Thu 2024.05.02

PECC: Problem Extraction and Coding Challenges

Patrick Haller, Jonas Golde, Alan Akbik

Recent advancements in large language models (LLMs) have showcased their exceptional abilities across various tasks, such as code generation, problem-solving and reasoning. Existing benchmarks evaluate tasks in isolation, yet the extent to which LLMs can understand prose-style tasks, identify the underlying problems, and then generate appropriate code solutions is still unexplored. Addressing this gap, we introduce PECC, a novel benchmark derived from Advent Of Code (AoC) challenges and Project Euler, including 2396 problems. Unlike conventional benchmarks, PECC requires LLMs to interpret narrative-embedded problems, extract requirements, and generate executable code. A key feature of our dataset is the complexity added by natural language prompting in chat-based evaluations, mirroring real-world instruction ambiguities. Results show varying model performance between narrative and neutral problems, with specific challenges in the Euler math-based subset with GPT-3.5-Turbo passing 50% of the AoC challenges and only 8% on the Euler problems. By probing the limits of LLMs' capabilities, our benchmark provides a framework to monitor and assess the subsequent progress of LLMs as a universal problem solver.

link: http://arxiv.org/abs/2404.18766v1

Learning with Norm Constrained, Over-parameterized, Two-layer Neural Networks

Fanghui Liu, Leello Dadi, Volkan Cevher

Recent studies show that a reproducing kernel Hilbert space (RKHS) is not a suitable space to model functions by neural networks as the curse of dimensionality (CoD) cannot be evaded when trying to approximate even a single ReLU neuron (Bach, 2017). In this paper, we study a suitable function space for over-parameterized two-layer neural networks with bounded norms (e.g., the path norm, the Barron norm) in the perspective of sample complexity and generalization properties. First, we show that the path norm (as well as the Barron norm) is able to obtain width-independence sample complexity bounds, which allows for uniform convergence guarantees. Based on this result, we derive the improved result of metric entropy for ϵ-covering up to $\mathcal{O}(\epsilon^{-\frac{2d}{d+2}})$ (d is the input dimension and the depending constant is at most polynomial order of d) via the convex hull technique, which demonstrates the separation with kernel methods with $\Omega(\epsilon^{-d})$ to learn the target function in a Barron space. Second, this metric entropy result allows for building a sharper generalization bound under a general moment hypothesis setting, achieving the rate at $\mathcal{O}(n^{-\frac{d+2}{2d+2}})$. Our analysis is novel in that it offers a sharper and refined estimation for metric entropy (with a clear dependence relationship on the dimension d) and unbounded sampling in the estimation of the sample error and the output error.

link: http://arxiv.org/abs/2404.18769v1

Saliency Suppressed, Semantics Surfaced: Visual Transformations in Neural Networks and the Brain

Gustaw Opie■ka, Jessica Loke, Steven Scholte

Deep learning algorithms lack human-interpretable accounts of how they transform raw visual input into a robust semantic understanding, which impedes comparisons between different architectures, training objectives, and the human brain. In this work, we take inspiration from neuroscience and employ representational approaches to shed light on how neural networks encode information at low (visual saliency) and high (semantic similarity) levels of abstraction. Moreover, we introduce a custom image dataset where we systematically manipulate salient and semantic information. We find that ResNets are more sensitive to saliency information than ViTs, when trained with object classification objectives. We uncover that networks suppress saliency in early layers, a process enhanced by natural language supervision (CLIP) in ResNets. CLIP also enhances semantic encoding in both architectures. Finally, we show that semantic encoding is a key factor in aligning

AI with human visual perception, while saliency suppression is a non-brain-like strategy.

