequences of each node, which can capture both the within-layer neighbors, structural role members, and cross-layer structural role members of a node. Additionally, the variant of RMNE extends the existing collaborative embedding method by unifying the structural role information into our method to obtain the role-based node representations. Finally, the proposed methods were evaluated on the network reconstruction, node classification, link prediction, and multi-class edge classification tasks. The experimental results on eight public, real-world multiplex networks demonstrate that the proposed methods outperform state-of-the-art baseline methods.

Dynamic Topic Models for Temporal Document Networks

- Delvin Ce Zhang, Hady Lauw
- abstract: Dynamic topic models explore the time evolution of topics in temporally accumulative corpora. While existing topic models focus on the dynamics of individual documents, we propose two neural topic models aimed at learning unified topic distributions that incorporate both document dynamics and network structure. For the first model, by adding a time dimension, we propose Time-Aware Optimal Transport, which measures the probability of a link between two differently timestamped documents using their semantic distance. Since the gradually evolving topological structure of network may also influence the establishment of a new link, for the second model, we further design a Temporal Point Process to capture the impact of historical neighbors on the current link formation at the network level. Experiments on four dynamic document networks demonstrate the advantage of our models in jointly modeling document dynamics and network adjacency.

Personalized Federated Learning via Variational Bayesian Inference

- Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, Yunfeng Shao
- abstract: Federated learning faces huge challenges from model overfitting due to the lack of data and statistical diversity among clients. To address these challenges, this paper proposes a novel personalized federated learning method via Bayesian variational inference named pFedBayes. To alleviate the overfitting, weight uncertainty is introduced to neural networks for clients and the server. To achieve personalization, each client updates its local distribution parameters by balancing its construction error over private data and its KL divergence with global distribution from the server. Theoretical analysis gives an upper bound of averaged generalization error and illustrates that the convergence rate of the generalization error is minimax optimal up to a logarithmic factor. Experiments show that the proposed method outperforms other advanced personalized methods on personalized models, e.g., pFedBayes respectively outperforms other SOTA algorithms by 1.25%, 0.42% and 11.71% on MNIST, FMNIST and CIFAR-10 under non-i.i.d. limited data.

Federated Learning with Label Distribution Skew via Logits Calibration

- Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Chao Wu
- abstract: Traditional federated optimization methods perform poorly with heterogeneous data (i.e. , accuracy reduction), especially for highly skewed data. In this paper, we investigate the label distribution skew in FL, where the distribution of labels varies across clients. First, we investigate the label distribution skew from a statistical view. We demonstrate both theoretically and empirically that previous methods based on softmax cross-entropy are not suitable, which can result in local models heavily overfitting to minority classes and missing classes. Additionally, we theoretically introduce a deviation bound to measure the deviation of the gradient after local update. At last, we propose FedLC (\textbf{Fed}erated learning via \textbf{L}ogits \textbf{C}alibration), which calibrates the logits before softmax cross-entropy according to the probability of occurrence of each class. FedLC applies a fine-grained calibrated cross-entropy loss to local update by adding a pairwise label margin. Extensive experiments on federated datasets and real-world datasets demonstrate that FedLC leads to a more accurate global model and much improved performance. Furthermore, integrating other FL methods into our approach can further enhance the performance of the global model.

Neural Network Weights Do Not Converge to Stationary Points: An Invariant Measure Perspective

- Jingzhao Zhang, Haochuan Li, Suvrit Sra, Ali Jadbabaie
- abstract: This work examines the deep disconnect between existing theoretical analyses of gradient-based algorithms and the practice of training deep neural networks. Specifically, we provide numerical evidence that in large-scale neural network training (e.g., ImageNet + ResNet101, and WT103 + TransformerXL models), the neural network's weights do not converge to stationary points where the gradient of the loss is zero. Remarkably, however, we observe that even though the weights do not converge to stationary points, the progress in minimizing the loss function halts and training loss stabilizes. Inspired by this observation, we propose a new perspective based on ergodic theory of dynamical systems to explain it. Rather than studying the evolution of weights, we study the evolution of the distribution of weights. We prove convergence of the distribution of weights to an approximate invariant measure, thereby explaining how the training loss can stabilize without weights necessarily converging to stationary points. We further discuss how this perspective can better align optimization theory with empirical observations in machine learning practice.

Beyond Worst-Case Analysis in Stochastic Approximation: Moment Estimation Improves Instance Complexity

- Jingzhao Zhang, Hongzhou Lin, Subhro Das, Suvrit Sra, Ali Jadbabaie
- abstract: We study oracle complexity of gradient based methods for stochastic approximation problems. Though in many settings optimal algorithms and tight lower bounds are known for such problems, these optimal algorithms do not achieve the best performance when used in practice. We address this theory-practice gap by focusing on instance-dependent complexity instead of worst case complexity. In particular, we first summarize known instance-dependent complexity results and categorize them into three levels. We identify the domination relation between different levels and propose a fourth instance-dependent bound that dominates existing ones. We then provide a sufficient condition according to which an adaptive algorithm with moment estimation can achieve the proposed bound without knowledge of noise levels. Our proposed algorithm and its analysis provide a theoretical justification for the success of moment estimation as it achieves improved instance complexity.

Deep and Flexible Graph Neural Architecture Search

- Wentao Zhang, Zheyu Lin, Yu Shen, Yang Li, Zhi Yang, Bin Cui
- abstract: Graph neural networks (GNNs) have been intensively applied to various graph-based applications. Despite their success, designing good GNN architectures is non-trivial, which heavily relies on lots of human efforts and domain knowledge. Although several attempts have been made in graph neural architecture search, they suffer from the following limitations: 1) fixed pipeline pattern of propagation (P) and (T) transformation operations; 2) restricted pipeline depth of GNN architectures. This paper proposes DFG-NAS, a novel method that searches for deep and flexible GNN architectures. Unlike most existing methods that focus on micro-architecture, DFG-NAS highlights another level of design: the search for macro-architectures of how atomic P and T are integrated and organized into a GNN. Concretely, DFG-NAS proposes a novel-designed search space for the P-T permutations and combinations based on the message-passing dis-aggregation, and defines various mutation strategies and employs the evolutionary algorithm to conduct an efficient and effective search. Empirical studies on four benchmark datasets demonstrate that DFG-NAS could find more powerful architectures than state-of-the-art manual designs and meanwhile are more efficient than the current graph neural architecture search approaches.

A Langevin-like Sampler for Discrete Distributions

- Ruqi Zhang, Xingchao Liu, Qiang Liu
- abstract: We propose discrete Langevin proposal (DLP), a simple and scalable gradient-based proposal for sampling complex high-dimensional discrete distributions. In contrast to Gibbs sampling-based methods, DLP is able to update all coordinates in parallel in a single step and the magnitude of changes is controlled by a stepsize. This allows a cheap and efficient exploration in the space of high-dimensional and strongly correlated variables. We prove the efficiency of DLP by showing that the asymptotic bias of its stationary distribution is zero for log-quadratic distributions, and is small for distributions that are close to being log-quadratic. With DLP, we develop several variants of sampling algorithms, including unadjusted, Metropolis-adjusted, stochastic and preconditioned versions. DLP outperforms many popular alternatives on a wide variety of tasks, including Ising models, restricted Boltzmann machines, deep energy-based models, binary neural networks and language generation.

Rich Feature Construction for the Optimization-Generalization Dilemma

- Jianyu Zhang, David Lopez-Paz, Leon Bottou
- abstract: There often is a dilemma between ease of optimization and robust out-of-distribution (OoD) generalization. For instance, many OoD methods rely on penalty terms whose optimization is challenging. They are either too strong to optimize reliably or too weak to achieve their goals. We propose to initialize the networks with a rich representation containing a palette of potentially useful features, ready to be used by even simple models. On the one hand, a rich representation provides a good initialization for the optimizer. On the other hand, it also provides an inductive bias that helps OoD generalization. Such a representation is constructed with the Rich Feature Construction (RFC) algorithm, also called the Bonsai algorithm, which consists of a succession of training episodes. During discovery episodes, we craft a multi-objective optimization criterion and its associated datasets in a manner that prevents the network from using the features constructed in the previous iterations. During synthesis episodes, we use knowledge distillation to force the network to simultaneously represent all the previously discovered features. Initializing the networks with Bonsai representations consistently helps six OoD methods achieve top performance on ColoredMNIST benchmark. The same technique substantially outperforms comparable results on the Wilds Camelyon17 task, eliminates the high result variance that plagues other methods, and makes hyperparameter tuning and model selection more reliable.

Generative Flow Networks for Discrete Probabilistic Modeling

- Dinghuai Zhang, Nikolay Malkin, Zhen Liu, Alexandra Volokhova, Aaron Courville, Yoshua Bengio
- abstract: We present energy-based generative flow networks (EB-GFN), a novel probabilistic modeling algorithm for high-dimensional discrete data. Building upon the theory of generative flow networks (GFlowNets), we model the generation process by a stochastic data construction policy and thus amortize expensive MCMC exploration into a fixed number of actions sampled from a GFlowNet. We show how GFlowNets can approximately perform large-block Gibbs sampling to mix between modes. We propose a framework to jointly train a GFlowNet with an energy function, so that the GFlowNet learns to sample from the energy distribution, while the energy learns with an approximate MLE objective with negative samples from the GFlowNet. We demonstrate EB-GFN's effectiveness on various probabilistic modeling tasks. Code is publicly available at https://github.com/zdhNarsil/EB_GFN.

