Sat 2024.05.04

Mixture of insighTful Experts (MoTE): The Synergy of Thought Chains and Expert Mixtures in Self-Alignment

Zhili Liu, Yunhao Gou, Kai Chen, Lanqing Hong, Jiahui Gao, Fei Mi, Yu Zhang, Zhenguo Li, Xin Jiang, Qun Liu, James T. Kwok

As the capabilities of large language models (LLMs) have expanded dramatically, aligning these models with human values presents a significant challenge, posing potential risks during deployment. Traditional alignment strategies rely heavily on human intervention, such as Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), or on the self-alignment capacities of LLMs, which usually require a strong LLM's emergent ability to improve its original bad answer. To address these challenges, we propose a novel self-alignment method that utilizes a Chain of Thought (CoT) approach, termed AlignCoT. This method encompasses stages of Question Analysis, Answer Guidance, and Safe Answer production. It is designed to enable LLMs to generate high-quality, safe responses throughout various stages of their development. Furthermore, we introduce the Mixture of insighTful Experts (MoTE) architecture, which applies the mixture of experts to enhance each component of the AlignCoT process, markedly increasing alignment efficiency. The MoTE approach not only outperforms existing methods in aligning LLMs with human values but also highlights the benefits of using self-generated data, revealing the dual benefits of improved alignment and training efficiency.

link: http://arxiv.org/abs/2405.00557v1

Cross-Cluster Networking to Support Extended Reality Services

Theodoros Theodoropoulos, Luis Rosa, Abderrahmane Boudi, Tarik Zakaria Benmerar, Antonios Makris, Tarik Taleb, Luis Cordeiro, Konstantinos Tserpes, JaeSeung Song

Extented Reality (XR) refers to a class of contemporary services that are intertwined with a plethora of rather demanding Quality of Service (QoS) and functional requirements. Despite Kubernetes being the de-facto standard in terms of deploying and managing contemporary containerized microservices, it lacks adequate support for cross-cluster networking, hindering service-to-service communication across diverse cloud domains. Although there are tools that may be leveraged alongside Kubernetes in order to establish multi-cluster deployments, each one of them comes with its drawbacks and limitations. The purpose of this article is to explore the various potential technologies that may facilitate multi-cluster deployments and to propose how they may be leveraged to provide a cross-cluster connectivity solution that caters to the intricacies of XR services. The proposed solution is based on the use of two open source frameworks, namely Cluster API for multi-cluster management, and Liqo for multi-cluster interconnectivity. The efficiency of this approach is evaluated in the context of two experiments. This work is the first attempt at proposing a solution for supporting multi-cluster deployments in a manner that is aligned with the requirements of XR services

link: http://arxiv.org/abs/2405.00558v1

NumLLM: Numeric-Sensitive Large Language Model for Chinese Finance

Huan-Yi Su, Ke Wu, Yu-Hao Huang, Wu-Jun Li

Recently, many works have proposed various financial large language models (FinLLMs) by pre-training from scratch or fine-tuning open-sourced LLMs on financial corpora. However, existing FinLLMs exhibit unsatisfactory performance in understanding financial text when numeric variables are involved in questions. In this paper, we propose a novel LLM, called numeric-sensitive large language model (NumLLM), for Chinese finance. We first construct a financial corpus from financial textbooks which is essential for improving numeric capability of LLMs during fine-tuning. After that, we train two individual low-rank adaptation (LoRA) modules by fine-tuning on our constructed financial corpus. One module is for adapting general-purpose LLMs to financial domain, and the other module is for enhancing the ability of NumLLM to understand financial text with numeric

variables. Lastly, we merge the two LoRA modules into the foundation model to obtain NumLLM for inference. Experiments on financial question-answering benchmark show that NumLLM can boost the performance of the foundation model and can achieve the best overall performance compared to all baselines, on both numeric and non-numeric questions.

