Fri 2024.02.23

Replicable Learning of Large-Margin Halfspaces

Alkis Kalavasis, Amin Karbasi, Kasper Green Larsen, Grigoris Velegkas, Felix Zhou

We provide efficient replicable algorithms for the problem of learning large-margin halfspaces. Our results improve upon the algorithms provided by Impagliazzo, Lei, Pitassi, and Sorrell [STOC, 2022]. We design the first dimension-independent replicable algorithms for this task which runs in polynomial time, is proper, and has strictly improved sample complexity compared to the one achieved by Impagliazzo et al. [2022] with respect to all the relevant parameters. Moreover, our first algorithm has sample complexity that is optimal with respect to the accuracy parameter ϵ. We also design an SGD-based replicable algorithm that, in some parameters' regimes, achieves better sample and time complexity than our first algorithm. Departing from the requirement of polynomial time algorithms, using the DP-to-Replicability reduction of Bun, Gaboardi, Hopkins, Impagliazzo, Lei, Pitassi, Sorrell, and Sivakumar [STOC, 2023], we show how to obtain a replicable algorithm for large-margin halfspaces with improved sample complexity with respect to the margin parameter τ, but running time doubly exponential in $1/\tau^2$ and worse sample complexity dependence on ϵ than one of our previous algorithms. We then design an improved algorithm with better sample complexity than all three of our previous algorithms and running time exponential in $1/\tau^{2}$.

link: http://arxiv.org/abs/2402.13857v1

Diversity-Aware k-Maximum Inner Product Search Revisited

Qiang Huang, Yanhao Wang, Yiqun Sun, Anthony K. H. Tung

The k-Maximum Inner Product Search (kMIPS) serves as a foundational component in recommender systems and various data mining tasks. However, while most existing kMIPS approaches prioritize the efficient retrieval of highly relevant items for users, they often neglect an equally pivotal facet of search results: \emph{diversity}. To bridge this gap, we revisit and refine the diversity-aware kMIPS (DkMIPS) problem by incorporating two well-known diversity objectives -- minimizing the average and maximum pairwise item similarities within the results -- into the original relevance objective. This enhancement, inspired by Maximal Marginal Relevance (MMR), offers users a controllable trade-off between relevance and diversity. We introduce \textsc{Greedy} and \textsc{DualGreedy}, two linear scan-based algorithms tailored for DkMIPS. They both achieve data-dependent approximations and, when aiming to minimize the average pairwise similarity, \textsc{DualGreedy} attains an approximation ratio of $1/4$ with an additive term for regularization. To further improve query efficiency, we integrate a lightweight Ball-Cone Tree (BC-Tree) index with the two algorithms. Finally, comprehensive experiments on ten real-world data sets demonstrate the efficacy of our proposed methods, showcasing their capability to efficiently deliver diverse and relevant search results to users.

link: http://arxiv.org/abs/2402.13858v1

Improving Efficiency of Iso-Surface Extraction on Implicit Neural Representations Using Uncertainty Propagation

Haoyu Li, Han-Wei Shen

Implicit Neural representations (INRs) are widely used for scientific data reduction and visualization by modeling the function that maps a spatial location to a data value. Without any prior knowledge about the spatial distribution of values, we are forced to sample densely from INRs to perform visualization tasks like iso-surface extraction which can be very computationally expensive. Recently, range analysis has shown promising results in improving the efficiency of geometric queries, such as ray casting and hierarchical mesh extraction, on INRs for 3D geometries by using arithmetic rules to bound the output range of the network within a spatial region. However, the analysis bounds are often too conservative for complex scientific data. In this paper, we present an improved technique for range analysis by revisiting the arithmetic rules and analyzing the

probability distribution of the network output within a spatial region. We model this distribution efficiently as a Gaussian distribution by applying the central limit theorem. Excluding low probability values, we are able to tighten the output bounds, resulting in a more accurate estimation of the value range, and hence more accurate identification of iso-surface cells and more efficient iso-surface extraction on INRs. Our approach demonstrates superior performance in terms of the iso-surface extraction time on four datasets compared to the original range analysis method and can also be generalized to other geometric query tasks.

