Sat 2024.04.06

On Few-Shot Prompting for Controllable Question-Answer Generation in Narrative Comprehension

Bernardo Leite, Henrique Lopes Cardoso

Question Generation aims to automatically generate questions based on a given input provided as context. A controllable question generation scheme focuses on generating questions with specific attributes, allowing better control. In this study, we propose a few-shot prompting strategy for controlling the generation of question-answer pairs from children's narrative texts. We aim to control two attributes: the question's explicitness and underlying narrative elements. With empirical evaluation, we show the effectiveness of controlling the generation process by employing few-shot prompting side by side with a reference model. Our experiments highlight instances where the few-shot strategy surpasses the reference model, particularly in scenarios such as semantic closeness evaluation and the diversity and coherency of question-answer pairs. However, these improvements are not always statistically significant. The code is publicly available at github.com/bernardoleite/few-shot-prompting-qg-control.

link: http://arxiv.org/abs/2404.02800v1

The RealHumanEval: Evaluating Large Language Models' Abilities to Support Programmers

Hussein Mozannar, Valerie Chen, Mohammed Alsobay, Subhro Das, Sebastian Zhao, Dennis Wei, Manish Nagireddy, Prasanna Sattigeri, Ameet Talwalkar, David Sontag

Evaluation of large language models (LLMs) for code has primarily relied on static benchmarks, including HumanEval (Chen et al., 2021), which measure the ability of LLMs to generate complete code that passes unit tests. As LLMs are increasingly used as programmer assistants, we study whether gains on existing benchmarks translate to gains in programmer productivity when coding with LLMs, including time spent coding. In addition to static benchmarks, we investigate the utility of preference metrics that might be used as proxies to measure LLM helpfulness, such as code acceptance or copy rates. To do so, we introduce RealHumanEval, a web interface to measure the ability of LLMs to assist programmers, through either autocomplete or chat support. We conducted a user study (N=213) using RealHumanEval in which users interacted with six LLMs of varying base model performance. Despite static benchmarks not incorporating humans-in-the-loop, we find that improvements in benchmark performance lead to increased programmer productivity; however gaps in benchmark versus human performance are not proportional -- a trend that holds across both forms of LLM support. In contrast, we find that programmer preferences do not correlate with their actual performance, motivating the need for better, human-centric proxy signals. We also open-source RealHumanEval to enable human-centric evaluation of new models and the study data to facilitate efforts to improve code models.

link: http://arxiv.org/abs/2404.02806v1

Generative-Contrastive Heterogeneous Graph Neural Network

Yu Wang, Lei Sang, Yi Zhang, Yiwen Zhang

Heterogeneous Graphs (HGs) can effectively model complex relationships in the real world by multi-type nodes and edges. In recent years, inspired by self-supervised learning, contrastive Heterogeneous Graphs Neural Networks (HGNNs) have shown great potential by utilizing data augmentation and discriminators for downstream tasks. However, data augmentation is still limited due to the discrete and abstract nature of graphs. To tackle the above limitations, we propose a novel \textit{Generative-Contrastive Heterogeneous Graph Neural Network (GC-HGNN)}. Specifically, we first propose a heterogeneous graph generative learning enhanced contrastive paradigm. This paradigm includes: 1) A contrastive view augmentation strategy by using masked autoencoder. 2) Position-aware and semantics-aware positive sample sampling strategy for generate hard negative samples. 3) A hierarchical contrastive learning strategy for capturing local and global information. Furthermore, the hierarchical contrastive learning and sampling strategies aim to constitute an enhanced discriminator under the generative-contrastive perspective. Finally, we compare our model with seventeen baselines on eight real-world datasets. Our model outperforms the latest contrastive and generative baselines on node classification and link prediction tasks. To reproduce our work, we have open-sourced our code at https://github.com/xxx.

link: http://arxiv.org/abs/2404.02810v1

GPU-Accelerated RSF Level Set Evolution for Large-Scale Microvascular Segmentation

Meher Niger, Helya Goharbavang, Taeyong Ahn, Emily K. Alley, Joshua D. Wythe, Guoning Chen, David Mayerich