link: http://arxiv.org/abs/2405.00566v1

Powering In-Database Dynamic Model Slicing for Structured Data Analytics

Lingze Zeng, Naili Xing, Shaofeng Cai, Gang Chen, Beng Chin Ooi, Jian Pei, Yuncheng Wu

Relational database management systems (RDBMS) are widely used for the storage and retrieval of structured data. To derive insights beyond statistical aggregation, we typically have to extract specific subdatasets from the database using conventional database operations, and then apply deep neural networks (DNN) training and inference on these respective subdatasets in a separate machine learning system. The process can be prohibitively expensive, especially when there are a combinatorial number of subdatasets extracted for different analytical purposes. This calls for efficient in-database support of advanced analytical methods In this paper, we introduce LEADS, a novel SQL-aware dynamic model slicing technique to customize models for subdatasets specified by SQL queries. LEADS improves the predictive modeling of structured data via the mixture of experts (MoE) technique and maintains inference efficiency by a SQL-aware gating network. At the core of LEADS is the construction of a general model with multiple expert sub-models via MoE trained over the entire database. This SQL-aware MoE technique scales up the modeling capacity, enhances effectiveness, and preserves efficiency by activating only necessary experts via the gating network during inference. Additionally, we introduce two regularization terms during the training process of LEADS to strike a balance between effectiveness and efficiency. We also design and build an in-database inference system, called INDICES, to support end-to-end advanced structured data analytics by non-intrusively incorporating LEADS onto PostgreSQL. Our extensive experiments on real-world datasets demonstrate that LEADS consistently outperforms baseline models, and INDICES delivers effective in-database analytics with a considerable reduction in inference latency compared to traditional solutions.

link: http://arxiv.org/abs/2405.00568v1

WEST GCN-LSTM: Weighted Stacked Spatio-Temporal Graph Neural Networks for Regional Traffic Forecasting

Theodoros Theodoropoulos, Angelos-Christos Maroudis, Antonios Makris, Konstantinos Tserpes

Regional traffic forecasting is a critical challenge in urban mobility, with applications to various fields such as the Internet of Everything. In recent years, spatio-temporal graph neural networks have achieved state-of-the-art results in the context of numerous traffic forecasting challenges. This work aims at expanding upon the conventional spatio-temporal graph neural network architectures in a manner that may facilitate the inclusion of information regarding the examined regions, as well as the populations that traverse them, in order to establish a more efficient prediction model. The end-product of this scientific endeavour is a novel spatio-temporal graph neural network architecture that is referred to as WEST (WEighted STacked) GCN-LSTM. Furthermore, the inclusion of the aforementioned information is conducted via the use of two novel dedicated algorithms that are referred to as the Shared Borders Policy and the Adjustable Hops Policy. Through information fusion and distillation, the proposed solution manages to significantly outperform its competitors in the frame of an experimental evaluation that consists of 19 forecasting models, across several datasets. Finally, an additional ablation study determined that each of the components of the proposed solution contributes towards enhancing its overall performance.

link: http://arxiv.org/abs/2405.00570v1

Spherical Linear Interpolation and Text-Anchoring for Zero-shot Composed Image Retrieval

Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, Ser-Nam Lim

Composed Image Retrieval (CIR) is a complex task that retrieves images using a query, which is configured with an image and a caption that describes desired modifications to that image. Supervised CIR approaches have shown strong performance, but their reliance on expensive manually-annotated datasets restricts their scalability and broader applicability. To address these issues, previous studies have proposed pseudo-word token-based Zero-Shot CIR (ZS-CIR) methods, which utilize a projection module to map images to word tokens. However, we conjecture that this approach has a downside: the projection module distorts the original image representation and confines the resulting composed embeddings to the text-side. In order to resolve this, we introduce a novel ZS-CIR method that uses Spherical Linear Interpolation (Slerp) to directly merge image and text representations by identifying an intermediate embedding of both. Furthermore, we introduce Text-Anchored-Tuning (TAT), a method that fine-tunes the image encoder while keeping the text encoder fixed. TAT closes the modality gap between images and text, making the Slerp process much more effective. Notably, the TAT method is not only efficient in terms of the scale of the training dataset and training time, but it also serves as an excellent initial checkpoint for training supervised CIR models, thereby highlighting its wider potential. The integration of the Slerp-based ZS-CIR with a TAT-tuned model enables our approach to deliver state-of-the-art retrieval performance across CIR benchmarks.

