Thu 2024.04.11

Enhancing Software Related Information Extraction with Generative Language Models through Single-Choice Question Answering

Wolfgang Otto, Sharmila Upadhyaya, Stefan Dietze

This paper describes our participation in the Shared Task on Software Mentions Disambiguation (SOMD), with a focus on improving relation extraction in scholarly texts through Generative Language Models (GLMs) using single-choice question-answering. The methodology prioritises the use of in-context learning capabilities of GLMs to extract software-related entities and their descriptive attributes, such as distributive information. Our approach uses Retrieval-Augmented Generation (RAG) techniques and GLMs for Named Entity Recognition (NER) and Attributive NER to identify relationships between extracted software entities, providing a structured solution for analysing software citations in academic literature. The paper provides a detailed description of our approach, demonstrating how using GLMs in a single-choice QA paradigm can greatly enhance IE methodologies. Our participation in the SOMD shared task highlights the importance of precise software citation practices and showcases our system's ability to overcome the challenges of disambiguating and extracting relationships between software mentions. This sets the groundwork for future research and development in this field.

link: http://arxiv.org/abs/2404.05587v1

MedExpQA: Multilingual Benchmarking of Large Language Models for Medical Question Answering

Iñigo Alonso, Maite Oronoz, Rodrigo Agerri

Large Language Models (LLMs) have the potential of facilitating the development of Artificial Intelligence technology to assist medical experts for interactive decision support, which has been demonstrated by their competitive performances in Medical QA. However, while impressive, the required quality bar for medical applications remains far from being achieved. Currently, LLMs remain challenged by outdated knowledge and by their tendency to generate hallucinated content. Furthermore, most benchmarks to assess medical knowledge lack reference gold explanations which means that it is not possible to evaluate the reasoning of LLMs predictions. Finally, the situation is particularly grim if we consider benchmarking LLMs for languages other than English which remains, as far as we know, a totally neglected topic. In order to address these shortcomings, in this paper we present MedExpQA, the first multilingual benchmark based on medical exams to evaluate LLMs in Medical Question Answering. To the best of our knowledge, MedExpQA includes for the first time reference gold explanations written by medical doctors which can be leveraged to establish various gold-based upper-bounds for comparison with LLMs performance. Comprehensive multilingual experimentation using both the gold reference explanations and Retrieval Augmented Generation (RAG) approaches show that performance of LLMs still has large room for improvement, especially for languages other than English. Furthermore, and despite using state-of-the-art RAG methods, our results also demonstrate the difficulty of obtaining and integrating readily available medical knowledge that may positively impact results on downstream evaluations for Medical Question Answering. So far the benchmark is available in four languages, but we hope that this work may encourage further development to other languages.

link: http://arxiv.org/abs/2404.05590v1

UniFL: Improve Stable Diffusion via Unified Feedback Learning

Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li, Xuefeng Xiao, Weilin Huang, Min Zheng, Lean Fu, Guanbin Li

Diffusion models have revolutionized the field of image generation, leading to the proliferation of high-quality models and diverse downstream applications. However, despite these significant advancements, the current competitive solutions still suffer from several limitations, including inferior visual quality, a lack of aesthetic appeal, and inefficient inference, without a comprehensive solution in sight. To address these challenges, we present UniFL, a unified framework that leverages feedback learning to enhance diffusion models comprehensively. UniFL stands out as a universal, effective, and generalizable solution applicable to various diffusion models, such as SD1.5 and SDXL. Notably, UniFL incorporates three key components: perceptual feedback learning, which enhances visual quality; decoupled feedback learning, which improves aesthetic appeal; and adversarial feedback learning, which optimizes inference speed. In-depth experiments and extensive user studies validate the superior performance of our proposed method in enhancing both the quality of generated models and their acceleration. For instance, UniFL surpasses ImageReward by 17% user preference in terms of generation quality and outperforms LCM and SDXL Turbo by 57% and 20% in 4-step inference. Moreover, we have verified the efficacy of our approach in downstream tasks, including Lora, ControlNet, and AnimateDiff.