Microvascular networks are challenging to model because these structures are currently near the diffraction limit for most advanced three-dimensional imaging modalities, including confocal and light sheet microscopy. This makes semantic segmentation difficult, because individual components of these networks fluctuate within the confines of individual pixels. Level set methods are ideally suited to solve this problem by providing surface and topological constraints on the resulting model, however these active contour techniques are extremely time intensive and impractical for terabyte-scale images. We propose a reformulation and implementation of the region-scalable fitting (RSF) level set model that makes it amenable to three-dimensional evaluation using both single-instruction multiple data (SIMD) and single-program multiple-data (SPMD) parallel processing. This enables evaluation of the level set equation on independent regions of the data set using graphics processing units (GPUs), making large-scale segmentation of high-resolution networks practical and inexpensive. We tested this 3D parallel RSF approach on multiple data sets acquired using state-of-the-art imaging techniques to acquire microvascular data, including micro-CT, light sheet fluorescence microscopy (LSFM) and milling microscopy. To assess the performance and accuracy of the RSF model, we conducted a Monte-Carlo-based validation technique to compare results to other segmentation methods. We also provide a rigorous profiling to show the gains in processing speed leveraging parallel hardware. This study showcases the practical application of the RSF model, emphasizing its utility in the challenging domain of segmenting large-scale high-topology network structures with a particular focus on building microvascular models.

link: http://arxiv.org/abs/2404.02813v1

A Survey of Optimization-based Task and Motion Planning: From Classical To Learning Approaches

Zhigen Zhao, Shuo Chen, Yan Ding, Ziyi Zhou, Shiqi Zhang, Danfei Xu, Ye Zhao

Task and Motion Planning (TAMP) integrates high-level task planning and low-level motion planning to equip robots with the autonomy to effectively reason over long-horizon, dynamic tasks. Optimization-based TAMP focuses on hybrid optimization approaches that define goal conditions via objective functions and are capable of handling open-ended goals, robotic dynamics, and physical interaction between the robot and the environment. Therefore, optimization-based TAMP is particularly suited to solve highly complex, contact-rich locomotion and manipulation problems. This survey provides a comprehensive review on optimization-based TAMP, covering (i) planning domain representations, including action description languages and temporal logic, (ii) individual solution strategies for components of TAMP, including AI planning and trajectory optimization (TO), and (iii) the dynamic interplay between logic-based task planning and model-based TO. A particular focus of this survey is to highlight the algorithm structures to efficiently solve TAMP, especially hierarchical and distributed approaches. Additionally, the survey emphasizes the synergy between the classical methods and contemporary learning-based innovations such as large language models. Furthermore, the future research directions for TAMP is discussed in this survey, highlighting both algorithmic and application-specific challenges.

link: http://arxiv.org/abs/2404.02817v1

Identifying Climate Targets in National Laws and Policies using Machine Learning

Matyas Juhasz, Tina Marchand, Roshan Melwani, Kalyan Dutia, Sarah Goodenough, Harrison Pim, Henry Franks

Quantified policy targets are a fundamental element of climate policy, typically characterised by domain-specific and technical language. Current methods for curating comprehensive views of global climate policy targets entail significant manual effort. At present there are few scalable methods for extracting climate targets from national laws or policies, which limits policymakers' and researchers' ability to (1) assess private and public sector alignment with global goals and (2) inform policy decisions. In this paper we present an approach for extracting mentions of climate targets from national laws and policies. We create an expert-annotated dataset identifying three categories of target ('Net Zero', 'Reduction' and 'Other' (e.g. renewable energy targets)) and train a classifier to reliably identify them in text. We investigate bias and equity impacts related to our model and identify specific years and country names as problematic features. Finally, we investigate the characteristics of the dataset produced by running this classifier on the Climate Policy Radar (CPR) dataset of global national climate laws and policies and UNFCCC submissions, highlighting the potential of automated and scalable data collection for existing climate policy databases and supporting further research. Our work represents a significant upgrade in the accessibility of these key climate policy elements for policymakers and researchers. We publish our model at https://huggingface.co/ClimatePolicyRadar/national-climate-targets and related dataset at https://huggingface.co/datasets/ClimatePolicyRadar/national-climate-targets.

link: http://arxiv.org/abs/2404.02822v2

Conifer: Improving Complex Constrained Instruction-Following Ability of Large Language Models

Haoran Sun, Lixin Liu, Junjie Li, Fengyu Wang, Baohua Dong, Ran Lin, Ruohui Huang

The ability of large language models (LLMs) to follow instructions is crucial to real-world applications. Despite recent advances, several studies have highlighted that LLMs struggle when faced with challenging instructions, especially those that include complex constraints, hindering their effectiveness in various tasks. To address this challenge, we introduce Conifer, a novel instruction tuning dataset, designed to enhance LLMs to follow multi-level instructions with complex constraints. Utilizing GPT-4, we curate the dataset by a series of LLM-driven refinement processes to ensure high quality. We also propose a progressive learning scheme that emphasizes an easy-to-hard progression, and learning from process feedback. Models trained with Conifer exhibit remarkable improvements in instruction-following abilities, especially for instructions with complex constraints. On several instruction-following benchmarks, our 7B model outperforms the state-of-the-art open-source 7B models, even exceeds the performance of models 10 times larger on certain metrics. All the code and Conifer dataset are available at https://www.github.com/ConiferLM/Conifer.

