Wed 2024.02.21

Zero shot VLMs for hate meme detection: Are we there yet?

Naquee Rizwan, Paramananda Bhaskar, Mithun Das, Swadhin Satyaprakash Majhi, Punyajoy Saha, Animesh Mukherjee

Multimedia content on social media is rapidly evolving, with memes gaining prominence as a distinctive form. Unfortunately, some malicious users exploit memes to target individuals or vulnerable communities, making it imperative to identify and address such instances of hateful memes. Extensive research has been conducted to address this issue by developing hate meme detection models. However, a notable limitation of traditional machine/deep learning models is the requirement for labeled datasets for accurate classification. Recently, the research community has witnessed the emergence of several visual language models that have exhibited outstanding performance across various tasks. In this study, we aim to investigate the efficacy of these visual language models in handling intricate tasks such as hate meme detection. We use various prompt settings to focus on zero-shot classification of hateful/harmful memes. Through our analysis, we observe that large VLMs are still vulnerable for zero-shot hate meme detection.

link: http://arxiv.org/abs/2402.12198v1

Dictionary Learning Improves Patch-Free Circuit Discovery in Mechanistic Interpretability: A Case Study on Othello-GPT

Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, Xipeng Qiu

Sparse dictionary learning has been a rapidly growing technique in mechanistic interpretability to attack superposition and extract more human-understandable features from model activations. We ask a further question based on the extracted more monosemantic features: How do we recognize circuits connecting the enormous amount of dictionary features? We propose a circuit discovery framework alternative to activation patching. Our framework suffers less from out-of-distribution and proves to be more efficient in terms of asymptotic complexity. The basic unit in our framework is dictionary features decomposed from all modules writing to the residual stream, including embedding, attention output and MLP output. Starting from any logit, dictionary feature or attention score, we manage to trace down to lower-level dictionary features of all tokens and compute their contribution to these more interpretable and local model behaviors. We dig in a small transformer trained on a synthetic task named Othello and find a number of human-understandable fine-grained circuits inside of it.

link: http://arxiv.org/abs/2402.12201v1

Heterogeneity-aware Cross-school Electives Recommendation: a Hybrid Federated Approach

Chengyi Ju, Jiannong Cao, Yu Yang, Zhen-Qun Yang, Ho Man Lee

In the era of modern education, addressing cross-school learner diversity is crucial, especially in personalized recommender systems for elective course selection. However, privacy concerns often limit cross-school data sharing, which hinders existing methods' ability to model sparse data and address heterogeneity effectively, ultimately leading to suboptimal recommendations. In response, we propose HFRec, a heterogeneity-aware hybrid federated recommender system designed for cross-school elective course recommendations. The proposed model constructs heterogeneous graphs for each school, incorporating various interactions and historical behaviors between students to integrate context and content information. We design an attention mechanism to capture heterogeneity-aware representations. Moreover, under a federated scheme, we train individual school-based models with adaptive learning settings to recommend tailored electives. Our HFRec model demonstrates its effectiveness in providing personalized elective recommendations while maintaining privacy, as it outperforms state-of-the-art models on both open-source and real-world datasets.

link: http://dx.doi.org/10.1109/ICDMW60847.2023.00191

MPI Implementation Profiling for Better Application Performance

Riley Shipley, Garrett Hooten, David Boehme, Derek Schafer, Anthony Skjellum, Olga Pearce

While application profiling has been a mainstay in the HPC community for years, profiling of MPI and other communication middleware has not received the same degree of exploration. This paper adds to the discussion of MPI profiling, contributing two general-purpose profiling methods as well as practical applications of these methods to an existing implementation. The ability to detect performance defects in MPI codes using these methods increases the potential of further research and development in communication optimization.

link: http://arxiv.org/abs/2402.12203v1

Enhancing Multilingual Capabilities of Large Language Models through Self-Distillation from Resource-Rich Languages

