Wed 2024.02.14

Online Sequential Decision-Making with Unknown Delays

Ping Wu, Heyan Huang, Zhengyang Liu

In the field of online sequential decision-making, we address the problem with delays utilizing the framework of online convex optimization (OCO), where the feedback of a decision can arrive with an unknown delay. Unlike previous research that is limited to Euclidean norm and gradient information, we propose three families of delayed algorithms based on approximate solutions to handle different types of received feedback. Our proposed algorithms are versatile and applicable to universal norms. Specifically, we introduce a family of Follow the Delayed Regularized Leader algorithms for feedback with full information on the loss function, a family of Delayed Mirror Descent algorithms for feedback with gradient information on the loss function and a family of Simplified Delayed Mirror Descent algorithms for feedback with the value information of the loss function's gradients at corresponding decision points. For each type of algorithm, we provide corresponding regret bounds under cases of general convexity and relative strong convexity, respectively. We also demonstrate the efficiency of each algorithm under different norms through concrete examples. Furthermore, our theoretical results are consistent with the current best bounds when degenerated to standard settings.

link: http://arxiv.org/abs/2402.07703v1

Signed Distance Field based Segmentation and Statistical Shape Modelling of the Left Atrial Appendage

Kristine Aavild Juhl, Jakob Slipsager, Ole de Backer, Klaus Kofoed, Oscar Camara, Rasmus Paulsen

Patients with atrial fibrillation have a 5-7 fold increased risk of having an ischemic stroke. In these cases, the most common site of thrombus localization is inside the left atrial appendage (LAA) and studies have shown a correlation between the LAA shape and the risk of ischemic stroke. These studies make use of manual measurement and qualitative assessment of shape and are therefore prone to large inter-observer discrepancies, which may explain the contradictions between the conclusions in different studies. We argue that quantitative shape descriptors are necessary to robustly characterize LAA morphology and relate to other functional parameters and stroke risk. Deep Learning methods are becoming standardly available for segmenting cardiovascular structures from high resolution images such as computed tomography (CT), but only few have been tested for LAA segmentation. Furthermore, the majority of segmentation algorithms produces non-smooth 3D models that are not ideal for further processing, such as statistical shape analysis or computational fluid modelling. In this paper we present a fully automatic pipeline for image segmentation, mesh model creation and statistical shape modelling of the LAA. The LAA anatomy is implicitly represented as a signed distance field (SDF), which is directly regressed from the CT image using Deep Learning. The SDF is further used for registering the LAA shapes to a common template and build a statistical shape model (SSM). Based on 106 automatically segmented LAAs, the built SSM reveals that the LAA shape can be quantified using approximately 5 PCA modes and allows the identification of two distinct shape clusters corresponding to the so-called chicken-wing and non-chicken-wing morphologies.

link: http://arxiv.org/abs/2402.07708v1

Optimization of Sparse Convolution for 3D-Point Cloud on GPUs with CUDA

Chester Luo, Kevin Lai

In recent years, there has been a significant increase in the utilization of deep learning methods, particularly convolutional neural networks (CNNs), which have emerged as the dominant approach in various domains that involve structured grid data, such as picture analysis and processing. Nevertheless, the exponential growth in the utilization of LiDAR and 3D sensors across many domains has resulted in an increased need for the analysis of 3D point clouds. The utilization of 3D point clouds is crucial in various applications, including object recognition and segmentation, as they offer a spatial depiction of things within a three-dimensional environment. In contrast to photos, point clouds exhibit sparsity and lack a regular grid, hence posing distinct processing and computational issues.

link: http://arxiv.org/abs/2402.07710v1

Model Collapse Demystified: The Case of Regression

Elvis Dohmatob, Yunzhen Feng, Julia Kempe

In the era of large language models like ChatGPT, the phenomenon of "model collapse" refers to the situation whereby as a model is trained recursively on data generated from previous generations of itself over time, its performance degrades until the model eventually becomes completely useless, i.e the model collapses. In this work, we study this phenomenon in the simplified setting of kernel regression and obtain results which show a clear crossover between where the model can cope with fake data, and a regime where the model's performance completely collapses. Under polynomial decaying spectral and source conditions, we obtain modified scaling laws which exhibit new crossover phenomena from fast to slow rates. We also propose a simple strategy based on adaptive regularization to mitigate model collapse. Our theoretical results are validated with experiments.