Neurotoxin: Durable Backdoors in Federated Learning

- Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, Joseph Gonzalez
- abstract: Federated learning (FL) systems have an inherent vulnerability to adversarial backdoor attacks during training due to their decentralized nature. The goal of the attacker is to implant backdoors in the learned model with poisoned updates such that at test time, the model's outputs can be fixed to a given target for certain inputs (e.g., if a user types "people from New York" into a mobile keyboard app that uses a backdoored next word prediction model, the model will autocomplete their sentence to "people in New York are rude"). Prior work has shown that backdoors can be inserted in FL, but these backdoors are not durable: they do not remain in the model after the attacker stops uploading poisoned updates because training continues, and in production FL systems an inserted backdoor may not survive until deployment. We propose Neurotoxin, a simple one-line backdoor attack that functions by attacking parameters that are changed less in magnitude during training. We conduct an exhaustive evaluation across ten natural language processing and computer vision tasks and find that we can double the durability of state of the art backdoors by adding a single line with Neurotoxin.

Making Linear MDPs Practical via Contrastive Representation Learning

- Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, Bo Dai
- abstract: It is common to address the curse of dimensionality in Markov decision processes (MDPs) by exploiting low-rank representations. This motivates much of the recent theoretical study on linear MDPs. However, most approaches require a given representation under unrealistic assumptions about the normalization of the decomposition or introduce unresolved computational challenges in practice. Instead, we consider an alternative definition of linear MDPs that automatically ensures normalization while allowing efficient representation learning via contrastive estimation. The framework also admits confidence-adjusted index algorithms, enabling an efficient and principled approach to incorporating optimism or pessimism in the face of uncertainty. To the best of our knowledge, this provides the first practical representation learning method for linear MDPs that achieves both strong theoretical guarantees and empirical performance. Theoretically, we prove that the proposed algorithm is sample efficient in both the online and offline settings. Empirically, we demonstrate superior performance over existing state-of-the-art model-based and model-free algorithms on several benchmarks.

NAFS: A Simple yet Tough-to-beat Baseline for Graph Representation Learning

- Wentao Zhang, Zeang Sheng, Mingyu Yang, Yang Li, Yu Shen, Zhi Yang, Bin Cui
- abstract: Recently, graph neural networks (GNNs) have shown prominent performance in graph representation learning by leveraging knowledge from both graph structure and node features. However, most of them have two major limitations. First, GNNs can learn higher-order structural information by stacking more layers but can not deal with large depth due to the over-smoothing issue. Second, it is not easy to apply these methods on large graphs due to the expensive computation cost and high memory usage. In this paper, we present node-adaptive feature smoothing (NAFS), a simple non-parametric method that constructs node representations without parameter learning. NAFS first extracts the features of each node with its neighbors of different hops by feature smoothing, and then adaptively combines the smoothed features. Besides, the constructed node representation can further be enhanced by the ensemble of smoothed features extracted via different smoothing strategies. We conduct experiments on four benchmark datasets on two different application scenarios: node clustering and link prediction. Remarkably, NAFS with feature ensemble outperforms the state-of-the-art GNNs on these tasks and mitigates the aforementioned two limitations of most learning-based GNN counterparts.

Correct-N-Contrast: a Contrastive Approach for Improving Robustness to Spurious Correlations

- Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, Christopher Re
- abstract: Spurious correlations pose a major challenge for robust machine learning. Models trained with empirical risk minimization (ERM) may learn to rely on correlations between class labels and spurious attributes, leading to poor performance on data groups without these correlations. This is challenging to address when the spurious attribute labels are unavailable. To improve worst-group performance on spuriously correlated data without training attribute labels, we propose Correct-N-Contrast (CNC), a contrastive approach to directly learn representations robust to spurious correlations. As ERM models can be good spurious attribute predictors, CNC works by (1) using a trained ERM model's outputs to identify samples with the same class but dissimilar spurious features, and (2) training a robust model with contrastive learning to learn similar representations for these samples. To support

CNC, we introduce new connections between worst-group error and a representation alignment loss that CNC aims to minimize. We empirically observe that worst-group error closely tracks with alignment loss, and prove that the alignment loss over a class helps upper-bound the class's worst-group vs. average error gap. On popular benchmarks, CNC reduces alignment loss drastically, and achieves state-of-the-art worst-group accuracy by 3.6% average absolute lift. CNC is also competitive with oracle methods that require group labels.

Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning approach

- Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, Wen Sun
- abstract: We present BRIEE, an algorithm for efficient reinforcement learning in Markov Decision Processes with block-structured dynamics (i.e., Block MDPs), where rich observations are generated from a set of unknown latent states. BRIEE interleaves latent states discovery, exploration, and exploitation together, and can provably learn a near-optimal policy with sample complexity scaling polynomially in the number of latent states, actions, and the time horizon, with no dependence on the size of the potentially infinite observation space. Empirically, we show that BRIEE is more sample efficient than the state-of-art Block MDP algorithm HOMER and other empirical RL baselines on challenging rich-observation combination lock problems which require deep exploration.

Partial Counterfactual Identification from Observational and Experimental Data

- Junzhe Zhang, Jin Tian, Elias Bareinboim
- abstract: This paper investigates the problem of bounding counterfactual queries from an arbitrary collection of observational and experimental distributions and qualitative knowledge about the underlying data-generating model represented in the form of a causal diagram. We show that all counterfactual distributions in an arbitrary structural causal model (SCM) with discrete observed domains could be generated by a canonical family of SCMs with the same causal diagram where unobserved (exogenous) variables are also discrete, taking values in finite domains. Utilizing the canonical SCMs, we translate the problem of bounding counterfactuals into that of polynomial programming whose solution provides optimal bounds for the counterfactual query. Solving such polynomial programs is in general computationally expensive. We then develop effective Monte Carlo algorithms to approximate optimal bounds from a combination of observational and experimental data. Our algorithms are validated extensively on synthetic and real-world datasets.

Set Norm and Equivariant Skip Connections: Putting the Deep in Deep Sets

- Lily Zhang, Veronica Tozzo, John Higgins, Rajesh Ranganath
- abstract: Permutation invariant neural networks are a promising tool for predictive modeling of set data. We show, however, that existing architectures struggle to perform well when they are deep. In this work, we mathematically and empirically analyze normalization layers and residual connections in the context of deep permutation invariant neural networks. We develop set norm, a normalization tailored for sets, and introduce the "clean path principle" for equivariant residual connections alongside a novel benefit of such connections, the reduction of information loss. Based on our analysis, we propose Deep Sets++ and Set Transformer++, deep models that reach comparable or better performance than their original counterparts on a diverse suite of tasks. We additionally introduce Flow-RBC, a new single-cell dataset and real-world application of permutation invariant prediction. We open-source our data and code here: https://github.com/rajesh-lab/deep_permutation_invariant.

Learning to Estimate and Refine Fluid Motion with Physical Dynamics

- Mingrui Zhang, Jianhong Wang, James B Tlhomole, Matthew Piggott
- abstract: Extracting information on fluid motion directly from images is challenging. Fluid flow represents a complex dynamic system governed by the Navier-Stokes equations. General optical flow methods are typically designed for rigid body motion, and thus struggle if applied to fluid motion estimation directly. Further, optical flow methods only focus on two consecutive frames without utilising historical temporal information, while the fluid motion (velocity field) can be considered a continuous trajectory constrained by time-dependent partial differential equations (PDEs). This discrepancy has the potential to induce physically inconsistent estimations. Here we propose an unsupervised learning based prediction-correction scheme for fluid flow estimation. An estimate is first given by a PDE-constrained optical flow predictor, which is then refined by a physical based corrector. The proposed approach outperforms optical flow methods and shows competitive results compared to existing supervised learning based methods on a benchmark dataset. Furthermore, the proposed approach can generalize to complex real-world fluid scenarios where ground truth information is effectively unknowable. Finally, experiments demonstrate that the physical corrector can refine flow estimates by mimicking the operator splitting method commonly utilised in fluid dynamical simulation.

A Branch and Bound Framework for Stronger Adversarial Attacks of ReLU Networks

- Huan Zhang, Shiqi Wang, Kaidi Xu, Yihan Wang, Suman Jana, Cho-Jui Hsieh, Zico Kolter
- abstract: Strong adversarial attacks are important for evaluating the true robustness of deep neural networks. Most existing attacks search in the input space, e.g., using gradient descent, and may miss adversarial examples due to non-convexity. In this work, we systematically search adversarial examples in the activation space of ReLU networks to tackle hard instances where none of the existing adversarial attacks succeed. Unfortunately, searching the activation space typically relies on generic mixed integer programming (MIP) solvers and is limited to small networks and easy problem instances. To improve scalability and practicability, we use branch and bound (BaB) with specialized GPU-based bound propagation methods, and propose a top-down beam-search approach to quickly identify the subspace that may contain adversarial examples. Moreover, we build an adversarial candidates pool using cheap attacks to further assist the search in activation space via diving techniques and a bottom-up large neighborhood search. Our adversarial attack framework, BaB-Attack, opens up a new opportunity for designing novel adversarial attacks not limited to searching the input space, and enables us to borrow techniques from integer programming theory and neural network verification. In experiments, we can successfully generate adversarial examples when existing attacks on input space fail. Compared to off-the-shelf MIP solver based attacks that requires significant computations, we outperform in both success rates and efficiency.