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, Yang Liu

While large language models (LLMs) have been pre-trained on multilingual corpora, their performance still lags behind in most languages compared to a few resource-rich languages. One common approach to mitigate this issue is to translate training data from resource-rich languages into other languages and then continue training. However, using the data obtained solely relying on translation while ignoring the original capabilities of LLMs across languages is not always effective, which we show will limit the performance of cross-lingual knowledge transfer. In this work, we propose SDRRL, a method based on Self-Distillation from Resource-Rich Languages that effectively improve multilingual performance by leveraging the internal capabilities of LLMs on resource-rich languages. We evaluate on different LLMs (LLaMA-2 and SeaLLM) and source languages across various comprehension and generation tasks, experimental results demonstrate that SDRRL can significantly enhance multilingual capabilities while minimizing the impact on original performance in resource-rich languages.

link: http://arxiv.org/abs/2402.12204v1

Polarization of Autonomous Generative AI Agents Under Echo Chambers

Masaya Ohagi

Online social networks often create echo chambers where people only hear opinions reinforcing their beliefs. An echo chamber often generates polarization, leading to conflicts caused by people with radical opinions, such as the January 6, 2021, attack on the US Capitol. The echo chamber has been viewed as a human-specific problem, but this implicit assumption is becoming less reasonable as large language models, such as ChatGPT, acquire social abilities. In response to this situation, we investigated the potential for polarization to occur among a group of autonomous AI agents based on generative language models in an echo chamber environment. We had AI agents discuss specific topics and analyzed how the group's opinions changed as the discussion progressed. As a result, we found that the group of agents based on ChatGPT tended to become polarized in echo chamber environments. The analysis of opinion transitions shows that this result is caused by ChatGPT's high prompt understanding ability to update its opinion by considering its own and surrounding agents' opinions. We conducted additional experiments to investigate under what specific conditions AI agents tended to polarize. As a result, we identified factors that strongly influence polarization, such as the agent's persona. These factors should be monitored to prevent the polarization of AI agents.

link: http://arxiv.org/abs/2402.12212v1

Copyleft for Alleviating AIGC Copyright Dilemma: What-if Analysis, Public Perception and Implications

Xinwei Guo, Yujun Li, Yafeng Peng, Xuetao Wei

As AIGC has impacted our society profoundly in the past years, ethical issues have received tremendous attention. The most urgent one is the AIGC copyright dilemma, which can immensely stifle the development of AIGC and greatly cost the entire society. Given the complexity of AIGC copyright governance and the fact that no perfect solution currently exists, previous work advocated copyleft on AI governance but without substantive analysis. In this paper, we take a step further to explore the feasibility of copyleft to alleviate the AIGC copyright dilemma. We conduct a mixed-methods study from two aspects: qualitatively, we use a formal what-if analysis to clarify the dilemma and provide case studies to show the feasibility of copyleft; quantitatively, we perform a carefully designed survey to find out how the public feels about copylefting AIGC. The key findings include: a) people generally perceive the dilemma, b) they prefer to use authorized AIGC under loose restriction, and c) they are positive to copyleft in AIGC and willing to use it in the future.

link: http://arxiv.org/abs/2402.12216v1

Reformatted Alignment

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, Pengfei Liu

The quality of finetuning data is crucial for aligning large language models (LLMs) with human values. Current methods to improve data quality are either labor-intensive or prone to factual errors caused by LLM hallucinations. This paper explores elevating the quality of existing instruction data to better align with human values, introducing a simple and effective approach named ReAlign, which reformats the responses of instruction data into a format that better aligns with pre-established criteria and the collated evidence. This approach minimizes human annotation, hallucination, and the difficulty in scaling, remaining orthogonal to existing alignment techniques. Experimentally, ReAlign significantly boosts the general alignment ability, math reasoning, factuality, and readability of the LLMs. Encouragingly, without introducing any additional data or advanced training techniques, and merely by reformatting the response, LLaMA-2-13B's mathematical reasoning ability on GSM8K can be improved from 46.77% to 56.63% in accuracy. Additionally, a mere 5% of ReAlign data yields a 67% boost in general alignment ability measured by the Alpaca dataset. This work highlights the need for further research into the science and mechanistic interpretability of LLMs. We have made the associated code and data publicly accessible to support future studies at https://github.com/GAIR-NLP/ReAlign.