link: http://arxiv.org/abs/2404.05595v1

Hook-in Privacy Techniques for gRPC-based Microservice Communication

Louis Loechel, Siar-Remzi Akbayin, Elias Grünewald, Jannis Kiesel, Inga Strelnikova, Thomas Janke, Frank Pallas

gRPC is at the heart of modern distributed system architectures. Based on HTTP/2 and Protocol Buffers, it provides highly performant, standardized, and polyglot communication across loosely coupled microservices and is increasingly preferred over REST- or GraphQL-based service APIs in practice. Despite its widespread adoption, gRPC lacks any advanced privacy techniques beyond transport encryption and basic token-based authentication. Such advanced techniques are, however, increasingly important for fulfilling regulatory requirements. For instance, anonymizing or otherwise minimizing (personal) data before responding to requests, or pre-processing data based on the purpose of the access may be crucial in certain usecases. In this paper, we therefore propose a novel approach for integrating such advanced privacy techniques into the gRPC framework in a practically viable way. Specifically, we present a general approach along with a working prototype that implements privacy techniques, such as data minimization and purpose limitation, in a configurable, extensible, and gRPC-native way utilizing a gRPC interceptor. We also showcase how to integrate this contribution into a realistic example of a food delivery use case. Alongside these implementations, a preliminary performance evaluation shows practical applicability with reasonable overheads. Altogether, we present a viable solution for integrating advanced privacy techniques into real-world gRPC-based microservice architectures, thereby facilitating regulatory compliance ``by design''.

link: http://arxiv.org/abs/2404.05598v1

The Argument for Meta-Modeling-Based Approaches to Hardware Generation Languages

Johannes Schreiner, Daniel Gerl, Robert Kunzelmann, Paritosh Kumar Sinha, Wolfgang Ecker

The rapid evolution of Integrated Circuit (IC) development necessitates innovative methodologies such as code generation to manage complexity and increase productivity. Using the right methodology for generator development to maximize the capability and, most notably, the feasibility of generators is a crucial part of this work. Meta-Modeling-based approaches drawing on the principles of Model Driven Architecture (MDA) are a promising methodology for generator development. The goal of this paper is to show why such an MDA-based approach can provide extremely powerful generators with minimal implementation effort and to demonstrate that this approach is a superior alternative to the most advanced hardware generation languages such as SpinalHDL and Chisel. For this purpose, this paper provides an in-depth comparison of the Meta-Modeling approach against these hardware generation languages, highlighting the unique advantages of a Meta-Modeling-based approach and summarizes the benefits.

link: http://arxiv.org/abs/2404.05599v1

SpeechAlign: Aligning Speech Generation to Human Preferences

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, Xipeng Qiu

Speech language models have significantly advanced in generating realistic speech, with neural codec language models standing out. However, the integration of human feedback to align speech outputs to human preferences is often neglected. This paper addresses this gap by first analyzing the distribution gap in codec language models, highlighting how it leads to discrepancies between the training and inference phases, which negatively affects performance. Then we explore leveraging learning from human feedback to bridge the distribution gap. We introduce SpeechAlign, an iterative self-improvement strategy that aligns speech language models to human preferences. SpeechAlign involves constructing a preference codec dataset contrasting golden codec tokens against synthetic tokens, followed by preference optimization to improve the codec language model. This cycle of improvement is carried out iteratively to steadily convert weak models to strong ones. Through both subjective and objective evaluations, we show that SpeechAlign can bridge the distribution gap and facilitating continuous self-improvement of the speech language model. Moreover, SpeechAlign exhibits robust generalization capabilities and works for smaller models. Code and models will be available at https://github.com/0nutation/SpeechGPT.

link: http://arxiv.org/abs/2404.05600v1

AI-Enabled System for Efficient and Effective Cyber Incident Detection and Response in Cloud Environments

Mohammed Ashfaaq M. Farzaan, Mohamed Chahine Ghanem, Ayman El-Hajjar, Deepthi N. Ratnayake

The escalating sophistication and volume of cyber threats in cloud environments necessitate a paradigm shift in strategies. Recognising the need for an automated and precise response to cyber threats, this research explores the application of AI and ML and proposes an AI-powered cyber incident response system for cloud environments. This system, encompassing Network Traffic Classification, Web Intrusion Detection, and post-incident Malware Analysis (built as a Flask application), achieves seamless integration across platforms like Google Cloud and Microsoft Azure. The findings from this research highlight the effectiveness of the Random Forest model, achieving an accuracy of 90% for the Network Traffic Classifier and 96% for the Malware Analysis Dual Model application. Our research highlights the strengths of AI-powered cyber security. The Random Forest model excels at classifying cyber threats, offering an efficient and robust solution. Deep learning models significantly improve accuracy, and their resource demands can be managed using cloud-based TPUs and GPUs. Cloud environments themselves provide a perfect platform for hosting these AI/ML systems, while container technology ensures both efficiency and scalability. These findings demonstrate the contribution of the AI-led system in guaranteeing a robust and scalable cyber incident response solution in the cloud.

