Tue 2024.02.27

Interpreting Predictive Probabilities: Model Confidence or Human Label Variation?

Joris Baan, Raquel Fernández, Barbara Plank, Wilker Aziz

With the rise of increasingly powerful and user-facing NLP systems, there is growing interest in assessing whether they have a good representation of uncertainty by evaluating the quality of their predictive distribution over outcomes. We identify two main perspectives that drive starkly different evaluation protocols. The first treats predictive probability as an indication of model confidence; the second as an indication of human label variation. We discuss their merits and limitations, and take the position that both are crucial for trustworthy and fair NLP systems, but that exploiting a single predictive distribution is limiting. We recommend tools and highlight exciting directions towards models with disentangled representations of uncertainty about predictions and uncertainty about human labels.

link: http://arxiv.org/abs/2402.16102v1

Informed Meta-Learning

Katarzyna Kobalczyk, Mihaela van der Schaar

In noisy and low-data regimes prevalent in real-world applications, an outstanding challenge of machine learning lies in effectively incorporating inductive biases that promote data efficiency and robustness. Meta-learning and informed ML stand out as two approaches for incorporating prior knowledge into the ML pipeline. While the former relies on a purely data-driven source of priors, the latter is guided by a formal representation of expert knowledge. This paper introduces a novel hybrid paradigm, informed meta-learning, seeking complementarity in cross-task knowledge sharing of humans and machines. We establish the foundational components of informed meta-learning and present a concrete instantiation of this framework--the Informed Neural Process. Through a series of illustrative and larger-scale experiments, we demonstrate the potential benefits of informed meta-learning in improving data efficiency and robustness to observational noise, task distribution shifts, and heterogeneity.

link: http://arxiv.org/abs/2402.16105v1

FuseChat: Knowledge Fusion of Chat Models

Fanqi Wan, Ziyi Yang, Longguang Zhong, Xiaojun Quan, Xinting Huang, Wei Bi

While training large language models (LLMs) from scratch can indeed lead to models with distinct capabilities and strengths, this approach incurs substantial costs and may lead to potential redundancy in competencies. An alternative strategy is to combine existing LLMs into a more robust LLM, thereby diminishing the necessity for expensive pre-training. However, due to the diverse architectures of LLMs, direct parameter blending proves to be unfeasible. Recently, \textsc{FuseLLM} introduced the concept of knowledge fusion to transfer the collective knowledge of multiple structurally varied LLMs into a target LLM through lightweight continual training. In this report, we extend the scalability and flexibility of the \textsc{FuseLLM} framework to realize the fusion of chat LLMs, resulting in \textsc{FuseChat}. \textsc{FuseChat} comprises two main stages. Firstly, we undertake knowledge fusion for structurally and scale-varied source LLMs to derive multiple target LLMs of identical structure and size via lightweight fine-tuning. Then, these target LLMs are merged within the parameter space, wherein we propose a novel method for determining the merging weights based on the variation ratio of parameter matrices before and after fine-tuning. We validate our approach using three prominent chat LLMs with diverse architectures and scales, namely \texttt{NH2-Mixtral-8x7B}, \texttt{NH2-Solar-10.7B}, and \texttt{OpenChat-3.5-7B}. Experimental results spanning various chat domains demonstrate the superiority of \texttt{\textsc{FuseChat}-7B} across a broad spectrum of chat LLMs at 7B and 34B scales, even surpassing \texttt{GPT-3.5 (March)} and approaching \texttt{Mixtral-8x7B-Instruct}. Our code, model weights, and data are openly accessible at \url{https://github.com/fanqiwan/FuseLLM}.

link: http://arxiv.org/abs/2402.16107v1

RoboCodeX: Multimodal Code Generation for Robotic Behavior Synthesis

Yao Mu, Junting Chen, Qinglong Zhang, Shoufa Chen, Qiaojun Yu, Chongjian Ge, Runjian Chen, Zhixuan Liang, Mengkang Hu, Chaofan Tao, Peize Sun, Haibao Yu, Chao Yang, Wenqi Shao, Wenhai Wang, Jifeng Dai, Yu Qiao, Mingyu Ding, Ping Luo

