Tue 2024.02.20

On Network Design and Planning 2.0 for Optical-computing-enabled Networks

Dao Thanh Hai, Isaac Woungang

In accommodating the continued explosive growth in Internet traffic, optical core networks have been evolving accordingly thanks to numerous technological and architectural innovations. From an architectural perspective, the adoption of optical-bypass networking in the last two decades has resulted in substantial cost savings, owning to the elimination of massive optical-electrical optical interfaces. In optical-bypass framework, the basic functions of optical nodes include adding (dropping) and cross-connecting transitional lightpaths. Moreover, in the process of cross-connecting transiting lightpaths through an intermediate node, these lightpaths must be separated from each other in either time, frequency or spatial domain, to avoid unwanted interference which deems to deteriorate the signal qualities. In light of recently enormous advances in photonic signal processing / computing technologies enabling the precisely controlled interference of optical channels for various computing functions, we propose a new architectural paradigm for future optical networks, namely, optical-computing-enabled networks. Our proposal is defined by the added capability of optical nodes permitting the superposition of transitional lightpaths for computing purposes to achieve greater capacity efficiency. Specifically, we present two illustrative examples highlighting the potential benefits of bringing about in-network optical computing functions which are relied on optical aggregation and optical XOR gate. The new optical computing capabilities armed at optical nodes therefore call for a radical change in formulating networking problems and designing accompanying algorithms, which are collectively referred to as optical network design and planning 2.0 so that the capital and operational efficiency could be fully unlocked.

link: http://arxiv.org/abs/2402.11618v1

Decoding News Narratives: A Critical Analysis of Large Language Models in Framing Bias Detection

Valeria Pastorino, Jasivan A. Sivakumar, Nafise Sadat Moosavi

This work contributes to the expanding research on the applicability of LLMs in social sciences by examining the performance of GPT-3.5 Turbo, GPT-4, and Flan-T5 models in detecting framing bias in news headlines through zero-shot, few-shot, and explainable prompting methods. A key insight from our evaluation is the notable efficacy of explainable prompting in enhancing the reliability of these models, highlighting the importance of explainable settings for social science research on framing bias. GPT-4, in particular, demonstrated enhanced performance in few-shot scenarios when presented with a range of relevant, in-domain examples. FLAN-T5's poor performance indicates that smaller models may require additional task-specific fine-tuning for identifying framing bias detection. Our study also found that models, particularly GPT-4, often misinterpret emotional language as an indicator of framing bias, underscoring the challenge of distinguishing between reporting genuine emotional expression and intentionally use framing bias in news headlines. We further evaluated the models on two subsets of headlines where the presence or absence of framing bias was either clear-cut or more contested, with the results suggesting that these models' can be useful in flagging potential annotation inaccuracies within existing or new datasets. Finally, the study evaluates the models in real-world conditions ("in the wild"), moving beyond the initial dataset focused on U.S. Gun Violence, assessing the models' performance on framed headlines covering a broad range of topics.

link: http://arxiv.org/abs/2402.11621v1

Logical Closed Loop: Uncovering Object Hallucinations in Large Vision-Language Models

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, Tieniu Tan

Object hallucination has been an Achilles' heel which hinders the broader applications of large vision-language models (LVLMs). Object hallucination refers to the phenomenon that the LVLMs claim non-existent objects in the image. To mitigate the object hallucinations, instruction tuning and external model-based detection methods have been proposed, which either require large-scare computational resources or depend on the detection result of external models. However, there remains an under-explored field to utilize the LVLM itself to alleviate object hallucinations. In this work, we adopt the intuition that the LVLM tends to respond logically consistently for existent objects but inconsistently for hallucinated objects. Therefore, we propose a Logical Closed Loop-based framework for Object Hallucination Detection and Mitigation, namely LogicCheckGPT. In specific, we devise logical consistency probing to raise questions with logical correlations, inquiring about attributes from objects and vice versa. Whether their responses can form a logical closed loop serves as an indicator of object hallucination. As a plug-and-play method, it can be seamlessly applied to all existing LVLMs. Comprehensive experiments conducted on three benchmarks across four LVLMs have demonstrated significant improvements brought by our method, indicating its effectiveness and generality.