A Simple yet Universal Strategy for Online Convex Optimization

- Lijun Zhang, Guanghui Wang, Jinfeng Yi, Tianbao Yang
- abstract: Recently, several universal methods have been proposed for online convex optimization, and attain minimax rates for multiple types of convex functions simultaneously. However, they need to design and optimize one surrogate loss for each type of functions, making it difficult to exploit the structure of the problem and utilize existing algorithms. In this paper, we propose a simple strategy for universal online convex optimization, which avoids these limitations. The key idea is to construct a set of experts to process the original online functions, and deploy a meta-algorithm over the linearized losses to aggregate predictions from experts. Specifically, the meta-algorithm is required to yield a second-order bound with excess losses, so that it can leverage strong convexity and exponential concavity to control the meta-regret. In this way, our strategy inherits the theoretical guarantee of any expert designed for strongly convex functions and exponentially concave functions, up to a double logarithmic factor. As a result, we can plug in off-the-shelf online solvers as black-box experts to deliver problem-dependent regret bounds. For general convex functions, it maintains the minimax optimality and also achieves a small-loss bound.

Low-Precision Stochastic Gradient Langevin Dynamics

- Ruqi Zhang, Andrew Gordon Wilson, Christopher De Sa
- abstract: While low-precision optimization has been widely used to accelerate deep learning, low-precision sampling remains largely unexplored. As a consequence, sampling is simply infeasible in many large-scale scenarios, despite providing remarkable benefits to generalization and uncertainty estimation for neural networks. In this paper, we provide the first study of low-precision Stochastic Gradient Langevin Dynamics (SGLD), showing that its costs can be significantly reduced without sacrificing performance, due to its intrinsic ability to handle system noise. We prove that the convergence of low-precision SGLD with full-precision gradient accumulators is less affected by the quantization error than its SGD counterpart in the strongly convex setting. To further enable low-precision gradient accumulators, we develop a new quantization function for SGLD that preserves the variance in each update step. We demonstrate that low-precision SGLD achieves comparable performance to full-precision SGLD with only 8 bits on a variety of deep learning tasks.

Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control

- Liang Zhang, Qiang Wu, Jun Shen, Linyuan Lü, Bo Du, Jianqing Wu
- abstract: Many studies confirmed that a proper traffic state representation is more important than complex algorithms for the classical traffic signal control (TSC) problem. In this paper, we (1) present a novel, flexible and efficient method, namely advanced max pressure (Advanced-MP), taking both running and queuing vehicles into consideration to decide whether to change current signal phase; (2) inventively design the traffic movement representation with the efficient pressure and effective running vehicles from Advanced-MP, namely advanced traffic state (ATS); and (3) develop a reinforcement learning (RL) based algorithm template, called Advanced-XLight, by combining ATS with the latest RL approaches, and generate two RL algorithms, namely "Advanced-MPLight" and "Advanced-CoLight" from Advanced-XLight. Comprehensive experiments on multiple real-world datasets show that: (1) the Advanced-MP outperforms baseline methods, and it is also efficient and reliable for deployment; and (2) Advanced-MPLight and Advanced-CoLight can achieve the state-of-the-art.

Uncertainty Modeling in Generative Compressed Sensing

- Yilang Zhang, Mengchu Xu, Xiaojun Mao, Jian Wang
- abstract: Compressed sensing (CS) aims to recover a high-dimensional signal with structural priors from its low-dimensional linear measurements. Inspired by the huge success of deep neural networks in modeling the priors of natural signals, generative neural networks have been recently used to replace the hand-crafted structural priors in CS. However, the reconstruction capability of the generative model is fundamentally limited by the range of its generator, typically a small subset of the signal space of interest. To break this bottleneck and thus reconstruct those out-of-range signals, this paper presents a novel method called CS-BGM that can effectively expands the range of generator. Specifically, CS-BGM introduces uncertainties to the latent variable and parameters of the generator, while adopting the variational inference (VI) and maximum a posteriori (MAP) to infer them. Theoretical analysis demonstrates that expanding the range of generators is necessary for reducing the reconstruction error in generative CS. Extensive experiments show a consistent improvement of CS-BGM over the baselines.

Building Robust Ensembles via Margin Boosting

- Dinghuai Zhang, Hongyang Zhang, Aaron Courville, Yoshua Bengio, Pradeep Ravikumar, Arun Sai Suggala
- abstract: In the context of adversarial robustness, a single model does not usually have enough power to defend against all possible adversarial attacks, and as a result, has sub-optimal robustness. Consequently, an emerging line of work has focused on learning an ensemble of neural networks to defend against adversarial attacks. In this work, we take a principled approach towards building robust ensembles. We view this problem from the perspective of margin-boosting and develop an algorithm for learning an ensemble with maximum margin. Through extensive empirical evaluation on benchmark datasets, we show that our algorithm not only outperforms existing ensembling techniques, but also large models trained in an end-to-end fashion. An important byproduct of our work is a margin-maximizing cross-entropy (MCE) loss, which is a better alternative to the standard cross-entropy (CE) loss. Empirically, we show that replacing the CE loss in state-of-the-art adversarial training techniques with our MCE loss leads to significant performance improvement.

Revisiting and Advancing Fast Adversarial Training Through The Lens of Bi-Level Optimization

- Yihua Zhang, Guanhua Zhang, Prashant Khanduri, Mingyi Hong, Shiyu Chang, Sijia Liu
- abstract: Adversarial training (AT) is a widely recognized defense mechanism to gain the robustness of deep neural networks against adversarial attacks. It is built on min-max optimization (MMO), where the minimizer (i.e., defender) seeks a robust model to minimize the worst-case training loss in the presence of adversarial examples crafted by the maximizer (i.e., attacker). However, the conventional MMO method makes AT hard to scale. Thus, Fast-AT and other recent algorithms attempt to simplify MMO by replacing its maximization step with the single gradient sign-based attack generation step. Although easy to implement, FAST-AT lacks theoretical guarantees, and its empirical performance is unsatisfactory due to the issue of robust catastrophic overfitting when training with strong adversaries. In this paper, we advance Fast-AT from the fresh perspective of bi-level optimization (BLO). We first show that the commonly-used Fast-AT is equivalent to using a stochastic gradient algorithm to solve a linearized BLO problem involving a sign operation. However, the discrete nature of the sign operation makes it difficult to understand the algorithm performance. Inspired by BLO, we design and analyze a new set of robust training algorithms termed Fast Bi-level AT (Fast-BAT), which effectively defends sign-based projected gradient descent (PGD) attacks without using any gradient sign method or explicit robust regularization. In practice, we show that our method yields substantial robustness improvements over multiple baselines across multiple models and datasets.

Off-Policy Fitted Q-Evaluation with Differentiable Function Approximators: Z-Estimation and Inference Theory

- Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, Mengdi Wang
- abstract: Off-Policy Evaluation (OPE) serves as one of the cornerstones in Reinforcement Learning (RL). Fitted Q Evaluation (FQE) with various function approximators, especially deep neural networks, has gained practical success. While statistical analysis has proved FQE to be minimax-optimal with tabular, linear and several nonparametric function families, its practical performance with more general function approximator is less theoretically understood. We focus on FQE with general differentiable function approximators, making our theory applicable to neural function approximations. We approach this problem using the Z-estimation theory and establish the following results: The FQE estimation error is asymptotically normal with explicit variance determined jointly by the tangent space of the function class at the ground truth, the reward structure, and the distribution shift due to off-policy learning; The finite-sample FQE error bound is dominated by the same variance term, and it can also be bounded by function class-dependent divergence, which measures how the off-policy distribution shift intertwines with the function approximator. In addition, we study bootstrapping FQE estimators for error distribution inference and estimating confidence intervals, accompanied by a Cramer-Rao lower bound that matches our upper bounds. The Z-estimation analysis provides a generalizable theoretical framework for studying off-policy estimation in RL and provides sharp statistical theory for FQE with differentiable function approximators.

ROCK: Causal Inference Principles for Reasoning about Commonsense Causality

- Jiayao Zhang, Hongming Zhang, Weijie Su, Dan Roth
- abstract: Commonsense causality reasoning (CCR) aims at identifying plausible causes and effects in natural language descriptions that are deemed reasonable by an average person. Although being of great academic and practical interest, this problem is still shadowed by the lack of a well-posed theoretical framework; existing work usually relies on deep language models wholeheartedly, and is potentially susceptible to confounding co-

occurrences. Motivated by classical causal principles, we articulate the central question of CCR and draw parallels between human subjects in observational studies and natural languages to adopt CCR to the potential-outcomes framework, which is the first such attempt for commonsense tasks. We propose a novel framework, ROCK, to Reason O(A)bout Commonsense K(C)ausality, which utilizes temporal signals as incidental supervision, and balances confounding effects using temporal propensities that are analogous to propensity scores. The ROCK implementation is modular and zero-shot, and demonstrates good CCR capabilities.