link: http://arxiv.org/abs/2405.00571v1

EALD-MLLM: Emotion Analysis in Long-sequential and De-identity videos with Multi-modal Large Language Model

Deng Li, Xin Liu, Bohao Xing, Baiqiang Xia, Yuan Zong, Bihan Wen, Heikki Kälviäinen

Emotion AI is the ability of computers to understand human emotional states. Existing works have achieved promising progress, but two limitations remain to be solved: 1) Previous studies have been more focused on short sequential video emotion analysis while overlooking long sequential video. However, the emotions in short sequential videos only reflect instantaneous emotions, which may be deliberately guided or hidden. In contrast, long sequential videos can reveal authentic emotions; 2) Previous studies commonly utilize various signals such as facial, speech, and even sensitive biological signals (e.g., electrocardiogram). However, due to the increasing demand for privacy, developing Emotion AI without relying on sensitive signals is becoming important. To address the aforementioned limitations, in this paper, we construct a dataset for Emotion Analysis in Long-sequential and De-identity videos called EALD by collecting and processing the sequences of athletes' post-match interviews. In addition to providing annotations of the overall emotional state of each video, we also provide the Non-Facial Body Language (NFBL) annotations for each player. NFBL is an inner-driven emotional expression and can serve as an identity-free clue to understanding the emotional state. Moreover, we provide a simple but effective baseline for further research. More precisely, we evaluate the Multimodal Large Language Models (MLLMs) with de-identification signals (e.g., visual, speech, and NFBLs) to perform emotion analysis. Our experimental results demonstrate that: 1) MLLMs can achieve comparable, even better performance than the supervised single-modal models, even in a zero-shot scenario; 2) NFBL is an important cue in long sequential emotion analysis. EALD will be available on the open-source platform.

link: http://arxiv.org/abs/2405.00574v1

The Real, the Better: Aligning Large Language Models with Online Human Behaviors

Guanying Jiang, Lingyong Yan, Haibo Shi, Dawei Yin

Large language model alignment is widely used and studied to avoid LLM producing unhelpful and harmful responses. However, the lengthy training process and predefined preference bias hinder adaptation to online diverse human preferences. To this end, this paper proposes an alignment framework, called Reinforcement Learning with Human Behavior (RLHB), to align LLMs by directly leveraging real online human behaviors. By taking the generative adversarial framework, the generator is trained to respond following expected human behavior; while the discriminator tries to verify whether the triplets of query, response, and human behavior come from real online

environments. Behavior modeling in natural-language form and the multi-model joint training mechanism enable an active and sustainable online alignment. Experimental results confirm the effectiveness of our proposed methods by both human and automatic evaluations.

link: http://arxiv.org/abs/2405.00578v1

GraCo: Granularity-Controllable Interactive Segmentation

Yian Zhao, Kehan Li, Zesen Cheng, Pengchong Qiao, Xiawu Zheng, Rongrong Ji, Chang Liu, Li Yuan, Jie Chen