link: http://arxiv.org/abs/2402.11622v1

SpeCrawler: Generating OpenAPI Specifications from API Documentation Using Large Language Models

Koren Lazar, Matan Vetzler, Guy Uziel, David Boaz, Esther Goldbraich, David Amid, Ateret Anaby-Tavor

In the digital era, the widespread use of APIs is evident. However, scalable utilization of APIs poses a challenge due to structure divergence observed in online API documentation. This underscores the need for automatic tools to facilitate API consumption. A viable approach involves the conversion of documentation into an API Specification format. While previous attempts have been made using rule-based methods, these approaches encountered difficulties in generalizing across diverse documentation. In this paper we introduce SpeCrawler, a comprehensive system that utilizes large language models (LLMs) to generate OpenAPI Specifications from diverse API documentation through a carefully crafted pipeline. By creating a standardized format for numerous APIs, SpeCrawler aids in streamlining integration processes within API orchestrating systems and facilitating the incorporation of tools into LLMs. The paper explores SpeCrawler's methodology, supported by empirical evidence and case studies, demonstrating its efficacy through LLM capabilities.

link: http://arxiv.org/abs/2402.11625v1

Metacognitive Retrieval-Augmented Large Language Models

Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, Zhicheng Dou

Retrieval-augmented generation have become central in natural language processing due to their efficacy in generating factual content. While traditional methods employ single-time retrieval, more recent approaches have shifted towards multi-time retrieval for multi-hop reasoning tasks. However, these strategies are bound by predefined reasoning steps, potentially leading to inaccuracies in response generation. This paper introduces MetaRAG, an approach that combines the retrieval-augmented generation process with metacognition. Drawing from cognitive psychology, metacognition allows an entity to self-reflect and critically evaluate its cognitive processes. By integrating this, MetaRAG enables the model to monitor, evaluate, and plan its response strategies, enhancing its introspective reasoning abilities. Through a three-step metacognitive regulation pipeline, the model can identify inadequacies in initial cognitive responses and fixes them. Empirical evaluations show that MetaRAG significantly outperforms existing methods.

link: http://arxiv.org/abs/2402.11626v1

Interactive Garment Recommendation with User in the Loop

Federico Becattini, Xiaolin Chen, Andrea Puccia, Haokun Wen, Xuemeng Song, Liqiang Nie, Alberto Del Bimbo

Recommending fashion items often leverages rich user profiles and makes targeted suggestions based on past history and previous purchases. In this paper, we work under the assumption that no prior knowledge is given about a user. We propose to build a user profile on the fly by integrating user reactions as we recommend complementary items to compose an outfit. We present a reinforcement learning agent capable of suggesting appropriate garments and ingesting user feedback so to improve its recommendations and maximize user satisfaction. To train such a model, we resort to a proxy model to be able to simulate having user feedback in the training loop. We experiment on the IQON3000 fashion dataset and we find that a reinforcement learning-based agent becomes capable of improving its recommendations by taking into account personal preferences. Furthermore, such task demonstrated to be hard for non-reinforcement models, that cannot exploit exploration during training.

link: http://arxiv.org/abs/2402.11627v1

Discrete Neural Algorithmic Reasoning

Gleb Rodionov, Liudmila Prokhorenkova

Neural algorithmic reasoning aims to capture computations with neural networks via learning the models to imitate the execution of classical algorithms. While common architectures are expressive enough to contain the correct model in the weights space, current neural reasoners are struggling to generalize well on out-of-distribution data. On the other hand, classical computations are not affected by distribution shifts as they can be described as transitions between discrete computational states. In this work, we propose to force neural reasoners to maintain the execution trajectory as a combination of finite predefined states. Trained with supervision on the algorithm's state transitions, such models are able to perfectly align with the original algorithm. To show this, we evaluate our approach on the SALSA-CLRS benchmark, where we get perfect test scores for all tasks. Moreover, the proposed architectural choice allows us to prove the correctness of the learned algorithms for any test data.