No-Regret Learning in Time-Varying Zero-Sum Games

- Mengxiao Zhang, Peng Zhao, Haipeng Luo, Zhi-Hua Zhou
- abstract: Learning from repeated play in a fixed two-player zero-sum game is a classic problem in game theory and online learning. We consider a variant of this problem where the game payoff matrix changes over time, possibly in an adversarial manner. We first present three performance measures to guide the algorithmic design for this problem: 1) the well-studied individual regret, 2) an extension of duality gap, and 3) a new measure called dynamic Nash Equilibrium regret, which quantifies the cumulative difference between the player's payoff and the minimax game value. Next, we develop a single parameter-free algorithm that simultaneously enjoys favorable guarantees under all these three performance measures. These guarantees are adaptive to different non-stationarity measures of the payoff matrices and, importantly, recover the best known results when the payoff matrix is fixed. Our algorithm is based on a two-layer structure with a meta-algorithm learning over a group of black-box base-learners satisfying a certain property, along with several novel ingredients specifically designed for the time-varying game setting. Empirical results further validate the effectiveness of our algorithm.

PLATON: Pruning Large Transformer Models with Upper Confidence Bound of Weight Importance

- Qingru Zhang, Simiao Zuo, Chen Liang, Alexander Bukharin, Pengcheng He, Weizhu Chen, Tuo Zhao
- abstract: Large Transformer-based models have exhibited superior performance in various natural language processing and computer vision tasks. However, these models contain enormous amounts of parameters, which restrict their deployment to real-world applications. To reduce the model size, researchers prune these models based on the weights' importance scores. However, such scores are usually estimated on mini-batches during training, which incurs large variability/uncertainty due to mini-batch sampling and complicated training dynamics. As a result, some crucial weights could be pruned by commonly used pruning methods because of such uncertainty, which makes training unstable and hurts generalization. To resolve this issue, we propose PLATON, which captures the uncertainty of importance scores by upper confidence bound of importance estimation. In particular, for the weights with low importance scores but high uncertainty, PLATON tends to retain them and explores their capacity. We conduct extensive experiments with several Transformer-based models on natural language understanding, question answering and image classification to validate the effectiveness of PLATON. Results demonstrate that PLATON manifests notable improvement under different sparsity levels. Our code is publicly available at https://github.com/QingruZhang/PLATON.

NysADMM: faster composite convex optimization via low-rank approximation

- Shipu Zhao, Zachary Frangella, Madeleine Udell
- abstract: This paper develops a scalable new algorithm, called NysADMM, to minimize a smooth convex loss function with a convex regularizer. NysADMM accelerates the inexact Alternating Direction Method of Multipliers (ADMM) by constructing a preconditioner for the ADMM subproblem from a randomized low-rank Nyström approximation. NysADMM comes with strong theoretical guarantees: it solves the ADMM subproblem in a constant number of iterations when the rank of the Nyström approximation is the effective dimension of the subproblem regularized Gram matrix. In practice, ranks much smaller than the effective dimension can succeed, so NysADMM uses an adaptive strategy to choose the rank that enjoys analogous guarantees. Numerical experiments on real-world datasets demonstrate that NysADMM can solve important applications, such as the lasso, logistic regression, and support vector machines, in half the time (or less) required by standard solvers. The breadth of problems on which NysADMM beats standard solvers is a surprise: it suggests that ADMM is a dominant paradigm for numerical optimization across a wide range of statistical learning problems that are usually solved with bespoke methods.

Toward Compositional Generalization in Object-Oriented World Modeling

- Linfeng Zhao, Lingzhi Kong, Robin Walters, Lawson L.S. Wong
- abstract: Compositional generalization is a critical ability in learning and decision-making. We focus on the setting of reinforcement learning in object-oriented environments to study compositional generalization in world modeling. We (1) formalize the compositional generalization problem with an algebraic approach and (2) study how a world model can achieve that. We introduce a conceptual environment, Object Library, and two instances, and deploy a principled pipeline to measure the generalization ability. Motivated by the formulation, we analyze several methods with exact or no compositional generalization ability using our framework, and design a differentiable approach, Homomorphic Object-oriented World Model (HOWM), that achieves soft but more efficient compositional generalization.

Dynamic Regret of Online Markov Decision Processes

- Peng Zhao, Long-Fei Li, Zhi-Hua Zhou
- abstract: We investigate online Markov Decision Processes (MDPs) with adversarially changing loss functions and known transitions. We choose dynamic regret as the performance measure, defined as the performance difference between the learner and any sequence of feasible changing policies. The measure is strictly stronger than the standard static regret that benchmarks the learner's performance with a fixed compared policy. We consider three foundational models of online MDPs, including episodic loop-free Stochastic Shortest Path (SSP), episodic SSP, and infinite-horizon MDPs. For the three models, we propose novel online ensemble algorithms and establish their dynamic regret guarantees respectively, in which the results for episodic (loop-free) SSP are provably minimax optimal in terms of time horizon and certain non-stationarity measure.

Learning to Solve PDE-constrained Inverse Problems with Graph Networks

- Qingqing Zhao, David B Lindell, Gordon Wetzstein
- abstract: Learned graph neural networks (GNNs) have recently been established as fast and accurate alternatives for principled solvers in simulating the dynamics of physical systems. In many application domains across science and engineering, however, we are not only interested in a forward simulation but also in solving inverse problems with constraints defined by a partial differential equation (PDE). Here we explore GNNs to solve such PDE-constrained inverse problems. Given a sparse set of measurements, we are interested in recovering the initial condition or parameters of the PDE. We demonstrate that GNNs combined with autodecoder-style priors are well-suited for these tasks, achieving more accurate estimates of initial conditions or physical parameters than other learned approaches when applied to the wave equation or Navier Stokes equations. We also demonstrate computational speedups of up to 90x using GNNs compared to principled solvers.

Learning from Counterfactual Links for Link Prediction

- Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, Meng Jiang
- abstract: Learning to predict missing links is important for many graph-based applications. Existing methods were designed to learn the association between observed graph structure and existence of link between a pair of nodes. However, the causal relationship between the two variables was largely

ignored for learning to predict links on a graph. In this work, we visit this factor by asking a counterfactual question: "would the link still exist if the graph structure became different from observation?" Its answer, counterfactual links, will be able to augment the graph data for representation learning. To create these links, we employ causal models that consider the information (i.e., learned representations) of node pairs as context, global graph structural properties as treatment, and link existence as outcome. We propose a novel data augmentation-based link prediction method that creates counterfactual links and learns representations from both the observed and counterfactual links. Experiments on benchmark data show that our graph learning method achieves state-of-the-art performance on the task of link prediction.

Global Optimization Networks

- Sen Zhao, Erez Louidor, Maya Gupta
- abstract: We consider the problem of estimating a good maximizer of a black-box function given noisy examples. We propose to fit a new type of function called a global optimization network (GON), defined as any composition of an invertible function and a unimodal function, whose unique global maximizer can be inferred in $\mathcal{O}(D)$ time, and used as the estimate. As an example way to construct GON functions, and interesting in its own right, we give new results for specifying multi-dimensional unimodal functions using lattice models with linear inequality constraints. We extend to conditional GONs that find a global maximizer conditioned on specified inputs of other dimensions. Experiments show the GON maximizers are statistically significantly better predictions than those produced by convex fits, GPR, or DNNs, and form more reasonable predictions for real-world problems.

Certified Robustness Against Natural Language Attacks by Causal Intervention

- Haiteng Zhao, Chang Ma, Xinshuai Dong, Anh Tuan Luu, Zhi-Hong Deng, Hanwang Zhang
- abstract: Deep learning models have achieved great success in many fields, yet they are vulnerable to adversarial examples. This paper follows a causal perspective to look into the adversarial vulnerability and proposes Causal Intervention by Semantic Smoothing (CISS), a novel framework towards robustness against natural language attacks. Instead of merely fitting observational data, CISS learns causal effects p(y|do(x)) by smoothing in the latent semantic space to make robust predictions, which scales to deep architectures and avoids tedious construction of noise customized for specific attacks. CISS is provably robust against word substitution attacks, as well as empirically robust even when perturbations are strengthened by unknown attack algorithms. For example, on YELP, CISS surpasses the runner-up by 6.8% in terms of certified robustness against word substitutions, and achieves 80.7% empirical robustness when syntactic attacks are integrated.

Efficient Learning for AlphaZero via Path Consistency

- Dengwei Zhao, Shikui Tu, Lei Xu
- abstract: In recent years, deep reinforcement learning have made great breakthroughs on board games. Still, most of the works require huge computational resources for a large scale of environmental interactions or self-play for the games. This paper aims at building powerful models under a limited amount of self-plays which can be utilized by a human throughout the lifetime. We proposes a learning algorithm built on AlphaZero, with its path searching regularised by a path consistency (PC) optimality, i.e., values on one optimal search path should be identical. Thus, the algorithm is shortly named PCZero. In implementation, historical trajectory and scouted search paths by MCTS makes a good balance between exploration and exploitation, which enhances the generalization ability effectively. PCZero obtains $94.1%$ winning rate against the champion of Hex Computer Olympiad in 2015 on 13×13 Hex, much higher than $84.3%$ by AlphaZero. The models consume only $900K$ self-play games, about the amount humans can study in a lifetime. The improvements by PCZero have been also generalized to Othello and Gomoku. Experiments also demonstrate the efficiency of PCZero under offline learning setting.