link: http://arxiv.org/abs/2402.12219v1

Bayesian Parameter-Efficient Fine-Tuning for Overcoming Catastrophic Forgetting

Haolin Chen, Philip N. Garner

Although motivated by the adaptation of text-to-speech synthesis models, we argue that more generic parameter-efficient fine-tuning (PEFT) is an appropriate framework to do such adaptation. However, catastrophic forgetting remains an issue with PEFT, damaging the pre-trained model's inherent capabilities. We demonstrate that existing Bayesian learning techniques can be applied to PEFT to prevent catastrophic forgetting as long as the parameter shift of the fine-tuned layers can be calculated differentiably. In a principled series of experiments on language modeling and speech synthesis tasks, we utilize established Laplace approximations, including diagonal and Kronecker factored approaches, to regularize PEFT with the low-rank adaptation (LoRA) and compare their performance in pre-training knowledge preservation. Our results demonstrate that catastrophic forgetting can be overcome by our methods without degrading the fine-tuning performance, and using the Kronecker factored approximations produces a better preservation of the pre-training knowledge than the diagonal ones.

link: http://arxiv.org/abs/2402.12220v1

CovRL: Fuzzing JavaScript Engines with Coverage-Guided Reinforcement Learning for LLM-based Mutation

Jueon Eom, Seyeon Jeong, Taekyoung Kwon

Fuzzing is an effective bug-finding technique but it struggles with complex systems like JavaScript engines that demand precise grammatical input. Recently, researchers have adopted language

models for context-aware mutation in fuzzing to address this problem. However, existing techniques are limited in utilizing coverage guidance for fuzzing, which is rather performed in a black-box manner. This paper presents a novel technique called CovRL (Coverage-guided Reinforcement Learning) that combines Large Language Models (LLMs) with reinforcement learning from coverage feedback. Our fuzzer, CovRL-Fuzz, integrates coverage feedback directly into the LLM by leveraging the Term Frequency-Inverse Document Frequency (TF-IDF) method to construct a weighted coverage map. This map is key in calculating the fuzzing reward, which is then applied to the LLM-based mutator through reinforcement learning. CovRL-Fuzz, through this approach, enables the generation of test cases that are more likely to discover new coverage areas, thus improving vulnerability detection while minimizing syntax and semantic errors, all without needing extra post-processing. Our evaluation results indicate that CovRL-Fuzz outperforms the state-of-the-art fuzzers in terms of code coverage and bug-finding capabilities: CovRL-Fuzz identified 48 real-world security-related bugs in the latest JavaScript engines, including 39 previously unknown vulnerabilities and 11 CVEs.

link: http://arxiv.org/abs/2402.12222v1

Pushing Auto-regressive Models for 3D Shape Generation at Capacity and Scalability

Xuelin Qian, Yu Wang, Simian Luo, Yinda Zhang, Ying Tai, Zhenyu Zhang, Chengjie Wang, Xiangyang Xue, Bo Zhao, Tiejun Huang, Yunsheng Wu, Yanwei Fu

Auto-regressive models have achieved impressive results in 2D image generation by modeling joint distributions in grid space. In this paper, we extend auto-regressive models to 3D domains, and seek a stronger ability of 3D shape generation by improving auto-regressive models at capacity and scalability simultaneously. Firstly, we leverage an ensemble of publicly available 3D datasets to facilitate the training of large-scale models. It consists of a comprehensive collection of approximately 900,000 objects, with multiple properties of meshes, points, voxels, rendered images, and text captions. This diverse labeled dataset, termed Objaverse-Mix, empowers our model to learn from a wide range of object variations. However, directly applying 3D auto-regression encounters critical challenges of high computational demands on volumetric grids and ambiguous auto-regressive order along grid dimensions, resulting in inferior quality of 3D shapes. To this end, we then present a novel framework Argus3D in terms of capacity. Concretely, our approach introduces discrete representation learning based on a latent vector instead of volumetric grids, which not only reduces computational costs but also preserves essential geometric details by learning the joint distributions in a more tractable order. The capacity of conditional generation can thus be realized by simply concatenating various conditioning inputs to the latent vector, such as point clouds, categories, images, and texts. In addition, thanks to the simplicity of our model architecture, we naturally scale up our approach to a larger model with an impressive 3.6 billion parameters, further enhancing the quality of versatile 3D generation. Extensive experiments on four generation tasks demonstrate that Argus3D can synthesize diverse and faithful shapes across multiple categories, achieving remarkable performance.