Robotic behavior synthesis, the problem of understanding multimodal inputs and generating precise physical control for robots, is an important part of Embodied AI. Despite successes in applying multimodal large language models for high-level understanding, it remains challenging to translate these conceptual understandings into detailed robotic actions while achieving generalization across various scenarios. In this paper, we propose a tree-structured multimodal code generation framework for generalized robotic behavior synthesis, termed RoboCodeX. RoboCodeX decomposes high-level human instructions into multiple object-centric manipulation units consisting of physical preferences such as affordance and safety constraints, and applies code generation to introduce generalization ability across various robotics platforms. To further enhance the capability to map conceptual and perceptual understanding into control commands, a specialized multimodal reasoning dataset is collected for pre-training and an iterative self-updating methodology is introduced for supervised fine-tuning. Extensive experiments demonstrate that RoboCodeX achieves state-of-the-art performance in both simulators and real robots on four different kinds of manipulation tasks and one navigation task.

link: http://arxiv.org/abs/2402.16117v1

DeepForge: Leveraging AI for Microstructural Control in Metal Forming via Model Predictive Control

Jan Petrik, Markus Bambach

This study presents a novel method for microstructure control in closed die hot forging that combines Model Predictive Control (MPC) with a developed machine learning model called DeepForge. DeepForge uses an architecture that combines 1D convolutional neural networks and gated recurrent units. It uses surface temperature measurements of a workpiece as input to predict microstructure changes during forging. The paper also details DeepForge's architecture and the finite element simulation model used to generate the data set, using a three-stroke forging process. The results demonstrate DeepForge's ability to predict microstructure with a mean absolute error of 0.4$\pm$0.3%. In addition, the study explores the use of MPC to adjust inter-stroke wait times, effectively counteracting temperature disturbances to achieve a target grain size of less than 35 microns within a specific 2D region of the workpiece. These results are then verified experimentally, demonstrating a significant step towards improved control and quality in forging processes where temperature can be used as an additional degree of freedom in the process.

link: http://arxiv.org/abs/2402.16119v1

Towards Accurate Post-training Quantization for Reparameterized Models

Luoming Zhang, Yefei He, Wen Fei, Zhenyu Lou, Weijia Wu, YangWei Ying, Hong Zhou

Model reparameterization is a widely accepted technique for improving inference speed without compromising performance. However, current Post-training Quantization (PTQ) methods often lead to significant accuracy degradation when applied to reparameterized models. This is primarily caused by channel-specific and sample-specific outliers, which appear only at specific samples and channels and impact on the selection of quantization parameters. To address this issue, we propose RepAPQ, a novel framework that preserves the accuracy of quantized reparameterization models. Different from previous frameworks using Mean Squared Error (MSE) as a measurement, we utilize Mean Absolute Error (MAE) to mitigate the influence of outliers on quantization parameters. Our framework comprises two main components: Quantization Protecting Reparameterization and Across-block Calibration. For effective calibration, Quantization Protecting Reparameterization combines multiple branches into a single convolution with an affine layer. During training, the affine layer accelerates convergence and amplifies the output of the convolution to better accommodate samples with outliers. Additionally, Across-block Calibration leverages the measurement of stage output as supervision to address the gradient problem introduced by MAE and enhance the interlayer correlation with quantization parameters. Comprehensive experiments demonstrate the effectiveness of RepAPQ across various models and tasks. Our framework outperforms previous methods by approximately 1\% for 8-bit PTQ and 2\% for 6-bit PTQ, showcasing its superior performance. The code is available at \url{https://github.com/ilur98/DLMC-QUANT}.

link: http://arxiv.org/abs/2402.16121v1

InstructEdit: Instruction-based Knowledge Editing for Large Language Models

Bozhong Tian, Siyuan Cheng, Xiaozhuan Liang, Ningyu Zhang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, Huajun Chen

Knowledge editing for large language models can offer an efficient solution to alter a model's behavior without negatively impacting the overall performance. However, the current approach encounters issues with limited generalizability across tasks, necessitating one distinct editor for each task, which significantly hinders the broader applications. To address this, we take the first step to analyze the multi-task generalization issue in knowledge editing. Specifically, we develop an instruction-based editing technique, termed InstructEdit, which facilitates the editor's adaptation to various task performances simultaneously using simple instructions. With only one unified editor for each LLM, we empirically demonstrate that InstructEdit can improve the editor's control, leading to an average 14.86% increase in Reliability in multi-task editing setting. Furthermore, experiments involving holdout unseen task illustrate that InstructEdit consistently surpass previous strong baselines. To further investigate the underlying mechanisms of instruction-based knowledge editing, we analyze the principal components of the editing gradient directions, which unveils that instructions can help control optimization direction with stronger OOD generalization. Code and datasets will be available in https://github.com/zjunlp/EasyEdit.