Penalizing Gradient Norm for Efficiently Improving Generalization in Deep Learning

- Yang Zhao, Hao Zhang, Xiuyuan Hu
- abstract: How to train deep neural networks (DNNs) to generalize well is a central concern in deep learning, especially for severely overparameterized networks nowadays. In this paper, we propose an effective method to improve the model generalization by additionally penalizing the gradient norm of loss function during optimization. We demonstrate that confining the gradient norm of loss function could help lead the optimizers towards finding flat minima. We leverage the first-order approximation to efficiently implement the corresponding gradient to fit well in the gradient descent framework. In our experiments, we confirm that when using our methods, generalization performance of various models could be improved on different datasets. Also, we show that the recent sharpness-aware minimization method (Foretet al., 2021) is a special, but not the best, case of our method, where the best case of our method could give new state-of-art performance on these tasks. Code is available at https://github.com/zhaoyang-0204/gnp.

Ripple Attention for Visual Perception with Sub-quadratic Complexity

- Lin Zheng, Huijie Pan, Lingpeng Kong
- abstract: Transformer architectures are now central to sequence modeling tasks. At its heart is the attention mechanism, which enables effective modeling of long-term dependencies in a sequence. Recently, transformers have been successfully applied in the computer vision domain, where 2D images are first segmented into patches and then treated as 1D sequences. Such linearization, however, impairs the notion of spatial locality in images, which bears important visual clues. To bridge the gap, we propose ripple attention, a sub-quadratic attention mechanism for vision transformers. Built upon the recent kernel-based efficient attention mechanisms, we design a novel dynamic programming algorithm that weights contributions of different tokens to a query with respect to their relative spatial distances in the 2D space in linear observed time. Extensive experiments and analyses demonstrate the effectiveness of ripple attention on various visual tasks.

Linear Complexity Randomized Self-attention Mechanism

- Lin Zheng, Chong Wang, Lingpeng Kong
- abstract: Recently, random feature attentions (RFAs) are proposed to approximate the softmax attention in linear time and space complexity by linearizing the exponential kernel. In this paper, we first propose a novel perspective to understand the bias in such approximation by recasting RFAs as self-normalized importance samplers. This perspective further sheds light on an unbiased estimator for the whole softmax attention, called randomized attention (RA). RA constructs positive random features via query-specific distributions and enjoys greatly improved approximation fidelity, albeit exhibiting quadratic complexity. By combining the expressiveness in RA and the efficiency in RFA, we develop a novel linear complexity self-attention mechanism called linear randomized attention (LARA). Extensive experiments across various domains demonstrate that RA and LARA significantly improve the performance of RFAs by a substantial margin.

Online Decision Transformer

- Qinqing Zheng, Amy Zhang, Aditya Grover
- abstract: Recent work has shown that offline reinforcement learning (RL) can be formulated as a sequence modeling problem (Chen et al., 2021; Janner et al., 2021) and solved via approaches similar to large-scale language modeling. However, any practical instantiation of RL also involves an online component, where policies pretrained on passive offline datasets are finetuned via task-specific interactions with the environment. We propose Online

Decision Transformers (ODT), an RL algorithm based on sequence modeling that blends offline pretraining with online finetuning in a unified framework. Our framework uses sequence-level entropy regularizers in conjunction with autoregressive modeling objectives for sample-efficient exploration and finetuning. Empirically, we show that ODT is competitive with the state-of-the-art in absolute performance on the D4RL benchmark but shows much more significant gains during the finetuning procedure.

Learning Efficient and Robust Ordinary Differential Equations via Invertible Neural Networks

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

HyperTransformer: Model Generation for Supervised and Semi-Supervised Few-Shot Learning

- Andrey Zhmoginov, Mark Sandler, Maksym Vladymyrov
- abstract: In this work we propose a HyperTransformer, a Transformer-based model for supervised and semi-supervised few-shot learning that generates weights of a convolutional neural network (CNN) directly from support samples. Since the dependence of a small generated CNN model on a specific task is encoded by a high-capacity Transformer model, we effectively decouple the complexity of the large task space from the complexity of individual tasks. Our method is particularly effective for small target CNN architectures where learning a fixed universal task-independent embedding is not optimal and better performance is attained when the information about the task can modulate all model parameters. For larger models we discover that generating the last layer alone allows us to produce competitive or better results than those obtained with state-of-the-art methods while being end-to-end differentiable.

Describing Differences between Text Distributions with Natural Language

- Ruiqi Zhong, Charlie Snell, Dan Klein, Jacob Steinhardt
- abstract: How do two distributions of text differ? Humans are slow at answering this, since discovering patterns might require tediously reading through hundreds of samples. We propose to automatically summarize the differences by “learning a natural language hypothesis": given two distributions D_{0} and D_{1}, we search for a description that is more often true for D_{1}, e.g., “is military-related." To tackle this problem, we fine-tune GPT-3 to propose descriptions with the prompt: “[samples of D_{0}] + [samples of D_{1}] + the difference between them is \underline{\space\space\space\space}". We then re-rank the descriptions by checking how often they hold on a larger set of samples with a learned verifier. On a benchmark of 54 real-world binary classification tasks, while GPT-3 Curie (13B) only generates a description similar to human annotation 7% of the time, the performance reaches 61% with fine-tuning and re-ranking, and our best system using GPT-3 Davinci (175B) reaches 76%. We apply our system to describe distribution shifts, debug dataset shortcuts, summarize unknown tasks, and label text clusters, and present analyses based on automatically generated descriptions.

Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets

- Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, Zhuoran Yang
- abstract: We study episodic two-player zero-sum Markov games (MGs) in the offline setting, where the goal is to find an approximate Nash equilibrium (NE) policy pair based on a dataset collected a priori. When the dataset does not have uniform coverage over all policy pairs, finding an approximate NE involves challenges in three aspects: (i) distributional shift between the behavior policy and the optimal policy, (ii) function approximation to handle large state space, and (iii) minimax optimization for equilibrium solving. We propose a pessimism-based algorithm, dubbed as pessimistic minimax value iteration (PMVI), which overcomes the distributional shift by constructing pessimistic estimates of the value functions for both players and outputs a policy pair by solving a correlated coarse equilibrium based on the two value functions. Furthermore, we establish a data-dependent upper bound on the suboptimality which recovers a sublinear rate without the assumption on uniform coverage of the dataset. We also prove an information-theoretical lower bound, which shows our upper bound is nearly minimax optimal, which suggests that the data-dependent term is intrinsic. Our theoretical results also highlight a notion of “relative uncertainty”, which characterizes the necessary and sufficient condition for achieving sample efficiency in offline MGs. To the best of our knowledge, we provide the first nearly minimax optimal result for offline MGs with function approximation.

Dimension-free Complexity Bounds for High-order Nonconvex Finite-sum Optimization

- Dongruo Zhou, Quanquan Gu
- abstract: Stochastic high-order methods for finding first-order stationary points in nonconvex finite-sum optimization have witnessed increasing interest in recent years, and various upper and lower bounds of the oracle complexity have been proved. However, under standard regularity assumptions, existing complexity bounds are all dimension-dependent (e.g., polylogarithmic dependence), which contrasts with the dimension-free complexity bounds for stochastic first-order methods and deterministic high-order methods. In this paper, we show that the polylogarithmic dimension dependence gap is not essential and can be closed. More specifically, we propose stochastic high-order algorithms with novel first-order and high-order derivative estimators, which can achieve dimension-free complexity bounds. With the access to p-th order derivatives of the objective function, we prove that our algorithm finds ϵ-stationary points with $O(n^{(2p-1)/(2p)}/\epsilon^{(p+1)/p})$ high-order oracle complexities, where n is the number of individual functions. Our result strictly improves the complexity bounds of existing high-order deterministic methods with respect to the dependence on n, and it is dimension-free compared with existing stochastic high-order methods.

A Hierarchical Bayesian Approach to Inverse Reinforcement Learning with Symbolic Reward Machines

- Weichao Zhou, Wenchao Li
- abstract: A misspecified reward can degrade sample efficiency and induce undesired behaviors in reinforcement learning (RL) problems. We propose symbolic reward machines for incorporating high-level task knowledge when specifying the reward signals. Symbolic reward machines augment existing reward machine formalism by allowing transitions to carry predicates and symbolic reward outputs. This formalism lends itself well to inverse reinforcement learning, whereby the key challenge is determining appropriate assignments to the symbolic values from a few expert demonstrations. We propose a hierarchical Bayesian approach for inferring the most likely assignments such that the concretized reward machine can discriminate expert demonstrated trajectories from other trajectories with high accuracy. Experimental results show that learned reward machines can significantly improve training efficiency for complex RL tasks and generalize well across different task environment configurations.

