Fri 2024.03.01

Large Language Models As Evolution Strategies

Robert Tjarko Lange, Yingtao Tian, Yujin Tang

Large Transformer models are capable of implementing a plethora of so-called in-context learning algorithms. These include gradient descent, classification, sequence completion, transformation, and improvement. In this work, we investigate whether large language models (LLMs), which never explicitly encountered the task of black-box optimization, are in principle capable of implementing evolutionary optimization algorithms. While previous works have solely focused on language-based task specification, we move forward and focus on the zero-shot application of LLMs to black-box optimization. We introduce a novel prompting strategy, consisting of least-to-most sorting of discretized population members and querying the LLM to propose an improvement to the mean statistic, i.e. perform a type of black-box recombination operation. Empirically, we find that our setup allows the user to obtain an LLM-based evolution strategy, which we call `EvoLLM', that robustly outperforms baseline algorithms such as random search and Gaussian Hill Climbing on synthetic BBOB functions as well as small neuroevolution tasks. Hence, LLMs can act as `plug-in' in-context recombination operators. We provide several comparative studies of the LLM's model size, prompt strategy, and context construction. Finally, we show that one can flexibly improve EvoLLM's performance by providing teacher algorithm information via instruction fine-tuning on previously collected teacher optimization trajectories.

link: http://arxiv.org/abs/2402.18381v1

Robust Quantification of Percent Emphysema on CT via Domain Attention: the Multi-Ethnic Study of Atherosclerosis (MESA) Lung Study

Xuzhe Zhang, Elsa D. Angelini, Eric A. Hoffman, Karol E. Watson, Benjamin M. Smith, R. Graham Barr, Andrew F. Laine

Robust quantification of pulmonary emphysema on computed tomography (CT) remains challenging for large-scale research studies that involve scans from different scanner types and for translation to clinical scans. Existing studies have explored several directions to tackle this challenge, including density correction, noise filtering, regression, hidden Markov measure field (HMMF) model-based segmentation, and volume-adjusted lung density. Despite some promising results, previous studies either required a tedious workflow or limited opportunities for downstream emphysema subtyping, limiting efficient adaptation on a large-scale study. To alleviate this dilemma, we developed an end-to-end deep learning framework based on an existing HMMF segmentation framework. We first demonstrate that a regular UNet cannot replicate the existing HMMF results because of the lack of scanner priors. We then design a novel domain attention block to fuse image feature with quantitative scanner priors which significantly improves the results.

link: http://arxiv.org/abs/2402.18383v1

The First Place Solution of WSDM Cup 2024: Leveraging Large Language Models for Conversational Multi-Doc QA

Yiming Li, Zhao Zhang

Conversational multi-doc question answering aims to answer specific questions based on the retrieved documents as well as the contextual conversations. In this paper, we introduce our winning approach for the "Conversational Multi-Doc QA" challenge in WSDM Cup 2024, which exploits the superior natural language understanding and generation capability of Large Language Models (LLMs). We first adapt LLMs to the task, then devise a hybrid training strategy to make the most of in-domain unlabeled data. Moreover, an advanced text embedding model is adopted to filter out potentially irrelevant documents and several approaches are designed and compared for the model ensemble. Equipped with all these techniques, our solution finally ranked 1st place in WSDM Cup 2024, surpassing its rivals to a large extent. The source codes have been released at https://github.com/zhangzhao219/WSDM-Cup-2024.

link: http://arxiv.org/abs/2402.18385v1

TrustRate: A Decentralized Platform for Hijack-Resistant Anonymous Reviews

Rohit Dwivedula, Sriram Sridhar, Sambhav Satija, Muthian Sivathanu, Nishanth Chandran, Divya Gupta, Satya Lokam

Reviews and ratings by users form a central component in several widely used products today (e.g., product reviews, ratings of online content, etc.), but today's platforms for managing such reviews are ad-hoc and vulnerable to various forms of tampering and hijack by fake reviews either by bots or motivated paid workers. We define a new metric called 'hijack-resistance' for such review platforms, and then present TrustRate, an end-to-end decentralized, hijack-resistant platform for authentic, anonymous, tamper-proof reviews. With a prototype implementation and evaluation at the scale of thousands of nodes, we demonstrate the efficacy and performance of our platform, towards a new paradigm for building products based on trusted reviews by end users without having to trust a single organization that manages the reviews.