link: http://arxiv.org/abs/2404.02823v1

BAdam: A Memory Efficient Full Parameter Training Method for Large Language Models

Qijun Luo, Hengxu Yu, Xiao Li

This work presents BAdam, an optimizer that leverages the block coordinate optimization framework with Adam as the inner solver. BAdam offers a memory efficient approach to the full parameter finetuning of large language models and reduces running time of the backward process thanks to the chain rule property. Experimentally, we apply BAdam to instruction-tune the Llama 2-7B model on the Alpaca-GPT4 dataset using a single RTX3090-24GB GPU. The results indicate that BAdam exhibits superior convergence behavior in comparison to LoRA and LOMO. Furthermore, our downstream performance evaluation of the instruction-tuned models using the MT-bench shows that BAdam modestly surpasses LoRA and more substantially outperforms LOMO. Finally, we compare BAdam with Adam on a medium-sized task, i.e., finetuning RoBERTa-large on the SuperGLUE benchmark. The results demonstrate that BAdam is capable of narrowing the performance gap with Adam. Our code is available at https://github.com/Ledzy/BAdam.

link: http://arxiv.org/abs/2404.02827v1

Enhancing Interpretability of Vertebrae Fracture Grading using Human-interpretable Prototypes

Poulami Sinhamahapatra, Suprosanna Shit, Anjany Sekuboyina, Malek Husseini, David Schinz, Nicolas Lenhart, Joern Menze, Jan Kirschke, Karsten Roscher, Stephan Guennemann

Vertebral fracture grading classifies the severity of vertebral fractures, which is a challenging task in medical imaging and has recently attracted Deep Learning (DL) models. Only a few works attempted to make such models human-interpretable despite the need for transparency and trustworthiness in critical use cases like DL-assisted medical diagnosis. Moreover, such models either rely on post-hoc methods or additional annotations. In this work, we propose a novel interpretable-by-design method, ProtoVerse, to find relevant sub-parts of vertebral fractures (prototypes) that reliably explain the model's decision in a human-understandable way. Specifically, we introduce a novel diversity-promoting loss to mitigate prototype repetitions in small datasets with intricate semantics. We have experimented with the VerSe'19 dataset and outperformed the existing prototype-based method. Further, our model provides superior interpretability against the post-hoc method. Importantly, expert radiologists validated the visual interpretability of our results, showing clinical applicability.

link: http://arxiv.org/abs/2404.02830v1

Empowering Biomedical Discovery with AI Agents

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, Marinka Zitnik

We envision 'AI scientists' as systems capable of skeptical learning and reasoning that empower biomedical research through collaborative agents that integrate machine learning tools with experimental platforms. Rather than taking humans out of the discovery process, biomedical AI agents combine human creativity and expertise with AI's ability to analyze large datasets, navigate hypothesis spaces, and execute repetitive tasks. AI agents are proficient in a variety of tasks, including self-assessment and planning of discovery workflows. These agents use large language models and generative models to feature structured memory for continual learning and use machine learning tools to incorporate scientific knowledge, biological principles, and theories. AI agents can impact areas ranging from hybrid cell simulation, programmable control of phenotypes, and the design of cellular circuits to the development of new therapies.

link: http://arxiv.org/abs/2404.02831v1

Retrieving Examples from Memory for Retrieval Augmented Neural Machine Translation: A Systematic Comparison

Maxime Bouthors, Josep Crego, Francois Yvon

Retrieval-Augmented Neural Machine Translation (RAMT) architectures retrieve examples from memory to guide the generation process. While most works in this trend explore new ways to exploit the retrieved examples, the upstream retrieval step is mostly unexplored. In this paper, we study the effect of varying retrieval methods for several translation architectures, to better understand the interplay between these two processes. We conduct experiments in two language pairs in a multi-domain setting and consider several downstream architectures based on a standard autoregressive model, an edit-based model, and a large language model with in-context learning. Our experiments show that the choice of the retrieval technique impacts the translation scores, with variance across architectures. We also discuss the effects of increasing the number and diversity of examples, which are mostly positive across the board.

link: http://arxiv.org/abs/2404.02835v1

Cherry on Top: Parameter Heterogeneity and Quantization in Large Language Models

Wanyun Cui, Qianle Wang

This paper reveals the phenomenon of parameter heterogeneity in large language models (LLMs). We find that a small subset of ``cherry'' parameters exhibit a disproportionately large influence on model performance, while the vast majority of parameters have minimal impact. This heterogeneity is found to be prevalent across different model families, scales, and types. Motivated by this observation, we propose CherryQ, a novel quantization method that unifies the optimization of mixed-precision parameters. CherryQ identifies and preserves the critical cherry parameters in high precision while aggressively quantizing the remaining parameters to low precision. Extensive experiments demonstrate the effectiveness of CherryQ. CherryQ outperforms existing quantization approaches in terms of perplexity and downstream task performance. Notably, our 3-bit quantized Vicuna-1.5 exhibits competitive performance compared to their 16-bit counterparts. These findings highlight the potential of CherryQ for enabling efficient deployment of LLMs by taking advantage of parameter heterogeneity.