On the Optimization Landscape of Neural Collapse under MSE Loss: Global Optimality with Unconstrained Features

- Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, Zhihui Zhu
- abstract: When training deep neural networks for classification tasks, an intriguing empirical phenomenon has been widely observed in the last-layer classifiers and features, where (i) the class means and the last-layer classifiers all collapse to the vertices of a Simplex Equiangular Tight Frame (ETF) up to scaling, and (ii) cross-example within-class variability of last-layer activations collapses to zero. This phenomenon is called Neural Collapse (NC), which seems to take place regardless of the choice of loss functions. In this work, we justify NC under the mean squared error (MSE) loss, where recent empirical evidence shows that it performs comparably or even better than the de-facto cross-entropy loss. Under a simplified unconstrained feature model, we provide the first global landscape analysis for vanilla nonconvex MSE loss and show that the (only!) global minimizers are neural collapse solutions, while all other critical points are strict saddles whose Hessian exhibit negative curvature directions. Furthermore, we justify the usage of rescaled MSE loss by probing the optimization landscape around the NC solutions, showing that the landscape can be improved by tuning the rescaling hyperparameters. Finally, our theoretical findings are experimentally verified on practical network architectures.

Model Agnostic Sample Reweighting for Out-of-Distribution Learning

- Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, Tong Zhang
- abstract: Distributionally robust optimization (DRO) and invariant risk minimization (IRM) are two popular methods proposed to improve out-of-distribution (OOD) generalization performance of machine learning models. While effective for small models, it has been observed that these methods can be vulnerable to overfitting with large overparameterized models. This work proposes a principled method, Model Agnostic samPLe rEweighting (MAPLE), to effectively address OOD problem, especially in overparameterized scenarios. Our key idea is to find an effective reweighting of the training samples so that the standard empirical risk minimization training of a large model on the weighted training data leads to superior OOD generalization performance. The overfitting issue is addressed by considering a bilevel formulation to search for the sample reweighting, in which the generalization complexity depends on the search space of sample weights instead of the model size. We present theoretical analysis in linear case to prove the insensitivity of MAPLE to model size, and empirically verify its superiority in surpassing state-of-the-art methods by a large margin.

Sparse Invariant Risk Minimization

- Xiao Zhou, Yong Lin, Weizhong Zhang, Tong Zhang
- abstract: Invariant Risk Minimization (IRM) is an emerging invariant feature extracting technique to help generalization with distributional shift. However, we find that there exists a basic and intractable contradiction between the model trainability and generalization ability in IRM. On one hand, recent studies on deep learning theory indicate the importance of large-sized or even overparameterized neural networks to make the model easy to train. On the other hand, unlike empirical risk minimization that can be benefited from overparameterization, our empirical and theoretical analyses show that the generalization ability of IRM is much easier to be demolished by overfitting caused by overparameterization. In this paper, we propose a simple yet effective paradigm named Sparse Invariant Risk Minimization (SparseIRM) to address this contradiction. Our key idea is to employ a global sparsity constraint as a defense to prevent spurious features from leaking in during the whole IRM process. Compared with sparisfy-after-training prototype by prior work which can discard invariant features, the global sparsity constraint limits the budget for feature selection and enforces SparseIRM to select the invariant features. We illustrate the benefit of SparseIRM through a theoretical analysis on a simple linear case. Empirically we demonstrate the power of SparseIRM through various datasets and models and surpass state-of-the-art methods with a gap up to 29%.

Prototype-Anchored Learning for Learning with Imperfect Annotations

- Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, Xiangyang Ji
- abstract: The success of deep neural networks greatly relies on the availability of large amounts of high-quality annotated data, which however are difficult or expensive to obtain. The resulting labels may be class imbalanced, noisy or human biased. It is challenging to learn unbiased classification models from imperfectly annotated datasets, on which we usually suffer from overfitting or underfitting. In this work, we thoroughly investigate the popular softmax loss and margin-based loss, and offer a feasible approach to tighten the generalization error bound by maximizing the minimal sample margin. We further derive the optimality condition for this purpose, which indicates how the class prototypes should be anchored. Motivated by theoretical analysis, we propose a simple yet effective method, namely prototype-anchored learning (PAL), which can be easily incorporated into various learning-based classification schemes to handle imperfect annotation. We verify the effectiveness of PAL on class-imbalanced learning and noise-tolerant learning by extensive experiments on synthetic and real-world datasets.

FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting

- Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, Rong Jin
- abstract: Long-term time series forecasting is challenging since prediction accuracy tends to decrease dramatically with the increasing horizon. Although Transformer-based methods have significantly improved state-of-the-art results for long-term forecasting, they are not only computationally expensive but more importantly, are unable to capture the global view of time series (e.g. overall trend). To address these problems, we propose to combine Transformer with the seasonal-trend decomposition method, in which the decomposition method captures the global profile of time series while Transformers capture more detailed structures. To further enhance the performance of Transformer for long-term prediction, we exploit the fact that most time series tend to have a sparse representation in a well-known basis such as Fourier transform, and develop a frequency enhanced Transformer. Besides being more effective, the proposed method, termed as Frequency Enhanced Decomposed Transformer (FEDformer), is more efficient than standard Transformer with a linear complexity to the sequence length. Our empirical studies with six benchmark datasets show that compared with state-of-the-art methods, Fedformer can reduce prediction error by 14.8% and 22.6% for multivariate and univariate time series, respectively. Code is publicly available at https://github.com/MAZiqing/FEDformer.

Probabilistic Bilevel Coreset Selection

- Xiao Zhou, Renjie Pi, Weizhong Zhang, Yong Lin, Zonghao Chen, Tong Zhang
- abstract: The goal of coreset selection in supervised learning is to produce a weighted subset of data, so that training only on the subset achieves similar performance as training on the entire dataset. Existing methods achieved promising results in resource-constrained scenarios such as continual learning and streaming. However, most of the existing algorithms are limited to traditional machine learning models. A few algorithms that can handle large models adopt greedy search approaches due to the difficulty in solving the discrete subset selection problem, which is computationally costly when coreset becomes larger and often produces suboptimal results. In this work, for the first time we propose a continuous probabilistic bilevel formulation of coreset selection by learning a probablistic weight for each training sample. The overall objective is posed as a bilevel optimization problem, where 1) the inner loop samples coresets and train the model to convergence and 2) the outer loop updates the sample probability progressively according to the model's performance. Importantly, we develop an efficient solver to the bilevel optimization problem via unbiased policy gradient without trouble of implicit differentiation. We theoretically prove the convergence of this training procedure and demonstrate the superiority of our algorithm against various coreset selection methods in various tasks, especially in more challenging label-noise and class-imbalance scenarios.

Approximate Frank-Wolfe Algorithms over Graph-structured Support Sets

- Baojian Zhou, Yifan Sun
- abstract: In this paper, we consider approximate Frank-Wolfe (FW) algorithms to solve convex optimization problems over graph-structured support sets where the linear minimization oracle (LMO) cannot be efficiently obtained in general. We first demonstrate that two popular approximation assumptions

(additive and multiplicative gap errors) are not applicable in that no cheap gap-approximate LMO oracle exists. Thus, approximate dual maximization oracles (DMO) are proposed, which approximate the inner product rather than the gap. We prove that the standard FW method using a δ-approximate DMO converges as $O((1-\delta) \sqrt{s}/\delta)$ in the worst case, and as $O(L/(\delta^2 t))$ over a δ-relaxation of the constraint set. Furthermore, when the solution is on the boundary, a variant of FW converges as $O(1/t^2)$ under the quadratic growth assumption. Our empirical results suggest that even these improved bounds are pessimistic, showing fast convergence in recovering real-world images with graph-structured sparsity.

Improving Adversarial Robustness via Mutual Information Estimation

- Dawei Zhou, Nannan Wang, Xinbo Gao, Bo Han, Xiaoyu Wang, Yibing Zhan, Tongliang Liu
- abstract: Deep neural networks (DNNs) are found to be vulnerable to adversarial noise. They are typically misled by adversarial samples to make wrong predictions. To alleviate this negative effect, in this paper, we investigate the dependence between outputs of the target model and input adversarial samples from the perspective of information theory, and propose an adversarial defense method. Specifically, we first measure the dependence by estimating the mutual information (MI) between outputs and the natural patterns of inputs (called natural MI) and MI between outputs and the adversarial patterns of inputs (called adversarial MI), respectively. We find that adversarial samples usually have larger adversarial MI and smaller natural MI compared with those w.r.t. natural samples. Motivated by this observation, we propose to enhance the adversarial robustness by maximizing the natural MI and minimizing the adversarial MI during the training process. In this way, the target model is expected to pay more attention to the natural pattern that contains objective semantics. Empirical evaluations demonstrate that our method could effectively improve the adversarial accuracy against multiple attacks.

Modeling Adversarial Noise for Adversarial Training

- Dawei Zhou, Nannan Wang, Bo Han, Tongliang Liu
- abstract: Deep neural networks have been demonstrated to be vulnerable to adversarial noise, promoting the development of defense against adversarial attacks. Motivated by the fact that adversarial noise contains well-generalizing features and that the relationship between adversarial data and natural data can help infer natural data and make reliable predictions, in this paper, we study to model adversarial noise by learning the transition relationship between adversarial labels (i.e. the flipped labels used to generate adversarial data) and natural labels (i.e. the ground truth labels of the natural data). Specifically, we introduce an instance-dependent transition matrix to relate adversarial labels and natural labels, which can be seamlessly embedded with the target model (enabling us to model stronger adaptive adversarial noise). Empirical evaluations demonstrate that our method could effectively improve adversarial accuracy.