link: http://dx.doi.org/10.1109/TVCG.2024.3365089

Kuaiji: the First Chinese Accounting Large Language Model

Jiayuan Luo, Songhua Yang, Xiaoling Qiu, Panyu Chen, Yufei Nai, Wenxuan Zeng, Wentao Zhang, Xinke Jiang

Large Language Models (LLMs) like ChatGPT and GPT-4 have demonstrated impressive proficiency in comprehending and generating natural language. However, they encounter difficulties when tasked with adapting to specialized domains such as accounting. To address this challenge, we introduce Kuaiji, a tailored Accounting Large Language Model. Kuaiji is meticulously fine-tuned using the Baichuan framework, which encompasses continuous pre-training and supervised fine-tuning processes. Supported by CAtAcctQA, a dataset containing large genuine accountant-client dialogues, Kuaiji exhibits exceptional accuracy and response speed. Our contributions encompass the creation of the first Chinese accounting dataset, the establishment of Kuaiji as a leading open-source Chinese accounting LLM, and the validation of its efficacy through real-world accounting scenarios.

link: http://arxiv.org/abs/2402.13866v1

Generative Probabilistic Time Series Forecasting and Applications in Grid Operations

Xinyi Wang, Lang Tong, Qing Zhao

Generative probabilistic forecasting produces future time series samples according to the conditional probability distribution given past time series observations. Such techniques are essential in risk-based decision-making and planning under uncertainty with broad applications in grid operations, including electricity price forecasting, risk-based economic dispatch, and stochastic optimizations. Inspired by Wiener and Kallianpur's innovation representation, we propose a weak innovation autoencoder architecture and a learning algorithm to extract independent and identically distributed innovation sequences from nonparametric stationary time series. We show that the weak innovation sequence is Bayesian sufficient, which makes the proposed weak innovation autoencoder a canonical architecture for generative probabilistic forecasting. The proposed technique is applied to forecasting highly volatile real-time electricity prices, demonstrating superior performance across multiple forecasting measures over leading probabilistic and point forecasting techniques.

link: http://arxiv.org/abs/2402.13870v1

An Explainable Transformer-based Model for Phishing Email Detection: A Large Language Model Approach

Mohammad Amaz Uddin, Iqbal H. Sarker

Phishing email is a serious cyber threat that tries to deceive users by sending false emails with the intention of stealing confidential information or causing financial harm. Attackers, often posing as trustworthy entities, exploit technological advancements and sophistication to make detection and prevention of phishing more challenging. Despite extensive academic research, phishing detection remains an ongoing and formidable challenge in the cybersecurity landscape. Large Language Models (LLMs) and Masked Language Models (MLMs) possess immense potential to offer innovative solutions to address long-standing challenges. In this research paper, we present an optimized, fine-tuned transformer-based DistilBERT model designed for the detection of phishing emails. In the detection process, we work with a phishing email dataset and utilize the

preprocessing techniques to clean and solve the imbalance class issues. Through our experiments, we found that our model effectively achieves high accuracy, demonstrating its capability to perform well. Finally, we demonstrate our fine-tuned model using Explainable-AI (XAI) techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and Transformer Interpret to explain how our model makes predictions in the context of text classification for phishing emails.

link: http://arxiv.org/abs/2402.13871v1

$\texttt{Se}^2$: $\textit{Se}$quential Example $\textit{Se}$lection for In-Context Learning

Haoyu Liu, Jianfeng Liu, Shaohan Huang, Yuefeng Zhan, Hao Sun, Weiwei Deng, Furu Wei, Qi Zhang

The remarkable capability of large language models (LLMs) for in-context learning (ICL) needs to be activated by demonstration examples. Prior work has extensively explored the selection of examples for ICL, predominantly following the "select then organize" paradigm, such approaches often neglect the internal relationships between examples and exist an inconsistency between the training and inference. In this paper, we formulate the problem as a $\textit{se}$quential $\textit{se}$lection problem and introduce $\texttt{Se}^2$, a sequential-aware method that leverages the LLM's feedback on varying context, aiding in capturing inter-relationships and sequential information among examples, significantly enriching the contextuality and relevance of ICL prompts. Meanwhile, we utilize beam search to seek and construct example sequences, enhancing both quality and diversity. Extensive experiments across 23 NLP tasks from 8 distinct categories illustrate that $\texttt{Se}^2$ markedly surpasses competitive baselines and achieves 42% relative improvement over random selection. Further in-depth analysis show the effectiveness of proposed strategies, highlighting $\texttt{Se}^2$'s exceptional stability and adaptability across various scenarios. Our code will be released to facilitate future research.

link: http://arxiv.org/abs/2402.13874v1

Scene Prior Filtering for Depth Map Super-Resolution

Zhengxue Wang, Zhiqiang Yan, Ming-Hsuan Yang, Jinshan Pan, Jian Yang, Ying Tai, Guangwei Gao

Multi-modal fusion is vital to the success of super-resolution of depth images. However, commonly used fusion strategies, such as addition and concatenation, fall short of effectively bridging the modal gap. As a result, guided image filtering methods have been introduced to mitigate this issue. Nevertheless, it is observed that their filter kernels usually encounter significant texture interference and edge inaccuracy. To tackle these two challenges, we introduce a Scene Prior Filtering network, SPFNet, which utilizes the priors surface normal and semantic map from large-scale models. Specifically, we design an All-in-one Prior Propagation that computes the similarity between multi-modal scene priors, \textit{i.e.}, RGB, normal, semantic, and depth, to reduce the texture interference. In addition, we present a One-to-one Prior Embedding that continuously embeds each single-modal prior into depth using Mutual Guided Filtering, further alleviating the texture interference while enhancing edges. Our SPFNet has been extensively evaluated on both real and synthetic datasets, achieving state-of-the-art performance.

link: http://arxiv.org/abs/2402.13876v1

Beyond Probabilities: Unveiling the Misalignment in Evaluating Large Language Models

Chenyang Lyu, Minghao Wu, Alham Fikri Aji

Large Language Models (LLMs) have demonstrated remarkable capabilities across various applications, fundamentally reshaping the landscape of natural language processing (NLP) research. However, recent evaluation frameworks often rely on the output probabilities of LLMs for predictions, primarily due to computational constraints, diverging from real-world LLM usage scenarios. While widely employed, the efficacy of these probability-based evaluation strategies remains an open research question. This study aims to scrutinize the validity of such

probability-based evaluation methods within the context of using LLMs for Multiple Choice Questions (MCQs), highlighting their inherent limitations. Our empirical investigation reveals that the prevalent probability-based evaluation method inadequately aligns with generation-based prediction. Furthermore, current evaluation frameworks typically assess LLMs through predictive tasks based on output probabilities rather than directly generating responses, owing to computational limitations. We illustrate that these probability-based approaches do not effectively correspond with generative predictions. The outcomes of our study can enhance the understanding of LLM evaluation methodologies and provide insights for future research in this domain.

link: http://arxiv.org/abs/2402.13887v1

Overcoming Saturation in Density Ratio Estimation by Iterated Regularization

Lukas Gruber, Markus Holzleitner, Johannes Lehner, Sepp Hochreiter, Werner Zellinger

Estimating the ratio of two probability densities from finitely many samples, is a central task in machine learning and statistics. In this work, we show that a large class of kernel methods for density ratio estimation suffers from error saturation, which prevents algorithms from achieving fast error convergence rates on highly regular learning problems. To resolve saturation, we introduce iterated regularization in density ratio estimation to achieve fast error rates. Our methods outperform its non-iteratively regularized versions on benchmarks for density ratio estimation as well as on large-scale evaluations for importance-weighted ensembling of deep unsupervised domain adaptation models.

link: http://arxiv.org/abs/2402.13891v1

Science Checker Reloaded: A Bidirectional Paradigm for Transparency and Logical Reasoning

Loïc Rakotoson, Sylvain Massip, Fréjus A. A. Laleye

Information retrieval is a rapidly evolving field. However it still faces significant limitations in the scientific and industrial vast amounts of information, such as semantic divergence and vocabulary gaps in sparse retrieval, low precision and lack of interpretability in semantic search, or hallucination and outdated information in generative models. In this paper, we introduce a two-block approach to tackle these hurdles for long documents. The first block enhances language understanding in sparse retrieval by query expansion to retrieve relevant documents. The second block deepens the result by providing comprehensive and informative answers to the complex question using only the information spread in the long document, enabling bidirectional engagement. At various stages of the pipeline, intermediate results are presented to users to facilitate understanding of the system's reasoning. We believe this bidirectional approach brings significant advancements in terms of transparency, logical thinking, and comprehensive understanding in the field of scientific information retrieval.

link: http://arxiv.org/abs/2402.13897v1

Non-asymptotic Convergence of Discrete-time Diffusion Models: New Approach and Improved Rate

Yuchen Liang, Peizhong Ju, Yingbin Liang, Ness Shroff

The denoising diffusion model emerges recently as a powerful generative technique that converts noise into data. Theoretical convergence guarantee has been mainly studied for continuous-time diffusion models, and has been obtained for discrete-time diffusion models only for distributions with bounded support in the literature. In this paper, we establish the convergence guarantee for substantially larger classes of distributions under discrete-time diffusion models and further improve the convergence rate for distributions with bounded support. In particular, we first establish the convergence rates for both smooth and general (possibly non-smooth) distributions having finite second moment. We then specialize our results to a number of interesting classes of distributions with explicit parameter dependencies, including distributions with Lipschitz scores, Gaussian mixture distributions, and distributions with bounded support. We further propose a novel accelerated sampler and show that it improves the convergence rates of the corresponding regular

sampler by orders of magnitude with respect to all system parameters. For distributions with bounded support, our result improves the dimensional dependence of the previous convergence rate by orders of magnitude. Our study features a novel analysis technique that constructs tilting factor representation of the convergence error and exploits Tweedie's formula for handling Taylor expansion power terms.

link: http://arxiv.org/abs/2402.13901v1

Dealing with unbounded gradients in stochastic saddle-point optimization

Gergely Neu, Nneka Okolo

We study the performance of stochastic first-order methods for finding saddle points of convex-concave functions. A notorious challenge faced by such methods is that the gradients can grow arbitrarily large during optimization, which may result in instability and divergence. In this paper, we propose a simple and effective regularization technique that stabilizes the iterates and yields meaningful performance guarantees even if the domain and the gradient noise scales linearly with the size of the iterates (and is thus potentially unbounded). Besides providing a set of general results, we also apply our algorithm to a specific problem in reinforcement learning, where it leads to performance guarantees for finding near-optimal policies in an average-reward MDP without prior knowledge of the bias span.

link: http://arxiv.org/abs/2402.13903v1

Calibrating Large Language Models with Sample Consistency

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, Chris Callison-Burch

Accurately gauging the confidence level of Large Language Models' (LLMs) predictions is pivotal for their reliable application. However, LLMs are often uncalibrated inherently and elude conventional calibration techniques due to their proprietary nature and massive scale. In this work, we explore the potential of deriving confidence from the distribution of multiple randomly sampled model generations, via three measures of consistency. We perform an extensive evaluation across various open and closed-source models on nine reasoning datasets. Results show that consistency-based calibration methods outperform existing post-hoc approaches. Meanwhile, we find that factors such as intermediate explanations, model scaling, and larger sample sizes enhance calibration, while instruction-tuning makes calibration more difficult. Moreover, confidence scores obtained from consistency have the potential to enhance model performance. Finally, we offer practical guidance on choosing suitable consistency metrics for calibration, tailored to the characteristics of various LMs.

link: http://arxiv.org/abs/2402.13904v1

Leveraging Collection-Wide Similarities for Unsupervised Document Structure Extraction

Gili Lior, Yoav Goldberg, Gabriel Stanovsky

Document collections of various domains, e.g., legal, medical, or financial, often share some underlying collection-wide structure, which captures information that can aid both human users and structure-aware models. We propose to identify the typical structure of document within a collection, which requires to capture recurring topics across the collection, while abstracting over arbitrary header paraphrases, and ground each topic to respective document locations. These requirements pose several challenges: headers that mark recurring topics frequently differ in phrasing, certain section headers are unique to individual documents and do not reflect the typical structure, and the order of topics can vary between documents. Subsequently, we develop an unsupervised graph-based method which leverages both inter- and intra-document similarities, to extract the underlying collection-wide structure. Our evaluations on three diverse domains in both English and Hebrew indicate that our method extracts meaningful collection-wide structure, and we hope that future work will leverage our method for multi-document applications and structure-aware models.

link: http://arxiv.org/abs/2402.13906v1