link: http://arxiv.org/abs/2402.11628v1

Neuromorphic Face Analysis: a Survey

Federico Becattini, Lorenzo Berlincioni, Luca Cultrera, Alberto Del Bimbo

Neuromorphic sensors, also known as event cameras, are a class of imaging devices mimicking the function of biological visual systems. Unlike traditional frame-based cameras, which capture fixed images at discrete intervals, neuromorphic sensors continuously generate events that represent changes in light intensity or motion in the visual field with high temporal resolution and low latency. These properties have proven to be interesting in modeling human faces, both from an effectiveness and a privacy-preserving point of view. Neuromorphic face analysis however is still a raw and unstructured field of research, with several attempts at addressing different tasks with no clear standard or benchmark. This survey paper presents a comprehensive overview of capabilities, challenges and emerging applications in the domain of neuromorphic face analysis, to outline promising directions and open issues. After discussing the fundamental working principles of neuromorphic vision and presenting an in-depth overview of the related research, we explore the current state of available data, standard data representations, emerging challenges, and limitations that require further investigation. This paper aims to highlight the recent process in this evolving field to provide to both experienced and newly come researchers an all-encompassing analysis of the state of the art along with its problems and shortcomings.

link: http://arxiv.org/abs/2402.11631v1

Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking dialogs

Arian Askari, Roxana Petcu, Chuan Meng, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas, Suzan Verberne

Identifying user intents in information-seeking dialogs is crucial for a system to meet user's information needs. Intent prediction (IP) is challenging and demands sufficient dialogs with

human-labeled intents for training. However, manually annotating intents is resource-intensive. While large language models (LLMs) have been shown to be effective in generating synthetic data, there is no study on using LLMs to generate intent-aware information-seeking dialogs. In this paper, we focus on leveraging LLMs for zero-shot generation of large-scale, open-domain, and intent-aware information-seeking dialogs. We propose SOLID, which has novel self-seeding and multi-intent self-instructing schemes. The former improves the generation quality by using the LLM's own knowledge scope to initiate dialog generation; the latter prompts the LLM to generate utterances sequentially, and mitigates the need for manual prompt design by asking the LLM to autonomously adapt its prompt instruction when generating complex multi-intent utterances. Furthermore, we propose SOLID-RL, which is further trained to generate a dialog in one step on the data generated by SOLID. We propose a length-based quality estimation mechanism to assign varying weights to SOLID-generated dialogs based on their quality during the training process of SOLID-RL. We use SOLID and SOLID-RL to generate more than 300k intent-aware dialogs, surpassing the size of existing datasets. Experiments show that IP methods trained on dialogs generated by SOLID and SOLID-RL achieve better IP quality than ones trained on human-generated dialogs.

link: http://arxiv.org/abs/2402.11633v1

Poisoning Federated Recommender Systems with Fake Users

Ming Yin, Yichang Xu, Minghong Fang, Neil Zhenqiang Gong

Federated recommendation is a prominent use case within federated learning, yet it remains susceptible to various attacks, from user to server-side vulnerabilities. Poisoning attacks are particularly notable among user-side attacks, as participants upload malicious model updates to deceive the global model, often intending to promote or demote specific targeted items. This study investigates strategies for executing promotion attacks in federated recommender systems. Current poisoning attacks on federated recommender systems often rely on additional information, such as the local training data of genuine users or item popularity. However, such information is challenging for the potential attacker to obtain. Thus, there is a need to develop an attack that requires no extra information apart from item embeddings obtained from the server. In this paper, we introduce a novel fake user based poisoning attack named PoisonFRS to promote the attacker-chosen targeted item in federated recommender systems without requiring knowledge about user-item rating data, user attributes, or the aggregation rule used by the server. Extensive experiments on multiple real-world datasets demonstrate that PoisonFRS can effectively promote the attacker-chosen targeted item to a large portion of genuine users and outperform current benchmarks that rely on additional information about the system. We further observe that the model updates from both genuine and fake users are indistinguishable within the latent space.

link: http://arxiv.org/abs/2402.11637v1

Stumbling Blocks: Stress Testing the Robustness of Machine-Generated Text Detectors Under Attacks

Yichen Wang, Shangbin Feng, Abe Bohan Hou, Xiao Pu, Chao Shen, Xiaoming Liu, Yulia Tsvetkov, Tianxing He