link: http://arxiv.org/abs/2404.05602v2

Self-Explainable Affordance Learning with Embodied Caption

Zhipeng Zhang, Zhimin Wei, Guolei Sun, Peng Wang, Luc Van Gool

In the field of visual affordance learning, previous methods mainly used abundant images or videos that delineate human behavior patterns to identify action possibility regions for object manipulation, with a variety of applications in robotic tasks. However, they encounter a main challenge of action ambiguity, illustrated by the vagueness like whether to beat or carry a drum, and the complexities involved in processing intricate scenes. Moreover, it is important for human intervention to rectify robot errors in time. To address these issues, we introduce Self-Explainable Affordance learning (SEA) with embodied caption. This innovation enables robots to articulate their intentions and bridge the gap between explainable vision-language caption and visual affordance learning. Due to a lack of appropriate dataset, we unveil a pioneering dataset and metrics tailored for this task, which integrates images, heatmaps, and embodied captions. Furthermore, we propose a novel model to effectively combine affordance grounding with self-explanation in a simple but efficient manner. Extensive quantitative and qualitative experiments demonstrate our method's effectiveness.

link: http://arxiv.org/abs/2404.05603v1

Technical Report: The Graph Spectral Token -- Enhancing Graph Transformers with Spectral Information

Zihan Pengmei, Zimu Li

Graph Transformers have emerged as a powerful alternative to Message-Passing Graph Neural Networks (MP-GNNs) to address limitations such as over-squashing of information exchange. However, incorporating graph inductive bias into transformer architectures remains a significant challenge. In this report, we propose the Graph Spectral Token, a novel approach to directly encode graph spectral information, which captures the global structure of the graph, into the transformer architecture. By parameterizing the auxiliary [CLS] token and leaving other tokens representing graph nodes, our method seamlessly integrates spectral information into the learning process. We benchmark the effectiveness of our approach by enhancing two existing graph transformers, GraphTrans and SubFormer. The improved GraphTrans, dubbed GraphTrans-Spec, achieves over 10% improvements on large graph benchmark datasets while maintaining efficiency comparable to MP-GNNs. SubFormer-Spec demonstrates strong performance across various datasets.

link: http://arxiv.org/abs/2404.05604v1

Graph Neural Networks Automated Design and Deployment on Device-Edge Co-Inference Systems

Ao Zhou, Jianlei Yang, Tong Qiao, Yingjie Qi, Zhi Yang, Weisheng Zhao, Chunming Hu

The key to device-edge co-inference paradigm is to partition models into computation-friendly and computation-intensive parts across the device and the edge, respectively. However, for Graph Neural Networks (GNNs), we find that simply partitioning without altering their structures can hardly achieve the full potential of the co-inference paradigm due to various computational-communication overheads of GNN operations over heterogeneous devices. We present GCoDE, the first automatic framework for GNN that innovatively Co-designs the architecture search and the mapping of each operation on Device-Edge hierarchies. GCoDE abstracts the device communication process into an explicit operation and fuses the search of architecture and the operations mapping in a unified space for joint-optimization. Also, the performance-awareness approach, utilized in the constraint-based search process of GCoDE, enables effective evaluation of architecture efficiency in diverse heterogeneous systems. We implement the co-inference engine and runtime dispatcher in GCoDE to enhance the deployment efficiency. Experimental results show that GCoDE can achieve up to $44.9\times$ speedup and 98.2% energy reduction compared to existing approaches across various applications and system configurations.

link: http://arxiv.org/abs/2404.05605v1

Learning Topology Uniformed Face Mesh by Volume Rendering for Multi-view Reconstruction

Yating Wang, Ran Yi, Ke Fan, Jinkun Hao, Jiangbo Lu, Lizhuang Ma

Face meshes in consistent topology serve as the foundation for many face-related applications, such as 3DMM constrained face reconstruction and expression retargeting. Traditional methods commonly acquire topology uniformed face meshes by two separate steps: multi-view stereo (MVS) to reconstruct shapes followed by non-rigid registration to align topology, but struggles with handling noise and non-lambertian surfaces. Recently neural volume rendering techniques have been rapidly evolved and shown great advantages in 3D reconstruction or novel view synthesis. Our goal is to leverage the superiority of neural volume rendering into multi-view reconstruction of face mesh with consistent topology. We propose a mesh volume rendering method that enables directly optimizing mesh geometry while preserving topology, and learning implicit features to model complex facial appearance from multi-view images. The key innovation lies in spreading sparse mesh features into the surrounding space to simulate radiance field required for volume rendering, which facilitates backpropagation of gradients from images to mesh geometry and implicit appearance features. Our proposed feature spreading module exhibits deformation invariance, enabling photorealistic rendering seamlessly after mesh editing. We conduct experiments on multi-view face image dataset to evaluate the reconstruction and implement an application for photorealistic rendering of animated face mesh.