link: http://arxiv.org/abs/2404.02837v1

I-Design: Personalized LLM Interior Designer

Ata Çelen, Guo Han, Konrad Schindler, Luc Van Gool, Iro Armeni, Anton Obukhov, Xi Wang

Interior design allows us to be who we are and live how we want - each design is as unique as our distinct personality. However, it is not trivial for non-professionals to express and materialize this since it requires aligning functional and visual expectations with the constraints of physical space; this renders interior design a luxury. To make it more accessible, we present I-Design, a personalized interior designer that allows users to generate and visualize their design goals through natural language communication. I-Design starts with a team of large language model agents that engage in dialogues and logical reasoning with one another, transforming textual user input into feasible scene graph designs with relative object relationships. Subsequently, an effective placement algorithm determines optimal locations for each object within the scene. The final design is then constructed in 3D by retrieving and integrating assets from an existing object database. Additionally, we propose a new evaluation protocol that utilizes a vision-language model and complements the design pipeline. Extensive quantitative and qualitative experiments show that I-Design outperforms existing methods in delivering high-quality 3D design solutions and aligning with abstract concepts that match user input, showcasing its advantages across detailed 3D arrangement and conceptual fidelity.

link: http://arxiv.org/abs/2404.02838v1

A Survey on Error-Bounded Lossy Compression for Scientific Datasets

Sheng Di, Jinyang Liu, Kai Zhao, Xin Liang, Robert Underwood, Zhaorui Zhang, Milan Shah, Yafan Huang, Jiajun Huang, Xiaodong Yu, Congrong Ren, Hanqi Guo, Grant Wilkins, Dingwen Tao, Jiannan Tian, Sian Jin, Zizhe Jian, Daoce Wang, MD Hasanur Rahman, Boyuan Zhang, Jon C. Calhoun, Guanpeng Li, Kazutomo Yoshii, Khalid Ayed Alharthi, Franck Cappello

Error-bounded lossy compression has been effective in significantly reducing the data storage/transfer burden while preserving the reconstructed data fidelity very well. Many error-bounded lossy compressors have been developed for a wide range of parallel and distributed use cases for years. These lossy compressors are designed with distinct compression models and design principles, such that each of them features particular pros and cons. In this paper we provide a comprehensive survey of emerging error-bounded lossy compression techniques for different use cases each involving big data to process. The key contribution is fourfold. (1) We summarize an insightful taxonomy of lossy compression into 6 classic compression models. (2) We provide a comprehensive survey of 10+ commonly used compression components/modules used in error-bounded lossy compressors. (3) We provide a comprehensive survey of 10+ state-of-the-art error-bounded lossy compressors as well as how they combine the various compression modules in their designs. (4) We provide a comprehensive survey of the lossy compression for 10+ modern scientific applications and use-cases. We believe this survey is useful to multiple communities including scientific applications, high-performance computing, lossy compression, and big data.

link: http://arxiv.org/abs/2404.02840v1

Scalable quantum detector tomography by high-performance computing

Timon Schapeler, Robert Schade, Michael Lass, Christian Plessl, Tim J. Bartley

At large scales, quantum systems may become advantageous over their classical counterparts at performing certain tasks. Developing tools to analyse these systems at the relevant scales, in a manner consistent with quantum mechanics, is therefore critical to benchmarking performance and characterising their operation. While classical computational approaches cannot perform like-for-like computations of quantum systems beyond a certain scale, classical high-performance computing (HPC) may nevertheless be useful for precisely these characterisation and certification tasks. By developing open-source customised algorithms using high-performance computing, we perform quantum tomography on a megascale quantum photonic detector covering a Hilbert space of 10^6. This requires finding 10^8 elements of the matrix corresponding to the positive operator valued measure (POVM), the quantum description of the detector, and is achieved in minutes of computation time. Moreover, by exploiting the structure of the problem, we achieve highly efficient parallel scaling, paving the way for quantum objects up to a system size of 10^{12} elements to be reconstructed using this method. In general, this shows that a consistent quantum mechanical description of quantum phenomena is applicable at everyday scales. More concretely, this enables the reconstruction of large-scale quantum sources, processes and detectors used in computation and sampling tasks, which may be necessary to prove their nonclassical character or quantum computational advantage.

link: http://arxiv.org/abs/2404.02844v1