link: http://arxiv.org/abs/2402.07712v1

Local Centrality Minimization with Quality Guarantees

Atsushi Miyauchi, Lorenzo Severini, Francesco Bonchi

Centrality measures, quantifying the importance of vertices or edges, play a fundamental role in network analysis. To date, triggered by some positive approximability results, a large body of work has been devoted to studying centrality maximization, where the goal is to maximize the centrality score of a target vertex by manipulating the structure of a given network. On the other hand, due to the lack of such results, only very little attention has been paid to centrality minimization, despite its practical usefulness. In this study, we introduce a novel optimization model for local centrality minimization, where the manipulation is allowed only around the target vertex. We prove the NP-hardness of our model and that the most intuitive greedy algorithm has a quite limited performance in terms of approximation ratio. Then we design two effective approximation algorithms: The first algorithm is a highly-scalable algorithm that has an approximation ratio unachievable by the greedy algorithm, while the second algorithm is a bicriteria approximation algorithm that solves a continuous relaxation based on the Lov\'asz extension, using a projected subgradient method. To the best of our knowledge, ours are the first polynomial-time algorithms with provable approximation guarantees for centrality minimization. Experiments using a variety of real-world networks demonstrate the effectiveness of our proposed algorithms: Our first algorithm is applicable to million-scale graphs and obtains much better solutions than those of scalable baselines, while our second algorithm is rather strong against adversarial instances.

link: http://arxiv.org/abs/2402.07718v1

LoRA-drop: Efficient LoRA Parameter Pruning based on Output Evaluation

Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao

Low-Rank Adaptation (LoRA) introduces auxiliary parameters for each layer to fine-tune the pre-trained model under limited computing resources. But it still faces challenges of resource consumption when scaling up to larger models. Previous studies employ pruning techniques by evaluating the importance of LoRA parameters for different layers to address the problem. However, these efforts only analyzed parameter features to evaluate their importance. Indeed, the output of LoRA related to the parameters and data is the factor that directly impacts the frozen model. To this end, we propose LoRA-drop which evaluates the importance of the parameters by analyzing the LoRA output. We retain LoRA for important layers and the LoRA of the other layers share the same parameters. Abundant experiments on NLU and NLG tasks demonstrate the effectiveness of LoRA-drop.

link: http://arxiv.org/abs/2402.07721v1

Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation

Benjamin Dupuis, Umut ■im■ekli

Understanding the generalization properties of heavy-tailed stochastic optimization algorithms has attracted increasing attention over the past years. While illuminating interesting aspects of stochastic optimizers by using heavy-tailed stochastic differential equations as proxies, prior works either provided expected generalization bounds, or introduced non-computable information theoretic terms. Addressing these drawbacks, in this work, we prove high-probability generalization bounds for heavy-tailed SDEs which do not contain any nontrivial information theoretic terms. To achieve this goal, we develop new proof techniques based on estimating the entropy flows associated with the so-called fractional Fokker-Planck equation (a partial differential equation that governs the evolution of the distribution of the corresponding heavy-tailed SDE). In addition to obtaining high-probability bounds, we show that our bounds have a better dependence on the dimension of parameters as compared to prior art. Our results further identify a phase transition phenomenon, which suggests that heavy tails can be either beneficial or harmful depending on the problem structure. We support our theory with experiments conducted in a variety of settings.

link: http://arxiv.org/abs/2402.07723v1

Unsupervised Sign Language Translation and Generation

Zhengsheng Guo, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Kehai Chen, Zhaopeng Tu, Yong Xu, Min Zhang

Motivated by the success of unsupervised neural machine translation (UNMT), we introduce an unsupervised sign language translation and generation network (USLNet), which learns from abundant single-modality (text and video) data without parallel sign language data. USLNet comprises two main components: single-modality reconstruction modules (text and video) that rebuild the input from its noisy version in the same modality and cross-modality back-translation modules (text-video-text and video-text-video) that reconstruct the input from its noisy version in the different modality using back-translation procedure.Unlike the single-modality back-translation procedure in text-based UNMT, USLNet faces the cross-modality discrepancy in feature representation, in which the length and the feature dimension mismatch between text and video sequences. We propose a sliding window method to address the issues of aligning variable-length text with video sequences. To our knowledge, USLNet is the first unsupervised sign language translation and generation model capable of generating both natural language text and sign language video in a unified manner. Experimental results on the BBC-Oxford Sign Language dataset (BOBSL) and Open-Domain American Sign Language dataset (OpenASL) reveal that USLNet achieves competitive results compared to supervised baseline models, indicating its effectiveness in sign language translation and generation.