link: http://arxiv.org/abs/2402.16123v1

AVI-Talking: Learning Audio-Visual Instructions for Expressive 3D Talking Face Generation

Yasheng Sun, Wenqing Chu, Hang Zhou, Kaisiyuan Wang, Hideki Koike

While considerable progress has been made in achieving accurate lip synchronization for 3D speech-driven talking face generation, the task of incorporating expressive facial detail synthesis aligned with the speaker's speaking status remains challenging. Our goal is to directly leverage the inherent style information conveyed by human speech for generating an expressive talking face that aligns with the speaking status. In this paper, we propose AVI-Talking, an Audio-Visual Instruction system for expressive Talking face generation. This system harnesses the robust contextual reasoning and hallucination capability offered by Large Language Models (LLMs) to instruct the realistic synthesis of 3D talking faces. Instead of directly learning facial movements from human speech, our two-stage strategy involves the LLMs first comprehending audio information and generating instructions implying expressive facial details seamlessly corresponding to the speech. Subsequently, a diffusion-based generative network executes these instructions. This two-stage process, coupled with the incorporation of LLMs, enhances model interpretability and provides users with flexibility to comprehend instructions and specify desired operations or modifications. Extensive experiments showcase the effectiveness of our approach in producing vivid talking faces with expressive facial movements and consistent emotional status.

link: http://arxiv.org/abs/2402.16124v1

A statistical method for crack detection in 3D concrete images

Vitalii Makogin, Duc Nguyen, Evgeny Spodarev

In practical applications, effectively segmenting cracks in large-scale computed tomography (CT) images holds significant importance for understanding the structural integrity of materials. However, classical methods and Machine Learning algorithms often incur high computational costs when dealing with the substantial size of input images. Hence, a robust algorithm is needed to pre-detect crack regions, enabling focused analysis and reducing computational overhead. The proposed approach addresses this challenge by offering a streamlined method for identifying crack regions in CT images with high probability. By efficiently identifying areas of interest, our algorithm allows for a more focused examination of potential anomalies within the material structure. Through comprehensive testing on both semi-synthetic and real 3D CT images, we validate the efficiency of our approach in enhancing crack segmentation while reducing computational resource requirements.

link: http://arxiv.org/abs/2402.16126v1

A VAE-based Framework for Learning Multi-Level Neural Granger-Causal Connectivity

Jiahe Lin, Huitian Lei, George Michailidis

Granger causality has been widely used in various application domains to capture lead-lag relationships amongst the components of complex dynamical systems, and the focus in extant literature has been on a single dynamical system. In certain applications in macroeconomics and neuroscience, one has access to data from a collection of related such systems, wherein the modeling task of interest is to extract the shared common structure that is embedded across them, as well as to identify the idiosyncrasies within individual ones. This paper introduces a Variational Autoencoder (VAE) based framework that jointly learns Granger-causal relationships amongst components in a collection of related-yet-heterogeneous dynamical systems, and handles the aforementioned task in a principled way. The performance of the proposed framework is evaluated on several synthetic data settings and benchmarked against existing approaches designed for individual system learning. The method is further illustrated on a real dataset involving time series data from a neurophysiological experiment and produces interpretable results.

link: http://arxiv.org/abs/2402.16131v1

LSTPrompt: Large Language Models as Zero-Shot Time Series Forecasters by Long-Short-Term Prompting

Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshavardhan Kamarthi, B. Aditya Prakash

Time-series forecasting (TSF) finds broad applications in real-world scenarios. Prompting off-the-shelf Large Language Models (LLMs) demonstrates strong zero-shot TSF capabilities while preserving computational efficiency. However, existing prompting methods oversimplify TSF as language next-token predictions, overlooking its dynamic nature and lack of integration with state-of-the-art prompt strategies such as Chain-of-Thought. Thus, we propose LSTPrompt, a novel approach for prompting LLMs in zero-shot TSF tasks. LSTPrompt decomposes TSF into short-term and long-term forecasting sub-tasks, tailoring prompts to each. LSTPrompt guides LLMs to regularly reassess forecasting mechanisms to enhance adaptability. Extensive evaluations demonstrate consistently better performance of LSTPrompt than existing prompting methods, and competitive results compared to foundation TSF models.