link: http://arxiv.org/abs/2402.12225v1

AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, Xipeng Qiu

We introduce AnyGPT, an any-to-any multimodal language model that utilizes discrete representations for the unified processing of various modalities, including speech, text, images, and music. AnyGPT can be trained stably without any alterations to the current large language model (LLM) architecture or training paradigms. Instead, it relies exclusively on data-level preprocessing, facilitating the seamless integration of new modalities into LLMs, akin to the incorporation of new languages. We build a multimodal text-centric dataset for multimodal alignment pre-training. Utilizing generative models, we synthesize the first large-scale any-to-any multimodal instruction dataset. It consists of 108k samples of multi-turn conversations that intricately interweave various modalities, thus equipping the model to handle arbitrary combinations of multimodal inputs and

outputs. Experimental results demonstrate that AnyGPT is capable of facilitating any-to-any multimodal conversation while achieving performance comparable to specialized models across all modalities, proving that discrete representations can effectively and conveniently unify multiple modalities within a language model. Demos are shown in https://junzhan2000.github.io/AnyGPT.github.io/

link: http://arxiv.org/abs/2402.12226v1

Diffusion Tempering Improves Parameter Estimation with Probabilistic Integrators for Ordinary Differential Equations

Jonas Beck, Nathanael Bosch, Michael Deistler, Kyra L. Kadhim, Jakob H. Macke, Philipp Hennig, Philipp Berens

Ordinary differential equations (ODEs) are widely used to describe dynamical systems in science, but identifying parameters that explain experimental measurements is challenging. In particular, although ODEs are differentiable and would allow for gradient-based parameter optimization, the nonlinear dynamics of ODEs often lead to many local minima and extreme sensitivity to initial conditions. We therefore propose diffusion tempering, a novel regularization technique for probabilistic numerical methods which improves convergence of gradient-based parameter optimization in ODEs. By iteratively reducing a noise parameter of the probabilistic integrator, the proposed method converges more reliably to the true parameters. We demonstrate that our method is effective for dynamical systems of different complexity and show that it obtains reliable parameter estimates for a Hodgkin-Huxley model with a practically relevant number of parameters.

link: http://arxiv.org/abs/2402.12231v1

Kernel KMeans clustering splits for end-to-end unsupervised decision trees

Louis Ohl, Pierre-Alexandre Mattei, Mickaël Leclercq, Arnaud Droit, Frédéric Precioso

Trees are convenient models for obtaining explainable predictions on relatively small datasets. Although there are many proposals for the end-to-end construction of such trees in supervised learning, learning a tree end-to-end for clustering without labels remains an open challenge. As most works focus on interpreting with trees the result of another clustering algorithm, we present here a novel end-to-end trained unsupervised binary tree for clustering: Kauri. This method performs a greedy maximisation of the kernel KMeans objective without requiring the definition of centroids. We compare this model on multiple datasets with recent unsupervised trees and show that Kauri performs identically when using a linear kernel. For other kernels, Kauri often outperforms the concatenation of kernel KMeans and a CART decision tree.

link: http://arxiv.org/abs/2402.12232v1

Empirical Study on Updating Key-Value Memories in Transformer Feed-forward Layers

Zihan Qiu, Zeyu Huang, Youcheng Huang, Jie Fu

The feed-forward networks (FFNs) in transformers are recognized as a group of key-value neural memories to restore abstract high-level knowledge. In this work, we conduct an empirical ablation study on updating keys (the 1st layer in the FFNs layer) or values (the 2nd layer in the FFNs layer). We compare those two methods in various knowledge editing and fine-tuning tasks of large language models to draw insights to understand FFNs further. Code is available at $\href{https://github.com/qiuzh20/Tuning-keys-v.s.-values}{this\,repo}$.

link: http://arxiv.org/abs/2402.12233v1