link: http://arxiv.org/abs/2402.07726v1

AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, Jingren Zhou

Recently, instruction-following audio-language models have received broad attention for human-audio interaction. However, the absence of benchmarks capable of evaluating audio-centric interaction capabilities has impeded advancements in this field. Previous models primarily focus on assessing different fundamental tasks, such as Automatic Speech Recognition (ASR), and lack an assessment of the open-ended generative capabilities centered around audio. Thus, it is challenging to track the progression in the Large Audio-Language Models (LALMs) domain and to provide guidance for future improvement. In this paper, we introduce AIR-Bench (\textbf{A}udio \textbf{I}nst\textbf{R}uction \textbf{Bench}mark), the first benchmark designed to evaluate the ability of LALMs to understand various types of audio signals (including human speech, natural sounds, and music), and furthermore, to interact with humans in the textual format. AIR-Bench encompasses two dimensions: \textit{foundation} and \textit{chat} benchmarks. The former consists of 19 tasks with approximately 19k single-choice questions, intending to inspect the basic single-task ability of LALMs. The latter one contains 2k instances of open-ended question-and-answer data, directly assessing the comprehension of the model on complex audio and its capacity to follow instructions. Both benchmarks require the model to generate hypotheses directly. We design a unified framework that leverages advanced language models, such as GPT-4, to evaluate the scores of generated hypotheses given the meta-information of the audio. Experimental results demonstrate a high level of consistency between GPT-4-based evaluation and human evaluation. By revealing the limitations of existing LALMs through evaluation results, AIR-Bench can provide insights into the direction of future research.

link: http://arxiv.org/abs/2402.07729v1

Graph Structure Inference with BAM: Introducing the Bilinear Attention Mechanism

Philipp Froehlich, Heinz Koeppl

In statistics and machine learning, detecting dependencies in datasets is a central challenge. We propose a novel neural network model for supervised graph structure learning, i.e., the process of learning a mapping between observational data and their underlying dependence structure. The model is trained with variably shaped and coupled simulated input data and requires only a single forward pass through the trained network for inference. By leveraging structural equation models and employing randomly generated multivariate Chebyshev polynomials for the simulation of training data, our method demonstrates robust generalizability across both linear and various types of non-linear dependencies. We introduce a novel bilinear attention mechanism (BAM) for explicit processing of dependency information, which operates on the level of covariance matrices of transformed data and respects the geometry of the manifold of symmetric positive definite matrices. Empirical evaluation demonstrates the robustness of our method in detecting a wide range of dependencies, excelling in undirected graph estimation and proving competitive in completed partially directed acyclic graph estimation through a novel two-step approach.

link: http://arxiv.org/abs/2402.07735v2

Universal link predictor by In-context Learning

Kaiwen Dong, Haitao Mao, Zhichun Guo, Nitesh V. Chawla

Link prediction is a crucial task in graph machine learning, where the goal is to infer missing or future links within a graph. Traditional approaches leverage heuristic methods based on widely observed connectivity patterns, offering broad applicability and generalizability without the need for model training. Despite their utility, these methods are limited by their reliance on human-derived heuristics and lack the adaptability of data-driven approaches. Conversely, parametric link predictors excel in automatically learning the connectivity patterns from data and achieving state-of-the-art but fail short to directly transfer across different graphs. Instead, it requires the cost of extensive training and hyperparameter optimization to adapt to the target graph. In this work, we introduce the Universal Link Predictor (UniLP), a novel model that combines the generalizability of heuristic approaches with the pattern learning capabilities of parametric models. UniLP is designed to autonomously identify connectivity patterns across diverse graphs, ready for immediate application to any unseen graph dataset without targeted training. We address the challenge of conflicting connectivity patterns-arising from the unique distributions of different graphs-through the implementation of In-context Learning (ICL). This approach allows UniLP to dynamically adjust to various target graphs based on contextual demonstrations, thereby avoiding negative transfer. Through rigorous experimentation, we demonstrate UniLP's effectiveness in adapting to new, unseen graphs at test time, showcasing its ability to perform comparably or even outperform parametric models that have been finetuned for specific datasets. Our findings highlight UniLP's potential to set a new standard in link prediction, combining the strengths of heuristic and parametric methods in a single, versatile framework.