The widespread use of large language models (LLMs) is increasing the demand for methods that detect machine-generated text to prevent misuse. The goal of our study is to stress test the detectors' robustness to malicious attacks under realistic scenarios. We comprehensively study the robustness of popular machine-generated text detectors under attacks from diverse categories: editing, paraphrasing, prompting, and co-generating. Our attacks assume limited access to the generator LLMs, and we compare the performance of detectors on different attacks under different budget levels. Our experiments reveal that almost none of the existing detectors remain robust under all the attacks, and all detectors exhibit different loopholes. Averaging all detectors, the performance drops by 35% across all attacks. Further, we investigate the reasons behind these defects and propose initial out-of-the-box patches to improve robustness.

link: http://arxiv.org/abs/2402.11638v1

In-Context Learning with Transformers: Softmax Attention Adapts to Function Lipschitzness

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, Sanjay Shakkottai

A striking property of transformers is their ability to perform in-context learning (ICL), a machine learning framework in which the learner is presented with a novel context during inference implicitly through some data, and tasked with making a prediction in that context. As such that learner must adapt to the context without additional training. We explore the role of softmax attention in an ICL setting where each context encodes a regression task. We show that an attention unit learns a window that it uses to implement a nearest-neighbors predictor adapted to the landscape of the pretraining tasks. Specifically, we show that this window widens with decreasing Lipschitzness and increasing label noise in the pretraining tasks. We also show that on low-rank, linear problems, the attention unit learns to project onto the appropriate subspace before inference. Further, we show that this adaptivity relies crucially on the softmax activation and thus cannot be replicated by the linear activation often studied in prior theoretical analyses.

link: http://arxiv.org/abs/2402.11639v1

Towards Versatile Graph Learning Approach: from the Perspective of Large Language Models

Lanning Wei, Jun Gao, Huan Zhao

Graph-structured data are the commonly used and have wide application scenarios in the real world. For these diverse applications, the vast variety of learning tasks, graph domains, and complex graph learning procedures present challenges for human experts when designing versatile graph learning approaches. Facing these challenges, large language models (LLMs) offer a potential solution due to the extensive knowledge and the human-like intelligence. This paper proposes a novel conceptual prototype for designing versatile graph learning methods with LLMs, with a particular focus on the ``where'' and ``how'' perspectives. From the ``where'' perspective, we summarize four key graph learning procedures, including task definition, graph data feature engineering, model selection and optimization, deployment and serving. We then explore the application scenarios of LLMs in these procedures across a wider spectrum. In the ``how'' perspective, we align the abilities of LLMs with the requirements of each procedure. Finally, we point out the promising directions that could better leverage the strength of LLMs towards versatile graph learning methods.

link: http://arxiv.org/abs/2402.11641v1

Quantum Image Denoising with Machine Learning: A Novel Approach to Improve Quantum Image Processing Quality and Reliability

Yew Kee Wonga, Yifan Zhou, Yan Shing Liang

Quantum Image Processing (QIP) is a field that aims to utilize the benefits of quantum computing for manipulating and analyzing images. However, QIP faces two challenges: the limitation of qubits and the presence of noise in a quantum machine. In this research we propose a novel approach to address the issue of noise in QIP. By training and employing a machine learning model that identifies and corrects the noise in quantum processed images, we can compensate for the noisiness caused by the machine and retrieve a processing result similar to that performed by a classical computer with higher efficiency. The model is trained by learning a dataset consisting of both existing processed images and quantum processed images from open access datasets. This model will be capable of providing us with the confidence level for each pixel and its potential original value. To assess the model's accuracy in compensating for loss and decoherence in QIP, we evaluate it using three metrics: Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Opinion Score (MOS). Additionally, we discuss the applicability of our model across domains well as its cost effectiveness compared to alternative methods.

link: http://arxiv.org/abs/2402.11645v1

Theoretical foundations for programmatic reinforcement learning

Guruprerana Shabadi, Nathanaël Fijalkow, Théo Matricon

The field of Reinforcement Learning (RL) is concerned with algorithms for learning optimal policies in unknown stochastic environments. Programmatic RL studies representations of policies as programs, meaning involving higher order constructs such as control loops. Despite attracting a lot of attention at the intersection of the machine learning and formal methods communities, very little is known on the theoretical front about programmatic RL: what are good classes of programmatic policies? How large are optimal programmatic policies? How can we learn them? The goal of this paper is to give first answers to these questions, initiating a theoretical study of programmatic RL.

link: http://arxiv.org/abs/2402.11650v1