link: http://arxiv.org/abs/2402.18386v1

Neuromorphic Event-Driven Semantic Communication in Microgrids

Xiaoguang Diao, Yubo Song, Subham Sahoo, Yuan Li

Synergies between advanced communications, computing and artificial intelligence are unraveling new directions of coordinated operation and resiliency in microgrids. On one hand, coordination among sources is facilitated by distributed, privacy-minded processing at multiple locations, whereas on the other hand, it also creates exogenous data arrival paths for adversaries that can lead to cyber-physical attacks amongst other reliability issues in the communication layer. This long-standing problem necessitates new intrinsic ways of exchanging information between converters through power lines to optimize the system's control performance. Going beyond the existing power and data co-transfer technologies that are limited by efficiency and scalability concerns, this paper proposes neuromorphic learning to implant communicative features using spiking neural networks (SNNs) at each node, which is trained collaboratively in an online manner simply using the power exchanges between the nodes. As opposed to the conventional neuromorphic sensors that operate with spiking signals, we employ an event-driven selective process to collect sparse data for training of SNNs. Finally, its multi-fold effectiveness and reliable performance is validated under simulation conditions with different microgrid topologies and components to establish a new direction in the sense-actuate-compute cycle for power electronic dominated grids and microgrids.

link: http://arxiv.org/abs/2402.18390v1

Unveiling the Potential of Robustness in Evaluating Causal Inference Models

Yiyan Huang, Cheuk Hang Leung, Siyi Wang, Yijun Li, Qi Wu

The growing demand for personalized decision-making has led to a surge of interest in estimating the Conditional Average Treatment Effect (CATE). The intersection of machine learning and causal inference has yielded various effective CATE estimators. However, deploying these estimators in practice is often hindered by the absence of counterfactual labels, making it challenging to select the desirable CATE estimator using conventional model selection procedures like cross-validation. Existing approaches for CATE estimator selection, such as plug-in and pseudo-outcome metrics, face two inherent challenges. Firstly, they are required to determine the metric form and the underlying machine learning models for fitting nuisance parameters or plug-in learners. Secondly, they lack a specific focus on selecting a robust estimator. To address these challenges, this paper introduces a novel approach, the Distributionally Robust Metric (DRM), for CATE estimator selection. The proposed DRM not only eliminates the need to fit additional models but also excels at selecting a robust CATE estimator. Experimental studies demonstrate the efficacy of the DRM method, showcasing its consistent effectiveness in identifying superior estimators while mitigating the risk of selecting inferior ones.

link: http://arxiv.org/abs/2402.18392v1

Evaluating Decision Optimality of Autonomous Driving via Metamorphic Testing

Mingfei Cheng, Yuan Zhou, Xiaofei Xie, Junjie Wang, Guozhu Meng, Kairui Yang

Autonomous Driving System (ADS) testing is crucial in ADS development, with the current primary focus being on safety. However, the evaluation of non-safety-critical performance, particularly the ADS's ability to make optimal decisions and produce optimal paths for autonomous vehicles (AVs), is equally vital to ensure the intelligence and reduce risks of AVs. Currently, there is little work dedicated to assessing ADSs' optimal decision-making performance due to the lack of corresponding oracles and the difficulty in generating scenarios with non-optimal decisions. In this paper, we focus on evaluating the decision-making quality of an ADS and propose the first method for detecting non-optimal decision scenarios (NoDSs), where the ADS does not compute optimal paths for AVs. Firstly, to deal with the oracle problem, we propose a novel metamorphic relation (MR) aimed at exposing violations of optimal decisions. The MR identifies the property that the ADS should retain optimal decisions when the optimal path remains unaffected by non-invasive changes. Subsequently, we develop a new framework, Decictor, designed to generate NoDSs efficiently. Decictor comprises three main components: Non-invasive Mutation, MR Check, and Feedback. The Non-invasive Mutation ensures that the original optimal path in the mutated scenarios is not affected, while the MR Check is responsible for determining whether non-optimal decisions are made. To enhance the effectiveness of identifying NoDSs, we design a feedback metric that combines both spatial and temporal aspects of the AV's movement. We evaluate Decictor on Baidu Apollo, an open-source and production-grade ADS. The experimental results validate the effectiveness of Decictor in detecting non-optimal decisions of ADSs. Our work provides valuable and original insights into evaluating the non-safety-critical performance of ADSs.

link: http://arxiv.org/abs/2402.18393v1

Decomposed Prompting: Unveiling Multilingual Linguistic Structure Knowledge in English-Centric Large Language Models

Ercong Nie, Shuzhou Yuan, Bolei Ma, Helmut Schmid, Michael Färber, Frauke Kreuter, Hinrich Schütze