link: http://arxiv.org/abs/2402.07738v1

Task-conditioned adaptation of visual features in multi-task policy learning

Pierre Marza, Laetitia Matignon, Olivier Simonin, Christian Wolf

Successfully addressing a wide variety of tasks is a core ability of autonomous agents, which requires flexibly adapting the underlying decision-making strategies and, as we argue in this work, also adapting the underlying perception modules. An analogical argument would be the human visual system, which uses top-down signals to focus attention determined by the current task. Similarly, in this work, we adapt pre-trained large vision models conditioned on specific downstream tasks in the context of multi-task policy learning. We introduce task-conditioned adapters that do not require finetuning any pre-trained weights, combined with a single policy trained with behavior cloning and capable of addressing multiple tasks. We condition the policy and visual adapters on task embeddings, which can be selected at inference if the task is known, or alternatively inferred from a set of example demonstrations. To this end, we propose a new optimization-based estimator. We evaluate the method on a wide variety of tasks of the CortexBench benchmark and show that, compared to existing work, it can be addressed with a single policy. In particular, we demonstrate that adapting visual features is a key design choice and that the method generalizes to unseen tasks given visual demonstrations.

link: http://arxiv.org/abs/2402.07739v1

Asking Multimodal Clarifying Questions in Mixed-Initiative Conversational Search

Yifei Yuan, Clemencia Siro, Mohammad Aliannejadi, Maarten de Rijke, Wai Lam

In mixed-initiative conversational search systems, clarifying questions are used to help users who struggle to express their intentions in a single query. These questions aim to uncover user's information needs and resolve query ambiguities. We hypothesize that in scenarios where multimodal information is pertinent, the clarification process can be improved by using non-textual information. Therefore, we propose to add images to clarifying questions and formulate the novel task of asking multimodal clarifying questions in open-domain, mixed-initiative conversational search systems. To facilitate research into this task, we collect a dataset named Melon that contains over 4k multimodal clarifying questions, enriched with over 14k images. We also propose a multimodal query clarification model named Marto and adopt a prompt-based, generative fine-tuning strategy to perform the training of different stages with different prompts. Several analyses are conducted to understand the importance of multimodal contents during the query clarification phase. Experimental results indicate that the addition of images leads to significant improvements of up to 90% in retrieval performance when selecting the relevant images. Extensive analyses are also performed to show the superiority of Marto compared with discriminative baselines in terms of effectiveness and efficiency.

link: http://arxiv.org/abs/2402.07742v1

Towards Unified Alignment Between Agents, Humans, and Environment

Zonghan Yang, An Liu, Zijun Liu, Kaiming Liu, Fangzhou Xiong, Yile Wang, Zeyuan Yang, Qingyuan Hu, Xinrui Chen, Zhenhe Zhang, Fuwen Luo, Zhicheng Guo, Peng Li, Yang Liu

The rapid progress of foundation models has led to the prosperity of autonomous agents, which leverage the universal capabilities of foundation models to conduct reasoning, decision-making, and environmental interaction. However, the efficacy of agents remains limited when operating in intricate, realistic environments. In this work, we introduce the principles of $\mathbf{U}$nified $\mathbf{A}$lignment for $\mathbf{A}$gents ($\mathbf{UA}^2$), which advocate for the simultaneous alignment of agents with human intentions, environmental dynamics, and self-constraints such as the limitation of monetary budgets. From the perspective of $\mathbf{UA}^2$, we review the current agent research and highlight the neglected factors in existing agent benchmarks and method candidates. We also conduct proof-of-concept studies by introducing realistic features to WebShop, including user profiles to demonstrate intentions, personalized reranking for complex environmental dynamics, and runtime cost statistics to reflect self-constraints. We then follow the principles of $\mathbf{UA}^2$ to propose an initial design of our agent, and benchmark its performance with several candidate baselines in the retrofitted WebShop. The extensive experimental results further prove the importance of the principles of $\mathbf{UA}^2$. Our research sheds light on the next steps of autonomous agent research with improved general problem-solving abilities.

link: http://arxiv.org/abs/2402.07744v1

Predictive Churn with the Set of Good Models

Jamelle Watson-Daniels, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C. Parkes, Berk Ustun

Machine learning models in modern mass-market applications are often updated over time. One of the foremost challenges faced is that, despite increasing overall performance, these updates may flip specific model predictions in unpredictable ways. In practice, researchers quantify the number of unstable predictions between models pre and post update -- i.e., predictive churn. In this paper, we study this effect through the lens of predictive multiplicity -- i.e., the prevalence of conflicting predictions over the set of near-optimal models (the Rashomon set). We show how traditional measures of predictive multiplicity can be used to examine expected churn over this set of prospective models -- i.e., the set of models that may be used to replace a baseline model in deployment. We present theoretical results on the expected churn between models within the Rashomon set from different perspectives. And we characterize expected churn over model updates via the Rashomon set, pairing our analysis with empirical results on real-world datasets -- showing how our approach can be used to better anticipate, reduce, and avoid churn in consumer-facing applications. Further, we show that our approach is useful even for models enhanced with uncertainty awareness.

link: http://arxiv.org/abs/2402.07745v1