link: http://arxiv.org/abs/2404.18772v1

A Universal Metric of Dataset Similarity for Cross-silo Federated Learning

Ahmed Elhussein, Gamze Gursoy

Federated Learning is increasingly used in domains such as healthcare to facilitate collaborative model training without data-sharing. However, datasets located in different sites are often non-identically distributed, leading to degradation of model performance in FL. Most existing methods for assessing these distribution shifts are limited by being dataset or task-specific. Moreover, these metrics can only be calculated by exchanging data, a practice restricted in many FL scenarios. To address these challenges, we propose a novel metric for assessing dataset similarity. Our metric exhibits several desirable properties for FL: it is dataset-agnostic, is calculated in a privacy-preserving manner, and is computationally efficient, requiring no model training. In this paper, we first establish a theoretical connection between our metric and training dynamics in FL. Next, we extensively evaluate our metric on a range of datasets including synthetic, benchmark, and medical imaging datasets. We demonstrate that our metric shows a robust and interpretable relationship with model performance and can be calculated in privacy-preserving manner. As the first federated dataset similarity metric, we believe this metric can better facilitate successful collaborations between sites.

link: http://arxiv.org/abs/2404.18773v1

Where on Earth Do Users Say They Are?: Geo-Entity Linking for Noisy Multilingual User Input

Tessa Masis, Brendan O'Connor

Geo-entity linking is the task of linking a location mention to the real-world geographic location. In this paper we explore the challenging task of geo-entity linking for noisy, multilingual social media data. There are few open-source multilingual geo-entity linking tools available and existing ones are often rule-based, which break easily in social media settings, or LLM-based, which are too expensive for large-scale datasets. We present a method which represents real-world locations as averaged embeddings from labeled user-input location names and allows for selective prediction via an interpretable confidence score. We show that our approach improves geo-entity linking on a global and multilingual social media dataset, and discuss progress and problems with evaluating at different geographic granularities.

link: http://arxiv.org/abs/2404.18784v1

Certification of Speaker Recognition Models to Additive Perturbations

Dmitrii Korzh, Elvir Karimov, Mikhail Pautov, Oleg Y. Rogov, Ivan Oseledets

Speaker recognition technology is applied in various tasks ranging from personal virtual assistants to secure access systems. However, the robustness of these systems against adversarial attacks, particularly to additive perturbations, remains a significant challenge. In this paper, we pioneer applying robustness certification techniques to speaker recognition, originally developed for the image domain. In our work, we cover this gap by transferring and improving randomized smoothing certification techniques against norm-bounded additive perturbations for classification and few-shot learning tasks to speaker recognition. We demonstrate the effectiveness of these methods on VoxCeleb 1 and 2 datasets for several models. We expect this work to improve voice-biometry robustness, establish a new certification benchmark, and accelerate research of certification methods in the audio domain.

link: http://arxiv.org/abs/2404.18791v1

Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, Patrick Lewis

As Large Language Models (LLMs) have become more advanced, they have outpaced our abilities to accurately evaluate their quality. Not only is finding data to adequately probe particular model properties difficult, but evaluating the correctness of a model's freeform generation alone is a challenge. To address this, many evaluations now rely on using LLMs themselves as judges to score the quality of outputs from other LLMs. Evaluations most commonly use a single large model like GPT4. While this method has grown in popularity, it is costly, has been shown to introduce intramodel bias, and in this work, we find that very large models are often unnecessary. We propose instead to evaluate models using a Panel of LLm evaluators (PoLL). Across three distinct judge settings and spanning six different datasets, we find that using a PoLL composed of a larger number of smaller models outperforms a single large judge, exhibits less intra-model bias due to its composition of disjoint model families, and does so while being over seven times less expensive.

link: http://arxiv.org/abs/2404.18796v2

Multi-Agent Synchronization Tasks

Rolando Fernandez, Garrett Warnell, Derrik E. Asher, Peter Stone

In multi-agent reinforcement learning (MARL), coordination plays a crucial role in enhancing agents' performance beyond what they could achieve through cooperation alone. The interdependence of agents' actions, coupled with the need for communication, leads to a domain where effective coordination is crucial. In this paper, we introduce and define $\textit{Multi-Agent Synchronization Tasks}$ (MSTs), a novel subset of multi-agent tasks. We describe one MST, that we call $\textit{Synchronized Predator-Prey}$, offering a detailed description that will serve as the basis for evaluating a selection of recent state-of-the-art (SOTA) MARL algorithms explicitly designed to address coordination challenges through the use of communication strategies. Furthermore, we present empirical evidence that reveals the limitations of the algorithms assessed to solve MSTs, demonstrating their inability to scale effectively beyond 2-agent coordination tasks in scenarios where communication is a requisite component. Finally, the results raise questions about the applicability of recent SOTA approaches for complex coordination tasks (i.e. MSTs) and prompt further exploration into the underlying causes of their limitations in this context.

link: http://arxiv.org/abs/2404.18798v1

A Partial Replication of MaskFormer in TensorFlow on TPUs for the TensorFlow Model Garden

Vishal Purohit, Wenxin Jiang, Akshath R. Ravikiran, James C. Davis

This paper undertakes the task of replicating the MaskFormer model a universal image segmentation model originally developed using the PyTorch framework, within the TensorFlow ecosystem, specifically optimized for execution on Tensor Processing Units (TPUs). Our implementation exploits the modular constructs available within the TensorFlow Model Garden (TFMG), encompassing elements such as the data loader, training orchestrator, and various architectural components, tailored and adapted to meet the specifications of the MaskFormer model. We address key challenges encountered during the replication, non-convergence issues, slow training, adaptation of loss functions, and the integration of TPU-specific functionalities. We verify our reproduced implementation and present qualitative results on the COCO dataset. Although our implementation meets some of the objectives for end-to-end reproducibility, we encountered challenges in replicating the PyTorch version of MaskFormer in TensorFlow. This replication process is not straightforward and requires substantial engineering efforts. Specifically, it necessitates the customization of various components within the TFMG, alongside thorough verification and hyper-parameter tuning. The replication is available at: https://github.com/PurdueDualityLab/tf-maskformer/tree/main/official/projects/maskformer

link: http://arxiv.org/abs/2404.18801v1

Unknown Script: Impact of Script on Cross-Lingual Transfer

Wondimagegnhue Tsegaye Tufa, Ilia Markov, Piek Vossen

Cross-lingual transfer has become an effective way of transferring knowledge between languages. In this paper, we explore an often-overlooked aspect in this domain: the influence of the source language of the base language model on transfer performance. We conduct a series of experiments to determine the effect of the script and tokenizer used in the pre-trained model on the performance of the downstream task. Our findings reveal the importance of the tokenizer as a stronger factor than the sharing of the script, the language typology match, and the model size.

link: http://arxiv.org/abs/2404.18810v1

Safe Reach Set Computation via Neural Barrier Certificates

Alessandro Abate, Sergiy Bogomolov, Alec Edwards, Kostiantyn Potomkin, Sadegh Soudjani, Paolo Zuliani

We present a novel technique for online safety verification of autonomous systems, which performs reachability analysis efficiently for both bounded and unbounded horizons by employing neural barrier certificates. Our approach uses barrier certificates given by parameterized neural networks that depend on a given initial set, unsafe sets, and time horizon. Such networks are trained efficiently offline using system simulations sampled from regions of the state space. We then employ a meta-neural network to generalize the barrier certificates to state space regions that are outside the training set. These certificates are generated and validated online as sound over-approximations of the reachable states, thus either ensuring system safety or activating appropriate alternative actions in unsafe scenarios. We demonstrate our technique on case studies from linear models to nonlinear control-dependent models for online autonomous driving scenarios.

link: http://arxiv.org/abs/2404.18813v1

Hiding from Facebook: An Encryption Protocol resistant to Correlation Attacks

Chen-Da Liu, Simone Santini

In many social networks, one publishes information that one wants to reveal (e.g., the photograph of some friends) together with information that may lead to privacy breaches (e.g., the name of these people). One might want to hide this sensitive information by encrypting it and sharing the decryption key only with trusted people, but this might not be enough. If the cipher associated to a face is always the same, correlation between the output of a face recognition system and the cipher can give useful clues and help train recognizers to identify untagged instances of the face. We refer to these as "correlation attacks". In this paper we present a coding system that attempts to counter correlation attacks by associating to each instance of a face a different encryption of the same tag in such a way that the correlation between different instances is minimal. In addition, we present a key distribution code that allows only the owner of the images to encode the tags, but allows a group of trusted friends to decode them.

link: http://arxiv.org/abs/2404.18817v1

Towards Extreme Image Compression with Latent Feature Guidance and Diffusion Prior

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, Jingwen Jiang

Compressing images at extremely low bitrates (below 0.1 bits per pixel (bpp)) is a significant challenge due to substantial information loss. Existing extreme image compression methods generally suffer from heavy compression artifacts or low-fidelity reconstructions. To address this problem, we propose a novel extreme image compression framework that combines compressive VAEs and pre-trained text-to-image diffusion models in an end-to-end manner. Specifically, we introduce a latent feature-guided compression module based on compressive VAEs. This module compresses images and initially decodes the compressed information into content variables. To enhance the alignment between content variables and the diffusion space, we introduce external guidance to modulate intermediate feature maps. Subsequently, we develop a conditional diffusion decoding module that leverages pre-trained diffusion models to further decode these content variables. To preserve the generative capability of pre-trained diffusion models, we keep their parameters fixed and use a control module to inject content information. We also design a space alignment loss to provide sufficient constraints for the latent feature-guided compression module. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches in terms of both visual performance and image fidelity at extremely low bitrates.

link: http://arxiv.org/abs/2404.18820v1

Control Policy Correction Framework for Reinforcement Learning-based Energy Arbitrage Strategies

Seyed Soroush Karimi Madahi, Gargya Gokhale, Marie-Sophie Verwee, Bert Claessens, Chris Develder

A continuous rise in the penetration of renewable energy sources, along with the use of the single imbalance pricing, provides a new opportunity for balance responsible parties to reduce their cost through energy arbitrage in the imbalance settlement mechanism. Model-free reinforcement learning (RL) methods are an appropriate choice for solving the energy arbitrage problem due to their outstanding performance in solving complex stochastic sequential problems. However, RL is rarely deployed in real-world applications since its learned policy does not necessarily guarantee safety during the execution phase. In this paper, we propose a new RL-based control framework for batteries to obtain a safe energy arbitrage strategy in the imbalance settlement mechanism. In our proposed control framework, the agent initially aims to optimize the arbitrage revenue. Subsequently, in the post-processing step, we correct (constrain) the learned policy following a knowledge distillation process based on properties that follow human intuition. Our post-processing step is a generic method and is not restricted to the energy arbitrage domain. We use the Belgian imbalance price of 2023 to evaluate the performance of our proposed framework. Furthermore, we deploy our proposed control framework on a real battery to show its capability in the real world.

link: http://arxiv.org/abs/2404.18821v2

Benchmarking Benchmark Leakage in Large Language Models

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, Pengfei Liu

Amid the expanding use of pre-training data, the phenomenon of benchmark dataset leakage has become increasingly prominent, exacerbated by opaque training processes and the often undisclosed inclusion of supervised data in contemporary Large Language Models (LLMs). This issue skews benchmark effectiveness and fosters potentially unfair comparisons, impeding the field's healthy development. To address this, we introduce a detection pipeline utilizing Perplexity and N-gram accuracy, two simple and scalable metrics that gauge a model's prediction precision on benchmark, to identify potential data leakages. By analyzing 31 LLMs under the context of mathematical reasoning, we reveal substantial instances of training even test set misuse, resulting in potentially unfair comparisons. These findings prompt us to offer several recommendations regarding model documentation, benchmark setup, and future evaluations. Notably, we propose the "Benchmark Transparency Card" to encourage clear documentation of benchmark utilization, promoting transparency and healthy developments of LLMs. we have made our leaderboard, pipeline implementation, and model predictions publicly available, fostering future research.

link: http://arxiv.org/abs/2404.18824v1