Despite the predominance of English in their training data, English-centric Large Language Models (LLMs) like GPT-3 and LLaMA display a remarkable ability to perform multilingual tasks, raising questions about the depth and nature of their cross-lingual capabilities. This paper introduces the decomposed prompting approach to probe the linguistic structure understanding of these LLMs in sequence labeling tasks. Diverging from the single text-to-text prompt, our method generates for each token of the input sentence an individual prompt which asks for its linguistic label. We assess our method on the Universal Dependencies part-of-speech tagging dataset for 38 languages, utilizing both English-centric and multilingual LLMs. Our findings show that decomposed prompting surpasses the iterative prompting baseline in efficacy and efficiency under zero- and few-shot settings. Further analysis reveals the influence of evaluation methods and the use of instructions in prompts. Our multilingual investigation shows that English-centric language models perform better on average than multilingual models. Our study offers insights into the multilingual transferability of English-centric LLMs, contributing to the understanding of their multilingual linguistic knowledge.

link: http://arxiv.org/abs/2402.18397v1

A Modular System for Enhanced Robustness of Multimedia Understanding Networks via Deep Parametric Estimation

Francesco Barbato, Umberto Michieli, Mehmet Kerim Yucel, Pietro Zanuttigh, Mete Ozay

In multimedia understanding tasks, corrupted samples pose a critical challenge, because when fed to machine learning models they lead to performance degradation. In the past, three groups of approaches have been proposed to handle noisy data: i) enhancer and denoiser modules to improve the quality of the noisy data, ii) data augmentation approaches, and iii) domain adaptation strategies. All the aforementioned approaches come with drawbacks that limit their applicability; the

first has high computational costs and requires pairs of clean-corrupted data for training, while the others only allow deployment of the same task/network they were trained on (\ie, when upstream and downstream task/network are the same). In this paper, we propose SyMPIE to solve these shortcomings. To this end, we design a small, modular, and efficient (just 2GFLOPs to process a Full HD image) system to enhance input data for robust downstream multimedia understanding with minimal computational cost. Our SyMPIE is pre-trained on an upstream task/network that should not match the downstream ones and does not need paired clean-corrupted samples. Our key insight is that most input corruptions found in real-world tasks can be modeled through global operations on color channels of images or spatial filters with small kernels. We validate our approach on multiple datasets and tasks, such as image classification (on ImageNetC, ImageNetC-Bar, VizWiz, and a newly proposed mixed corruption benchmark named ImageNetC-mixed) and semantic segmentation (on Cityscapes, ACDC, and DarkZurich) with consistent improvements of about 5\% relative accuracy gain across the board. The code of our approach and the new ImageNetC-mixed benchmark will be made available upon publication.

link: http://dx.doi.org/10.1145/3625468.3647623

A Cognitive Evaluation Benchmark of Image Reasoning and Description for Large Vision Language Models

Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, Yanyi Chen

Large Vision Language Models (LVLMs), despite their recent success, are hardly comprehensively tested for their cognitive abilities. Inspired by the prevalent use of the "Cookie Theft" task in human cognition test, we propose a novel evaluation benchmark to evaluate high-level cognitive ability of LVLMs using images with rich semantics. It defines eight reasoning capabilities and consists of an image description task and a visual question answering task. Our evaluation on well-known LVLMs shows that there is still a large gap in cognitive ability between LVLMs and humans.

link: http://arxiv.org/abs/2402.18409v2

Unsupervised Cross-Domain Image Retrieval via Prototypical Optimal Transport

Bin Li, Ye Shi, Qian Yu, Jingya Wang

Unsupervised cross-domain image retrieval (UCIR) aims to retrieve images sharing the same category across diverse domains without relying on labeled data. Prior approaches have typically decomposed the UCIR problem into two distinct tasks: intra-domain representation learning and cross-domain feature alignment. However, these segregated strategies overlook the potential synergies between these tasks. This paper introduces ProtoOT, a novel Optimal Transport formulation explicitly tailored for UCIR, which integrates intra-domain feature representation learning and cross-domain alignment into a unified framework. ProtoOT leverages the strengths of the K-means clustering method to effectively manage distribution imbalances inherent in UCIR. By utilizing K-means for generating initial prototypes and approximating class marginal distributions, we modify the constraints in Optimal Transport accordingly, significantly enhancing its performance in UCIR scenarios. Furthermore, we incorporate contrastive learning into the ProtoOT framework to further improve representation learning. This encourages local semantic consistency among features with similar semantics, while also explicitly enforcing separation between features and unmatched prototypes, thereby enhancing global discriminativeness. ProtoOT surpasses existing state-of-the-art methods by a notable margin across benchmark datasets. Notably, on DomainNet, ProtoOT achieves an average P@200 enhancement of 24.44%, and on Office-Home, it demonstrates a P@15 improvement of 12.12%. Code is available at https://github.com/HCVLAB/ProtoOT.