Interactive Segmentation (IS) segments specific objects or parts in the image according to user input. Current IS pipelines fall into two categories: single-granularity output and multi-granularity output. The latter aims to alleviate the spatial ambiguity present in the former. However, the multi-granularity output pipeline suffers from limited interaction flexibility and produces redundant results. In this work, we introduce Granularity-Controllable Interactive Segmentation (GraCo), a novel approach that allows precise control of prediction granularity by introducing additional parameters to input. This enhances the customization of the interactive system and eliminates redundancy while resolving ambiguity. Nevertheless, the exorbitant cost of annotating multi-granularity masks and the lack of available datasets with granularity annotations make it difficult for models to acquire the necessary guidance to control output granularity. To address this problem, we design an any-granularity mask generator that exploits the semantic property of the pre-trained IS model to automatically generate abundant mask-granularity pairs without requiring additional manual annotation. Based on these pairs, we propose a granularity-controllable learning strategy that efficiently imparts the granularity controllability to the IS model. Extensive experiments on intricate scenarios at object and part levels demonstrate that our GraCo has significant advantages over previous methods. This highlights the potential of GraCo to be a flexible annotation tool, capable of adapting to diverse segmentation scenarios. The project page: https://zhao-yian.github.io/GraCo.

link: http://arxiv.org/abs/2405.00587v1

Are Models Biased on Text without Gender-related Language?

Catarina G Belém, Preethi Seshadri, Yasaman Razeghi, Sameer Singh

Gender bias research has been pivotal in revealing undesirable behaviors in large language models, exposing serious gender stereotypes associated with occupations, and emotions. A key observation in prior work is that models reinforce stereotypes as a consequence of the gendered correlations that are present in the training data. In this paper, we focus on bias where the effect from training data is unclear, and instead address the question: Do language models still exhibit gender bias in non-stereotypical settings? To do so, we introduce UnStereoEval (USE), a novel framework tailored for investigating gender bias in stereotype-free scenarios. USE defines a sentence-level score based on pretraining data statistics to determine if the sentence contain minimal word-gender associations. To systematically benchmark the fairness of popular language models in stereotype-free scenarios, we utilize USE to automatically generate benchmarks without any gender-related language. By leveraging USE's sentence-level score, we also repurpose prior gender bias benchmarks (Winobias and Winogender) for non-stereotypical evaluation. Surprisingly, we find low fairness across all 28 tested models. Concretely, models demonstrate fair behavior in only 9%-41% of stereotype-free sentences, suggesting that bias does not solely stem from the presence of gender-related words. These results raise important questions about where underlying model biases come from and highlight the need for more systematic and comprehensive bias evaluation. We release the full dataset and code at https://ucinlp.github.io/unstereo-eval.

link: http://arxiv.org/abs/2405.00588v1

How founder motivations, goals, and actions influence early trajectories of online communities

Sanjay R. Kairam, Jeremy Foote

Online communities offer their members various benefits, such as information access, social and emotional support, and entertainment. Despite the important role that founders play in shaping communities, prior research has focused primarily on what drives users to participate and contribute; the motivations and goals of founders remain underexplored. To uncover how and why online communities get started, we present findings from a survey of 951 recent founders of Reddit communities. We find that topical interest is the most common motivation for community creation, followed by motivations to exchange information, connect with others, and self-promote. Founders have heterogeneous goals for their nascent communities, but they tend to privilege community quality and engagement over sheer growth. These differences in founders' early attitudes towards their communities help predict not only the community-building actions that they pursue, but also the ability of their communities to attract visitors, contributors, and subscribers over the first 28 days. We end with a discussion of the implications for researchers, designers, and founders of online communities.