link: http://arxiv.org/abs/2402.16132v1

What Generative Artificial Intelligence Means for Terminological Definitions

Antonio San Martín

This paper examines the impact of Generative Artificial Intelligence (GenAI) on the creation and consumption of terminological definitions. GenAI tools like ChatGPT present a mix of benefits and drawbacks compared to traditional terminological resources. ChatGPT excels in providing context-specific meanings in an interactive and customized fashion but faces challenges with accuracy. Terminological definitions in recognized resources will likely survive because of their reliability. From the point of view of the terminologist, tools like ChatGPT enable AI-assisted terminography, including post-editing terminography, as an approach blending AI efficiency with human expertise for faster definition creation.

link: http://arxiv.org/abs/2402.16139v1

PeriodicLoRA: Breaking the Low-Rank Bottleneck in LoRA Optimization

Xiangdi Meng, Damai Dai, Weiyao Luo, Zhe Yang, Shaoxiang Wu, Xiaochen Wang, Peiyi Wang, Qingxiu Dong, Liang Chen, Zhifang Sui

Supervised fine-tuning is the most common method to adapt large language models (LLMs) to downstream tasks, but full fine-tuning LLMs requires massive computational resources. Recently, parameter-efficient fine-tuning (PEFT) methods have been widely studied due to its cost-effectiveness. LoRA is one of the most widely used methods, which assumes that the optimization process is essentially low-dimensional. Although LoRA fine-tuning is effective, there is still a performance gap compared to full fine-tuning, since its weight update is limited to low-rank matrices. In order to break the low-rank bottleneck in LoRA Optimization, we propose PeriodicLoRA (PLoRA), which accumulates low-rank update matrices multiple times to achieve a higher update rank. PLoRA has multiple training stages. During each stage, we still update only the LoRA weights. However, at the end of each stage, we unload the LoRA weights into the backbone parameters and then reinitialize the LoRA states. Experimental results show that PLoRA has stronger learning ability, approximately 1.8 times that of LoRA's learning ability at most, but it does not increase memory usage. Further, we introduce a momentum-based unloading strategy for PLoRA to mitigate the training instability.

link: http://arxiv.org/abs/2402.16141v1

From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility

Pravneet Kaur, Gautam Siddharth Kashyap, Ankit Kumar, Md Tabrez Nafis, Sandeep Kumar, Vikrant Shokeen

This groundbreaking study explores the expanse of Large Language Models (LLMs), such as Generative Pre-Trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) across varied domains ranging from technology, finance, healthcare to education. Despite their established prowess in Natural Language Processing (NLP), these LLMs have not been systematically examined for their impact on domains such as fitness, and holistic well-being, urban planning, climate modelling as well as disaster management. This review paper, in addition to furnishing a comprehensive analysis of the vast expanse and extent of LLMs' utility in diverse domains, recognizes the research gaps and realms where the potential of LLMs is yet to be harnessed. This study uncovers innovative ways in which LLMs can leave a mark in the fields like fitness and wellbeing, urban planning, climate modelling and disaster response which could inspire future researches and applications in the said avenues.

link: http://arxiv.org/abs/2402.16142v1

Egalitarian Price of Fairness for Indivisible Goods

Karen Frilya Celine, Muhammad Ayaz Dzulfikar, Ivan Adrian Koswara

In the context of fair division, the concept of price of fairness has been introduced to quantify the loss of welfare when we have to satisfy some fairness condition. In other words, it is the price we have to pay to guarantee fairness. Various settings of fair division have been considered previously; we extend to the setting of indivisible goods by using egalitarian welfare as the welfare measure, instead of the commonly used utilitarian welfare. We provide lower and upper bounds for various fairness and efficiency conditions such as envy-freeness up to one good (EF1) and maximum Nash welfare (MNW).

link: http://dx.doi.org/10.1007/978-981-99-7019-3_3