link: http://arxiv.org/abs/2404.05606v1

A Training-Free Plug-and-Play Watermark Framework for Stable Diffusion

Guokai Zhang, Lanjun Wang, Yuting Su, An-An Liu

Nowadays, the family of Stable Diffusion (SD) models has gained prominence for its high quality outputs and scalability. This has also raised security concerns on social media, as malicious users can create and disseminate harmful content. Existing approaches involve training components or entire SDs to embed a watermark in generated images for traceability and responsibility attribution. However, in the era of AI-generated content (AIGC), the rapid iteration of SDs renders retraining with watermark models costly. To address this, we propose a training-free plug-and-play watermark framework for SDs. Without modifying any components of SDs, we embed diverse watermarks in the latent space, adapting to the denoising process. Our experimental findings reveal that our method effectively harmonizes image quality and watermark invisibility. Furthermore, it performs robustly under various attacks. We also have validated that our method is generalized to multiple versions of SDs, even without retraining the watermark model.

link: http://arxiv.org/abs/2404.05607v1

KaMPIng: Flexible and (Near) Zero-overhead C++ Bindings for MPI

Demian Hespe, Lukas Hübner, Florian Kurpicz, Peter Sanders, Matthias Schimek, Daniel Seemaier, Christoph Stelz, Tim Niklas Uhl

The Message-Passing Interface (MPI) and C++ form the backbone of high-performance computing, but MPI only provides C and Fortran bindings. While this offers great language interoperability, high-level programming languages like C++ make software development quicker and less error-prone. We propose novel C++ language bindings that cover all abstraction levels from low-level MPI calls to convenient STL-style bindings, where most parameters are inferred from a small subset of parameters, by bringing named parameters to C++. This enables rapid prototyping and fine-tuning runtime behavior and memory management. A flexible type system and additional safeness guarantees help to prevent programming errors. By exploiting C++'s template-metaprogramming capabilities, this has (near) zero-overhead, as only required code paths are generated at compile time. We demonstrate that our library is a strong foundation for a future distributed standard library using multiple application benchmarks, ranging from text-book sorting algorithms to phylogenetic interference.

link: http://arxiv.org/abs/2404.05610v1

Deep Representation Learning for Multi-functional Degradation Modeling of Community-dwelling Aging Population

Suiyao Chen, Xinyi Liu, Yulei Li, Jing Wu, Handong Yao

As the aging population grows, particularly for the baby boomer generation, the United States is witnessing a significant increase in the elderly population experiencing multifunctional disabilities. These disabilities, stemming from a variety of chronic diseases, injuries, and impairments, present a complex challenge due to their multidimensional nature, encompassing both physical and cognitive aspects. Traditional methods often use univariate regression-based methods to model and predict single degradation conditions and assume population homogeneity, which is inadequate to address the complexity and diversity of aging-related degradation. This study introduces a novel framework for multi-functional degradation modeling that captures the multidimensional (e.g., physical and cognitive) and heterogeneous nature of elderly disabilities. Utilizing deep learning, our approach predicts health degradation scores and uncovers latent heterogeneity from elderly health histories, offering both efficient estimation and explainable insights into the diverse effects and causes of aging-related degradation. A real-case study demonstrates the effectiveness and marks a pivotal contribution to accurately modeling the intricate dynamics of elderly degradation, and addresses the healthcare challenges in the aging population.

link: http://arxiv.org/abs/2404.05613v1

MULTIFLOW: Shifting Towards Task-Agnostic Vision-Language Pruning

Matteo Farina, Massimiliano Mancini, Elia Cunegatti, Gaowen Liu, Giovanni Iacca, Elisa Ricci

While excellent in transfer learning, Vision-Language models (VLMs) come with high computational costs due to their large number of parameters. To address this issue, removing parameters via model pruning is a viable solution. However, existing techniques for VLMs are task-specific, and thus require pruning the network from scratch for each new task of interest. In this work, we explore a new direction: Task-Agnostic Vision-Language Pruning (TA-VLP). Given a pretrained VLM, the goal is to find a unique pruned counterpart transferable to multiple unknown downstream tasks. In this challenging setting, the transferable representations already encoded in the pretrained model are a key aspect to preserve. Thus, we propose Multimodal Flow Pruning (MULTIFLOW), a first, gradient-free, pruning framework for TA-VLP where: (i) the importance of a parameter is expressed in terms of its magnitude and its information flow, by incorporating the saliency of the neurons it connects; and (ii) pruning is driven by the emergent (multimodal) distribution of the VLM parameters after pretraining. We benchmark eight state-of-the-art pruning algorithms in the context of TA-VLP, experimenting with two VLMs, three vision-language tasks, and three pruning ratios. Our experimental results show that MULTIFLOW outperforms recent sophisticated, combinatorial competitors in the vast majority of the cases, paving the way towards addressing TA-VLP. The code is publicly available at https://github.com/FarinaMatteo/multiflow.

link: http://arxiv.org/abs/2404.05621v1