link: http://arxiv.org/abs/2402.18411v1

Prediction of recurrence free survival of head and neck cancer using PET/CT radiomics and clinical information

Mona Furukawa, Daniel R. McGowan, Bart■omiej W. Papie■

The 5-year survival rate of Head and Neck Cancer (HNC) has not improved over the past decade and one common cause of treatment failure is recurrence. In this paper, we built Cox proportional hazard (CoxPH) models that predict the recurrence free survival (RFS) of oropharyngeal HNC patients. Our models utilise both clinical information and multimodal radiomics features extracted from tumour regions in Computed Tomography (CT) and Positron Emission Tomography (PET). Furthermore, we were one of the first studies to explore the impact of segmentation accuracy on the predictive power of the extracted radiomics features, through under- and over-segmentation study. Our models were trained using the HEad and neCK TumOR (HECKTOR) challenge data, and the best performing model achieved a concordance index (C-index) of 0.74 for the model utilising clinical information and multimodal CT and PET radiomics features, which compares favourably with the model that only used clinical information (C-index of 0.67). Our under- and over-segmentation study confirms that segmentation accuracy affects radiomics extraction, however, it affects PET and CT differently.

link: http://arxiv.org/abs/2402.18417v1

Can GPT Improve the State of Prior Authorization via Guideline Based Automated Question Answering?

Shubham Vatsal, Ayush Singh, Shabnam Tafreshi

Health insurance companies have a defined process called prior authorization (PA) which is a health plan cost-control process that requires doctors and other healthcare professionals to get clearance in advance from a health plan before performing a particular procedure on a patient in order to be eligible for payment coverage. For health insurance companies, approving PA requests for patients in the medical domain is a time-consuming and challenging task. One of those key challenges is validating if a request matches up to certain criteria such as age, gender, etc. In this work, we evaluate whether GPT can validate numerous key factors, in turn helping health plans reach a decision drastically faster. We frame it as a question answering task, prompting GPT to answer a question from patient electronic health record. We experiment with different conventional prompting techniques as well as introduce our own novel prompting technique. Moreover, we report qualitative assessment by humans on the natural language generation outputs from our approach. Results show that our method achieves superior performance with the mean weighted F1 score of 0.61 as compared to its standard counterparts.

link: http://arxiv.org/abs/2402.18419v1

Emotion Classification in Low and Moderate Resource Languages

Shabnam Tafreshi, Shubham Vatsal, Mona Diab

It is important to be able to analyze the emotional state of people around the globe. There are 7100+ active languages spoken around the world and building emotion classification for each language is labor intensive. Particularly for low-resource and endangered languages, building emotion classification can be quite challenging. We present a cross-lingual emotion classifier, where we train an emotion classifier with resource-rich languages (i.e. \textit{English} in our work) and transfer the learning to low and moderate resource languages. We compare and contrast two approaches of transfer learning from a high-resource language to a low or moderate-resource language. One approach projects the annotation from a high-resource language to low and moderate-resource language in parallel corpora and the other one uses direct transfer from high-resource language to the other languages. We show the efficacy of our approaches on 6 languages: Farsi, Arabic, Spanish, Ilocano, Odia, and Azerbaijani. Our results indicate that our approaches outperform random baselines and transfer emotions across languages successfully. For all languages, the direct cross-lingual transfer of emotion yields better results. We also create annotated emotion-labeled resources for four languages: Farsi, Azerbaijani, Ilocano and Odia.

link: http://arxiv.org/abs/2402.18424v1

A Relational Inductive Bias for Dimensional Abstraction in Neural Networks

Declan Campbell, Jonathan D. Cohen

The human cognitive system exhibits remarkable flexibility and generalization capabilities, partly due to its ability to form low-dimensional, compositional representations of the environment. In contrast, standard neural network architectures often struggle with abstract reasoning tasks, overfitting, and requiring extensive data for training. This paper investigates the impact of the relational bottleneck -- a mechanism that focuses processing on relations among inputs -- on the learning of factorized representations conducive to compositional coding and the attendant flexibility of processing. We demonstrate that such a bottleneck not only improves generalization and learning efficiency, but also aligns network performance with human-like behavioral biases. Networks trained with the relational bottleneck developed orthogonal representations of feature dimensions latent in the dataset, reflecting the factorized structure thought to underlie human cognitive flexibility. Moreover, the relational network mimics human biases towards regularity without pre-specified symbolic primitives, suggesting that the bottleneck fosters the emergence of abstract representations that confer flexibility akin to symbols.

link: http://arxiv.org/abs/2402.18426v1