Contrastive Learning with Boosted Memorization

- Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, Ya Zhang
- abstract: Self-supervised learning has achieved a great success in the representation learning of visual and textual data. However, the current methods are mainly validated on the well-curated datasets, which do not exhibit the real-world long-tailed distribution. Recent attempts to consider self-supervised long-tailed learning are made by rebalancing in the loss perspective or the model perspective, resembling the paradigms in the supervised long-tailed learning. Nevertheless, without the aid of labels, these explorations have not shown the expected significant promise due to the limitation in tail sample discovery or the heuristic structure design. Different from previous works, we explore this direction from an alternative perspective, i.e., the data perspective, and propose a novel Boosted Contrastive Learning (BCL) method. Specifically, BCL leverages the memorization effect of deep neural networks to automatically drive the information discrepancy of the sample views in contrastive learning, which is more efficient to enhance the long-tailed learning in the label-unaware context. Extensive experiments on a range of benchmark datasets demonstrate the effectiveness of BCL over several state-of-the-art methods. Our code is available at https://github.com/MediaBrain-SJTU/BCL.

Understanding The Robustness in Vision Transformers

- Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, Jose M. Alvarez
- abstract: Recent studies show that Vision Transformers (ViTs) exhibit strong robustness against various corruptions. Although this property is partly attributed to the self-attention mechanism, there is still a lack of an explanatory framework towards a more systematic understanding. In this paper, we examine the role of self-attention in learning robust representations. Our study is motivated by the intriguing properties of self-attention in visual grouping which indicate that self-attention could promote improved mid-level representation and robustness. We thus propose a family of fully attentional networks (FANs) that incorporate self-attention in both token mixing and channel processing. We validate the design comprehensively on various hierarchical backbones. Our model with a DeiT architecture achieves a state-of-the-art 47.6% mCE on ImageNet-C with 29M parameters. We also demonstrate significantly improved robustness in two downstream tasks: semantic segmentation and object detection

VLUE: A Multi-Task Multi-Dimension Benchmark for Evaluating Vision-Language Pre-training

- Wangchunshu Zhou, Yan Zeng, Shizhe Diao, Xinsong Zhang
- abstract: Recent advances in vision-language pre-training (VLP) have demonstrated impressive performance in a range of vision-language (VL) tasks. However, there exist several challenges for measuring the community's progress in building general multi-modal intelligence. First, most of the downstream VL datasets are annotated using raw images that are already seen during pre-training, which may result in an overestimation of current VLP models' generalization ability. Second, recent VLP work mainly focuses on absolute performance but overlooks the efficiency-performance trade-off, which is also an important indicator for measuring progress. To this end, we introduce the Vision-Language Understanding Evaluation (VLUE) benchmark, a multi-task multi-dimension benchmark for evaluating the generalization capabilities and the efficiency-performance trade-off ("Pareto SOTA") of VLP models. We demonstrate that there is a sizable generalization gap for all VLP models when testing on out-of-distribution test sets annotated on images from a more diverse distribution that spreads across cultures. Moreover, we find that measuring the efficiency-performance trade-off of VLP models leads to complementary insights for several design choices of VLP. We release the VLUE benchmark to promote research on building vision-language models that generalize well to images unseen during pre-training and are practical in terms of efficiency-performance trade-off.

Detecting Corrupted Labels Without Training a Model to Predict

- Zhaowei Zhu, Zihao Dong, Yang Liu
- abstract: Label noise in real-world datasets encodes wrong correlation patterns and impairs the generalization of deep neural networks (DNNs). It is critical to find efficient ways to detect corrupted patterns. Current methods primarily focus on designing robust training techniques to prevent DNNs from memorizing corrupted patterns. These approaches often require customized training processes and may overfit corrupted patterns, leading to a performance drop in detection. In this paper, from a more data-centric perspective, we propose a training-free solution to detect corrupted labels. Intuitively, "closer" instances are more likely to share the same clean label. Based on the neighborhood information, we propose two methods: the first one uses "local voting" via checking the noisy label consensuses of nearby features. The second one is a ranking-based approach that scores each instance and filters out a guaranteed number of instances that are likely to be corrupted. We theoretically analyze how the quality of features affects the local voting and provide guidelines for tuning neighborhood size. We also prove the worst-case error bound for the ranking-based method. Experiments with both synthetic and real-world label noise demonstrate our training-free solutions consistently and significantly improve most of the training-based baselines. Code is available at github.com/UCSC-REAL/SimiFeat.

Contextual Bandits with Large Action Spaces: Made Practical

- Yinglun Zhu, Dylan J Foster, John Langford, Paul Mineiro
- abstract: A central problem in sequential decision making is to develop algorithms that are practical and computationally efficient, yet support the use of flexible, general-purpose models. Focusing on the contextual bandit problem, recent progress provides provably efficient algorithms with strong empirical performance when the number of possible alternatives ("actions") is small, but guarantees for decision making in large, continuous action spaces have remained elusive, leading to a significant gap between theory and practice. We present the first efficient, general-purpose algorithm for contextual bandits with continuous, linearly structured action spaces. Our algorithm makes use of computational oracles for (i) supervised learning, and (ii) optimization over the action space, and achieves sample complexity, runtime, and memory independent of the size of the action space. In addition, it is simple and practical. We perform a large-scale empirical evaluation, and show that our approach typically enjoys superior performance and efficiency compared to standard baselines.

Neural-Symbolic Models for Logical Queries on Knowledge Graphs

- Zhaocheng Zhu, Mikhail Galkin, Zuobai Zhang, Jian Tang
- abstract: Answering complex first-order logic (FOL) queries on knowledge graphs is a fundamental task for multi-hop reasoning. Traditional symbolic methods traverse a complete knowledge graph to extract the answers, which provides good interpretation for each step. Recent neural methods learn geometric embeddings for complex queries. These methods can generalize to incomplete knowledge graphs, but their reasoning process is hard to interpret. In this paper, we propose Graph Neural Network Query Executor (GNN-QE), a neural-symbolic model that enjoys the advantages of both worlds. GNN-QE decomposes a complex FOL query into relation projections and logical operations over fuzzy sets, which provides interpretability for intermediate variables. To reason about the missing links, GNN-QE adapts a graph neural network from knowledge graph completion to execute the relation projections, and models the logical operations with product fuzzy logic. Experiments on 3 datasets show that GNN-QE significantly improves over previous state-of-the-art models in answering FOL queries. Meanwhile, GNN-QE can predict the number of answers without explicit supervision, and provide visualizations for intermediate variables.

Topology-aware Generalization of Decentralized SGD

- Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, Dacheng Tao
- abstract: This paper studies the algorithmic stability and generalizability of decentralized stochastic gradient descent (D-SGD). We prove that the consensus model learned by D-SGD is $\mathcal{O}{(m/N\unaryplus1/m\unaryplus\lambda^2)}$-stable in expectation in the non-convex non-smooth setting, where N is the total sample size of the whole system, m is the worker number, and $1\unaryminus\lambda$ is the spectral gap that measures the connectivity of the communication topology. These results then deliver an $\mathcal{O}{(1/N\unaryplus{({(m^{-1}\lambda^2)}^{\frac{\alpha}{2}}\unaryplus m^{\unaryminus\alpha})}/{N^{1\unaryminus\frac{\alpha}{2}}})}$ in-average generalization bound, which is non-vacuous even when λ is closed to 1, in contrast to vacuous as suggested by existing literature on the projected version of D-SGD. Our theory indicates that the generalizability of D-SGD has a positive correlation with the spectral gap, and can explain why consensus control in initial training phase can ensure better generalization. Experiments of VGG-11 and ResNet-18 on CIFAR-10, CIFAR-100 and Tiny-ImageNet justify our theory. To our best knowledge, this is the first work on the topology-aware generalization of vanilla D-SGD. Code is available at \url{https://github.com/Raiden-Zhu/Generalization-of-DSGD}.

Resilient and Communication Efficient Learning for Heterogeneous Federated Systems

- Zhuangdi Zhu, Junyuan Hong, Steve Drew, Jiayu Zhou
- abstract: The rise of Federated Learning (FL) is bringing machine learning to edge computing by utilizing data scattered across edge devices. However, the heterogeneity of edge network topologies and the uncertainty of wireless transmission are two major obstructions of FL's wide application in edge computing, leading to prohibitive convergence time and high communication cost. In this work, we propose an FL scheme to address both challenges simultaneously. Specifically, we enable edge devices to learn self-distilled neural networks that are readily prunable to arbitrary sizes, which capture the knowledge of the learning domain in a nested and progressive manner. Not only does our approach tackle system heterogeneity by serving edge devices with varying model architectures, but it also alleviates the issue of connection uncertainty by allowing transmitting part of the model parameters under faulty network connections, without wasting the contributing knowledge of the transmitted parameters. Extensive empirical studies show that under system heterogeneity and network instability, our approach demonstrates significant resilience and higher communication efficiency compared to the state-of-the-art.