link: http://dx.doi.org/10.1145/3613904.3642269

Investigating Automatic Scoring and Feedback using Large Language Models

Gloria Ashiya Katuka, Alexander Gain, Yen-Yun Yu

Automatic grading and feedback have been long studied using traditional machine learning and deep learning techniques using language models. With the recent accessibility to high performing large language models (LLMs) like LLaMA-2, there is an opportunity to investigate the use of these LLMs for automatic grading and feedback generation. Despite the increase in performance, LLMs require significant computational resources for fine-tuning and additional specific adjustments to enhance their performance for such tasks. To address these issues, Parameter Efficient Fine-tuning (PEFT) methods, such as LoRA and QLoRA, have been adopted to decrease memory and computational requirements in model fine-tuning. This paper explores the efficacy of PEFT-based quantized models, employing classification or regression head, to fine-tune LLMs for automatically assigning continuous numerical grades to short answers and essays, as well as generating corresponding feedback. We conducted experiments on both proprietary and open-source datasets for our tasks. The results show that prediction of grade scores via finetuned LLMs are highly accurate, achieving less than 3% error in grade percentage on average. For providing graded feedback fine-tuned 4-bit quantized LLaMA-2 13B models outperform competitive base models and achieve high similarity with subject matter expert feedback in terms of high BLEU and ROUGE scores and qualitatively in terms of feedback. The findings from this study provide important insights into the impacts of the emerging capabilities of using quantization approaches to fine-tune LLMs for various downstream tasks, such as automatic short answer scoring and feedback generation at comparatively lower costs and latency.

link: http://arxiv.org/abs/2405.00602v1

A Preprocessing and Evaluation Toolbox for Trajectory Prediction Research on the Drone Datasets

Theodor Westny, Björn Olofsson, Erik Frisk

The availability of high-quality datasets is crucial for the development of behavior prediction algorithms in autonomous vehicles. This paper highlights the need for standardizing the use of certain datasets for motion forecasting research to simplify comparative analysis and proposes a set of tools and practices to achieve this. Drawing on extensive experience and a comprehensive review of current literature, we summarize our proposals for preprocessing, visualizing, and evaluation in the form of an open-sourced toolbox designed for researchers working on trajectory prediction problems. The clear specification of necessary preprocessing steps and evaluation metrics is intended to alleviate development efforts and facilitate the comparison of results across different studies. The toolbox is available at: https://github.com/westny/dronalize.

link: http://arxiv.org/abs/2405.00604v1

Addressing Topic Granularity and Hallucination in Large Language Models for Topic Modelling

Yida Mu, Peizhen Bai, Kalina Bontcheva, Xingyi Song

Large language models (LLMs) with their strong zero-shot topic extraction capabilities offer an alternative to probabilistic topic modelling and closed-set topic classification approaches. As zero-shot topic extractors, LLMs are expected to understand human instructions to generate relevant and non-hallucinated topics based on the given documents. However, LLM-based topic modelling approaches often face difficulties in generating topics with adherence to granularity as specified in human instructions, often resulting in many near-duplicate topics. Furthermore, methods for addressing hallucinated topics generated by LLMs have not yet been investigated. In this paper, we focus on addressing the issues of topic granularity and hallucinations for better LLM-based topic modelling. To this end, we introduce a novel approach that leverages Direct Preference Optimisation (DPO) to fine-tune open-source LLMs, such as Mistral-7B. Our approach does not rely on traditional human annotation to rank preferred answers but employs a reconstruction pipeline to modify raw topics generated by LLMs, thus enabling a fast and efficient training and inference framework. Comparative experiments show that our fine-tuning approach not only significantly improves the LLM's capability to produce more coherent, relevant, and precise topics, but also reduces the number of hallucinated topics.

link: http://arxiv.org/abs/2405.00611v1

Multigroup Robustness

Lunjia Hu, Charlotte Peale, Judy Hanwen Shen

To address the shortcomings of real-world datasets, robust learning algorithms have been designed to overcome arbitrary and indiscriminate data corruption. However, practical processes of gathering data may lead to patterns of data corruption that are localized to specific partitions of the training dataset. Motivated by critical applications where the learned model is deployed to make predictions about people from a rich collection of overlapping subpopulations, we initiate the study of multigroup robust algorithms whose robustness guarantees for each subpopulation only degrade with the amount of data corruption inside that subpopulation. When the data corruption is not distributed uniformly over subpopulations, our algorithms provide more meaningful robustness guarantees than standard guarantees that are oblivious to how the data corruption and the affected subpopulations are related. Our techniques establish a new connection between multigroup fairness and robustness.

link: http://arxiv.org/abs/2405.00614v1