Sat 2024.02.17

LoraRetriever: Input-Aware LoRA Retrieval and Composition for Mixed Tasks in the Wild

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, Fei Wu

Low-Rank Adaptation (LoRA) provides an effective yet efficient solution for fine-tuning large language models (LLM). The modular and plug-and-play nature of LoRA enables the integration of diverse domain-specific LoRAs to enhance the capabilities of LLMs. Previous research on exploiting multiple LoRAs either focuses on specific isolated downstream tasks or fixes the selection of LoRAs during training. However, in real-world scenarios, LLMs receive diverse prompts covering different tasks, and the pool of candidate LoRAs is often dynamically updated. To bridge this gap, we propose LoraRetriever, a retrieve-then-compose framework that adaptively retrieves and composes multiple LoRAs according to the input prompts. LoraRetriever contains three main components: firstly, identifying and retrieving LoRAs relevant to the given input; secondly, formulating strategies for effectively integrating the retrieved LoRAs; and thirdly, developing efficient batch inference to accommodate heterogeneous requests. Experimental results indicate that LoraRetriever consistently outperforms the baselines, highlighting its practical effectiveness and versatility.

link: http://arxiv.org/abs/2402.09997v1

Privacy Attacks in Decentralized Learning

Abdellah El Mrini, Edwige Cyffers, Aurélien Bellet

Decentralized Gradient Descent (D-GD) allows a set of users to perform collaborative learning without sharing their data by iteratively averaging local model updates with their neighbors in a network graph. The absence of direct communication between non-neighbor nodes might lead to the belief that users cannot infer precise information about the data of others. In this work, we demonstrate the opposite, by proposing the first attack against D-GD that enables a user (or set of users) to reconstruct the private data of other users outside their immediate neighborhood. Our approach is based on a reconstruction attack against the gossip averaging protocol, which we then extend to handle the additional challenges raised by D-GD. We validate the effectiveness of our attack on real graphs and datasets, showing that the number of users compromised by a single or a handful of attackers is often surprisingly large. We empirically investigate some of the factors that affect the performance of the attack, namely the graph topology, the number of attackers, and their position in the graph.

link: http://arxiv.org/abs/2402.10001v1

MM-Point: Multi-View Information-Enhanced Multi-Modal Self-Supervised 3D Point Cloud Understanding

Hai-Tao Yu, Mofei Song

In perception, multiple sensory information is integrated to map visual information from 2D views onto 3D objects, which is beneficial for understanding in 3D environments. But in terms of a single 2D view rendered from different angles, only limited partial information can be provided.The richness and value of Multi-view 2D information can provide superior self-supervised signals for 3D objects. In this paper, we propose a novel self-supervised point cloud representation learning method, MM-Point, which is driven by intra-modal and inter-modal similarity objectives. The core of MM-Point lies in the Multi-modal interaction and transmission between 3D objects and multiple 2D views at the same time. In order to more effectively simultaneously perform the consistent cross-modal objective of 2D multi-view information based on contrastive learning, we further propose Multi-MLP and Multi-level Augmentation strategies. Through carefully designed transformation strategies, we further learn Multi-level invariance in 2D Multi-views. MM-Point demonstrates state-of-the-art (SOTA) performance in various downstream tasks. For instance, it achieves a peak accuracy of 92.4% on the synthetic dataset ModelNet40, and a top accuracy of

87.8% on the real-world dataset ScanObjectNN, comparable to fully supervised methods. Additionally, we demonstrate its effectiveness in tasks such as few-shot classification, 3D part segmentation and 3D semantic segmentation.

link: http://arxiv.org/abs/2402.10002v1

Zero-Shot Unsupervised and Text-Based Audio Editing Using DDPM Inversion

Hila Manor, Tomer Michaeli

Editing signals using large pre-trained models, in a zero-shot manner, has recently seen rapid advancements in the image domain. However, this wave has yet to reach the audio domain. In this paper, we explore two zero-shot editing techniques for audio signals, which use DDPM inversion on pre-trained diffusion models. The first, adopted from the image domain, allows text-based editing. The second, is a novel approach for discovering semantically meaningful editing directions without supervision. When applied to music signals, this method exposes a range of musically interesting modifications, from controlling the participation of specific instruments to improvisations on the melody. Samples can be found on our examples page in https://hilamanor.github.io/AudioEditing/ and code can be found in https://github.com/hilamanor/AudioEditing/ .

link: http://arxiv.org/abs/2402.10009v1

Clifford Group Equivariant Simplicial Message Passing Networks

Cong Liu, David Ruhe, Floor Eijkelboom, Patrick Forré

We introduce Clifford Group Equivariant Simplicial Message Passing Networks, a method for steerable E(n)-equivariant message passing on simplicial complexes. Our method integrates the expressivity of Clifford group-equivariant layers with simplicial message passing, which is topologically more intricate than regular graph message passing. Clifford algebras include higher-order objects such as bivectors and trivectors, which express geometric features (e.g., areas, volumes) derived from vectors. Using this knowledge, we represent simplex features through geometric products of their vertices. To achieve efficient simplicial message passing, we share the parameters of the message network across different dimensions. Additionally, we restrict the final message to an aggregation of the incoming messages from different dimensions, leading to what we term shared simplicial message passing. Experimental results show that our method is able to outperform both equivariant and simplicial graph neural networks on a variety of geometric tasks.

link: http://arxiv.org/abs/2402.10011v1

Bridging the Empirical-Theoretical Gap in Neural Network Formal Language Learning Using Minimum Description Length

Nur Lan, Emmanuel Chemla, Roni Katzir

Neural networks offer good approximation to many tasks but consistently fail to reach perfect generalization, even when theoretical work shows that such perfect solutions can be expressed by certain architectures. Using the task of formal language learning, we focus on one simple formal language and show that the theoretically correct solution is in fact not an optimum of commonly used objectives -- even with regularization techniques that according to common wisdom should lead to simple weights and good generalization (L1, L2) or other meta-heuristics (early-stopping, dropout). However, replacing standard targets with the Minimum Description Length objective (MDL) results in the correct solution being an optimum.

link: http://arxiv.org/abs/2402.10013v1

SAWEC: Sensing-Assisted Wireless Edge Computing

Khandaker Foysal Haque, Francesca Meneghello, Md. Ebtidaul Karim, Francesco Restuccia

Emerging mobile virtual reality (VR) systems will require to continuously perform complex computer vision tasks on ultra-high-resolution video frames through the execution of deep neural networks (DNNs)-based algorithms. Since state-of-the-art DNNs require computational power that is excessive for mobile devices, techniques based on wireless edge computing (WEC) have been

recently proposed. However, existing WEC methods require the transmission and processing of a high amount of video data which may ultimately saturate the wireless link. In this paper, we propose a novel Sensing-Assisted Wireless Edge Computing (SAWEC) paradigm to address this issue. SAWEC leverages knowledge about the physical environment to reduce the end-to-end latency and overall computational burden by transmitting to the edge server only the relevant data for the delivery of the service. Our intuition is that the transmission of the portion of the video frames where there are no changes with respect to previous frames can be avoided. Specifically, we leverage wireless sensing techniques to estimate the location of objects in the environment and obtain insights about the environment dynamics. Hence, only the part of the frames where any environmental change is detected is transmitted and processed. We evaluated SAWEC by using a 10K 360° camera with a Wi-Fi 6 sensing system operating at 160 MHz and performing localization and tracking. We perform experiments in an anechoic chamber and a hall room with two human subjects in six different setups. Experimental results show that SAWEC reduces the channel occupation, and end-to-end latency by 93.81%, and 96.19% respectively while improving the instance segmentation performance by 46.98% with respect to state-of-the-art WEC approaches. For reproducibility purposes, we pledge to share our whole dataset and code repository.

link: http://arxiv.org/abs/2402.10021v1

Self-Augmented In-Context Learning for Unsupervised Word Translation

Yaoyiran Li, Anna Korhonen, Ivan Vuli■

Recent work has shown that, while large language models (LLMs) demonstrate strong word translation or bilingual lexicon induction (BLI) capabilities in few-shot setups, they still cannot match the performance of 'traditional' mapping-based approaches in the unsupervised scenario where no seed translation pairs are available, especially for lower-resource languages. To address this challenge with LLMs, we propose self-augmented in-context learning (SAIL) for unsupervised BLI: starting from a zero-shot prompt, SAIL iteratively induces a set of high-confidence word translation pairs for in-context learning (ICL) from an LLM, which it then reapplies to the same LLM in the ICL fashion. Our method shows substantial gains over zero-shot prompting of LLMs on two established BLI benchmarks spanning a wide range of language pairs, also outperforming mapping-based baselines across the board. In addition to achieving state-of-the-art unsupervised BLI performance, we also conduct comprehensive analyses on SAIL and discuss its limitations.

link: http://arxiv.org/abs/2402.10024v1

Hybrid CNN Bi-LSTM neural network for Hyperspectral image classification

Alok Ranjan Sahoo, Pavan Chakraborty

Hyper spectral images have drawn the attention of the researchers for its complexity to classify. It has nonlinear relation between the materials and the spectral information provided by the HSI image. Deep learning methods have shown superiority in learning this nonlinearity in comparison to traditional machine learning methods. Use of 3-D CNN along with 2-D CNN have shown great success for learning spatial and spectral features. However, it uses comparatively large number of parameters. Moreover, it is not effective to learn inter layer information. Hence, this paper proposes a neural network combining 3-D CNN, 2-D CNN and Bi-LSTM. The performance of this model has been tested on Indian Pines(IP) University of Pavia(PU) and Salinas Scene(SA) data sets. The results are compared with the state of-the-art deep learning-based models. This model performed better in all three datasets. It could achieve 99.83, 99.98 and 100 percent accuracy using only 30 percent trainable parameters of the state-of-art model in IP, PU and SA datasets respectively.

link: http://arxiv.org/abs/2402.10026v1

Diffusion Models Meet Contextual Bandits with Large Action Spaces

Imad Aouali

Efficient exploration is a key challenge in contextual bandits due to the large size of their action space, where uninformed exploration can result in computational and statistical inefficiencies. Fortunately, the rewards of actions are often correlated and this can be leveraged to explore them

efficiently. In this work, we capture such correlations using pre-trained diffusion models; upon which we design diffusion Thompson sampling (dTS). Both theoretical and algorithmic foundations are developed for dTS, and empirical evaluation also shows its favorable performance.

link: http://arxiv.org/abs/2402.10028v1

Investigation of Federated Learning Algorithms for Retinal Optical Coherence Tomography Image Classification with Statistical Heterogeneity

Sanskar Amgain, Prashant Shrestha, Sophia Bano, Ignacio del Valle Torres, Michael Cunniffe, Victor Hernandez, Phil Beales, Binod Bhattarai

Purpose: We apply federated learning to train an OCT image classifier simulating a realistic scenario with multiple clients and statistical heterogeneous data distribution where data in the clients lack samples of some categories entirely. Methods: We investigate the effectiveness of FedAvg and FedProx to train an OCT image classification model in a decentralized fashion, addressing privacy concerns associated with centralizing data. We partitioned a publicly available OCT dataset across multiple clients under IID and Non-IID settings and conducted local training on the subsets for each client. We evaluated two federated learning methods, FedAvg and FedProx for these settings. Results: Our experiments on the dataset suggest that under IID settings, both methods perform on par with training on a central data pool. However, the performance of both algorithms declines as we increase the statistical heterogeneity across the client data, while FedProx consistently performs better than FedAvg in the increased heterogeneity settings. Conclusion: Despite the effectiveness of federated learning in the utilization of private data across multiple medical institutions, the large number of clients and heterogeneous distribution of labels deteriorate the performance of both algorithms. Notably, FedProx appears to be more robust to the increased heterogeneity.

link: http://arxiv.org/abs/2402.10035v1

Predictive Linear Online Tracking for Unknown Targets

Anastasios Tsiamis, Aren Karapetyan, Yueshan Li, Efe C. Balta, John Lygeros

In this paper, we study the problem of online tracking in linear control systems, where the objective is to follow a moving target. Unlike classical tracking control, the target is unknown, non-stationary, and its state is revealed sequentially, thus, fitting the framework of online non-stochastic control. We consider the case of quadratic costs and propose a new algorithm, called predictive linear online tracking (PLOT). The algorithm uses recursive least squares with exponential forgetting to learn a time-varying dynamic model of the target. The learned model is used in the optimal policy under the framework of receding horizon control. We show the dynamic regret of PLOT scales with $\mathcal{O}(\sqrt{TV_T})$, where V_T is the total variation of the target dynamics and T is the time horizon. Unlike prior work, our theoretical results hold for non-stationary targets. We implement PLOT on a real quadrotor and provide open-source software, thus, showcasing one of the first successful applications of online control methods on real hardware.

link: http://arxiv.org/abs/2402.10036v1

RS-DPO: A Hybrid Rejection Sampling and Direct Preference Optimization Method for Alignment of Large Language Models

Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, Prathap Ramachandra

Reinforcement learning from human feedback (RLHF) has been extensively employed to align large language models with user intent. However, proximal policy optimization (PPO) based RLHF is occasionally unstable requiring significant hyperparameter finetuning, and computationally expensive to maximize the estimated reward during alignment. Recently, direct preference optimization (DPO) is proposed to address those challenges. However, DPO relies on contrastive responses generated from human annotator and alternative LLM, instead of the policy model, limiting the effectiveness of the RLHF. In this paper, we addresses both challenges by systematically combining rejection sampling (RS) and DPO. Our proposed method, RS-DPO, initiates with the development of a supervised fine-tuned policy model (SFT). A varied set of k

responses per prompt are sampled directly from the SFT model. RS-DPO identifies pairs of contrastive samples based on their reward distribution. Finally, we apply DPO with the contrastive samples to align the model to human preference. Our experiments indicate that our proposed method effectively fine-tunes LLMs with limited resource environments, leading to improved alignment with user intent. Furthermore, it outperforms existing methods, including RS, PPO, and DPO.

link: http://arxiv.org/abs/2402.10038v1

Feature Accentuation: Revealing 'What' Features Respond to in Natural Images

Chris Hamblin, Thomas Fel, Srijani Saha, Talia Konkle, George Alvarez

Efforts to decode neural network vision models necessitate a comprehensive grasp of both the spatial and semantic facets governing feature responses within images. Most research has primarily centered around attribution methods, which provide explanations in the form of heatmaps, showing where the model directs its attention for a given feature. However, grasping 'where' alone falls short, as numerous studies have highlighted the limitations of those methods and the necessity to understand 'what' the model has recognized at the focal point of its attention. In parallel, 'Feature visualization' offers another avenue for interpreting neural network features. This approach synthesizes an optimal image through gradient ascent, providing clearer insights into 'what' features respond to. However, feature visualizations only provide one global explanation per feature; they do not explain why features activate for particular images. In this work, we introduce a new method to the interpretability tool-kit, 'feature accentuation', which is capable of conveying both where and what in arbitrary input images induces a feature's response. At its core, feature accentuation is image-seeded (rather than noise-seeded) feature visualization. We find a particular combination of parameterization, augmentation, and regularization yields naturalistic visualizations that resemble the seed image and target feature simultaneously. Furthermore, we validate these accentuations are processed along a natural circuit by the model. We make our precise implementation of feature accentuation available to the community as the Faccent library, an extension of Lucent.

link: http://arxiv.org/abs/2402.10039v1

How to validate average calibration for machine learning regression tasks ?

Pascal Pernot

Average calibration of the uncertainties of machine learning regression tasks can be tested in two ways. One way is to estimate the calibration error (CE) as the difference between the mean absolute error (MSE) and the mean variance (MV) or mean squared uncertainty. The alternative is to compare the mean squared z-scores or scaled errors (ZMS) to 1. Both approaches might lead to different conclusion, as illustrated on an ensemble of datasets from the recent machine learning uncertainty quantification literature. It is shown here that the CE is very sensitive to the distribution of uncertainties, and notably to the presence of outlying uncertainties, and that it cannot be used reliably for calibration testing. By contrast, the ZMS statistic does not present this sensitivity issue and offers the most reliable approach in this context. Implications for the validation of conditional calibration are discussed.

link: http://arxiv.org/abs/2402.10043v1