On Numerical Integration in Neural Ordinary Differential Equations

- Aiqing Zhu, Pengzhan Jin, Beibei Zhu, Yifa Tang
- abstract: The combination of ordinary differential equations and neural networks, i.e., neural ordinary differential equations (Neural ODE), has been widely studied from various angles. However, deciphering the numerical integration in Neural ODE is still an open challenge, as many researches demonstrated that numerical integration significantly affects the performance of the model. In this paper, we propose the inverse modified differential equations (IMDE) to clarify the influence of numerical integration on training Neural ODE models. IMDE is determined by the learning task and the employed ODE solver. It is shown that training a Neural ODE model actually returns a close approximation of the IMDE, rather than the true ODE. With the help of IMDE, we deduce that (i) the discrepancy between the learned model and the true ODE is bounded by the sum of discretization error and learning loss; (ii) Neural ODE using non-symplectic numerical integration fail to learn conservation laws theoretically. Several experiments are performed to numerically verify our theoretical analysis.

When AUC meets DRO: Optimizing Partial AUC for Deep Learning with Non-Convex Convergence Guarantee

- Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, Tianbao Yang
- abstract: In this paper, we propose systematic and efficient gradient-based methods for both one-way and two-way partial AUC (pAUC) maximization that are applicable to deep learning. We propose new formulations of pAUC surrogate objectives by using the distributionally robust optimization (DRO) to define the loss for each individual positive data. We consider two formulations of DRO, one of which is based on conditional-value-at-risk (CVaR) that yields a non-smooth but exact estimator for pAUC, and another one is based on a KL divergence regularized DRO that yields an inexact but smooth (soft) estimator for pAUC. For both one-way and two-way pAUC maximization, we propose two algorithms and prove their convergence for optimizing their two formulations, respectively. Experiments demonstrate the effectiveness of the proposed algorithms for pAUC maximization for deep learning on various datasets.

Contextual Bandits with Smooth Regret: Efficient Learning in Continuous Action Spaces

- Yinglun Zhu, Paul Mineiro
- abstract: Designing efficient general-purpose contextual bandit algorithms that work with large—or even infinite—action spaces would facilitate application to important scenarios such as information retrieval, recommendation systems, and continuous control. While obtaining standard regret guarantees can be hopeless, alternative regret notions have been proposed to tackle the large action setting. We propose a smooth regret notion for

contextual bandits, which dominates previously proposed alternatives. We design a statistically and computationally efficient algorithm—for the proposed smooth regret—that works with general function approximation under standard supervised oracles. We also present an adaptive algorithm that automatically adapts to any smoothness level. Our algorithms can be used to recover the previous minimax/Pareto optimal guarantees under the standard regret, e.g., in bandit problems with multiple best arms and Lipschitz/H{ö}lder bandits. We conduct large-scale empirical evaluations demonstrating the efficacy of our proposed algorithms.

Residual-Based Sampling for Online Outlier-Robust PCA

- Tianhao Zhu, Jie Shen
- abstract: Outlier-robust principal component analysis (ORPCA) has been broadly applied in scientific discovery in the last decades. In this paper, we study online ORPCA, an important variant that addresses the practical challenge that the data points arrive in a sequential manner and the goal is to recover the underlying subspace of the clean data with one pass of the data. Our main contribution is the first provable algorithm that enjoys comparable recovery guarantee to the best known batch algorithm, while significantly improving upon the state-of-the-art online ORPCA algorithms. The core technique is a robust version of the residual norm which, informally speaking, leverages not only the importance of a data point, but also how likely it behaves as an outlier.

Region-Based Semantic Factorization in GANs

- Jiapeng Zhu, Yujun Shen, Yinghao Xu, Deli Zhao, Qifeng Chen
- abstract: Despite the rapid advancement of semantic discovery in the latent space of Generative Adversarial Networks (GANs), existing approaches either are limited to finding global attributes or rely on a number of segmentation masks to identify local attributes. In this work, we present a highly efficient algorithm to factorize the latent semantics learned by GANs concerning an arbitrary image region. Concretely, we revisit the task of local manipulation with pre-trained GANs and formulate region-based semantic discovery as a dual optimization problem. Through an appropriately defined generalized Rayleigh quotient, we manage to solve such a problem without any annotations or training. Experimental results on various state-of-the-art GAN models demonstrate the effectiveness of our approach, as well as its superiority over prior arts regarding precise control, region robustness, speed of implementation, and simplicity of use.

Beyond Images: Label Noise Transition Matrix Estimation for Tasks with Lower-Quality Features

- Zhaowei Zhu, Jialu Wang, Yang Liu
- abstract: The label noise transition matrix, denoting the transition probabilities from clean labels to noisy labels, is crucial for designing statistically robust solutions. Existing estimators for noise transition matrices, e.g., using either anchor points or clusterability, focus on computer vision tasks that are relatively easier to obtain high-quality representations. We observe that tasks with lower-quality features fail to meet the anchor-point or clusterability condition, due to the coexistence of both uninformative and informative representations. To handle this issue, we propose a generic and practical information-theoretic approach to down-weight the less informative parts of the lower-quality features. This improvement is crucial to identifying and estimating the label noise transition matrix. The salient technical challenge is to compute the relevant information-theoretical metrics using only noisy labels instead of clean ones. We prove that the celebrated f-mutual information measure can often preserve the order when calculated using noisy labels. We then build our transition matrix estimator using this distilled version of features. The necessity and effectiveness of the proposed method are also demonstrated by evaluating the estimation error on a varied set of tabular data and text classification tasks with lower-quality features. Code is available at github.com/UCSC-REAL/BeyondImages.

Towards Uniformly Superhuman Autonomy via Subdominance Minimization

- Brian Ziebart, Sanjiban Choudhury, Xinyan Yan, Paul Vernaza
- abstract: Prevalent imitation learning methods seek to produce behavior that matches or exceeds average human performance. This often prevents achieving expert-level or superhuman performance when identifying the better demonstrations to imitate is difficult. We instead assume demonstrations are of varying quality and seek to induce behavior that is unambiguously better (i.e., Pareto dominant or minimally subdominant) than all human demonstrations. Our minimum subdominance inverse optimal control training objective is primarily defined by high quality demonstrations; lower quality demonstrations, which are more easily dominated, are effectively ignored instead of degrading imitation. With increasing probability, our approach produces superhuman behavior incurring lower cost than demonstrations on the demonstrator's unknown cost function{—}even if that cost function differs for each demonstration. We apply our approach on a computer cursor pointing task, producing behavior that is 78% superhuman, while minimizing demonstration suboptimality provides 50% superhuman behavior{—}and only 72% even after selective data cleaning.

Inductive Matrix Completion: No Bad Local Minima and a Fast Algorithm

- Pini Zilber, Boaz Nadler
- abstract: The inductive matrix completion (IMC) problem is to recover a low rank matrix from few observed entries while incorporating prior knowledge about its row and column subspaces. In this work, we make three contributions to the IMC problem: (i) we prove that under suitable conditions, the IMC optimization landscape has no bad local minima; (ii) we derive a simple scheme with theoretical guarantees to estimate the rank of the unknown matrix; and (iii) we propose GNIMC, a simple Gauss-Newton based method to solve the IMC problem, analyze its runtime and derive for it strong recovery guarantees. The guarantees for GNIMC are sharper in several aspects than those available for other methods, including a quadratic convergence rate, fewer required observed entries and stability to errors or deviations from low-rank. Empirically, given entries observed uniformly at random, GNIMC recovers the underlying matrix substantially faster than several competing methods.

Counterfactual Prediction for Outcome-Oriented Treatments

- Hao Zou, Bo Li, Jiangang Han, Shuiping Chen, Xuetao Ding, Peng Cui
- abstract: Large amounts of efforts have been devoted into learning counterfactual treatment outcome under various settings, including binary/continuous/multiple treatments. Most of these literature aims to minimize the estimation error of counterfactual outcome for the whole treatment space. However, in most scenarios when the counterfactual prediction model is utilized to assist decision-making, people are only concerned with the small fraction of treatments that can potentially induce superior outcome (i.e. outcome-oriented treatments). This gap of objective is even more severe when the number of possible treatments is large, for example under the continuous treatment setting. To overcome it, we establish a new objective of optimizing counterfactual prediction on outcome-oriented treatments, propose a novel Outcome-Oriented Sample Re-weighting (OOSR) method to make the predictive model concentrate more on outcome-oriented treatments, and theoretically analyze that our method can improve treatment selection towards the optimal one. Extensive experimental results on both synthetic datasets and semi-synthetic datasets demonstrate the effectiveness of our method.

SpaceMAP: Visualizing High-Dimensional Data by Space Expansion

- Xinrui Zu, Qian Tao
- abstract: Dimensionality reduction (DR) of high-dimensional data is of theoretical and practical interest in machine learning. However, there exist intriguing, non-intuitive discrepancies between the geometry of high- and low-dimensional space. We look into such discrepancies and propose a novel

visualization method called Space-based Manifold Approximation and Projection (SpaceMAP). Our method establishes an analytical transformation on distance metrics between spaces to address the “crowding problem" in DR. With the proposed equivalent extended distance (EED), we are able to match the capacity of high- and low-dimensional space in a principled manner. To handle complex data with different manifold properties, we propose hierarchical manifold approximation to model the similarity function in a data-specific manner. We evaluated SpaceMAP on a range of synthetic and real datasets with varying manifold properties, and demonstrated its excellent performance in comparison with classical and state-of-the-art DR methods. In particular, the concept of space expansion provides a generic framework for understanding nonlinear DR methods including the popular t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection