Fri 2024.04.19

HLAT: High-quality Large Language Model Pre-trained on AWS Trainium

Haozheng Fan, Hao Zhou, Guangtai Huang, Parameswaran Raman, Xinwei Fu, Gaurav Gupta, Dhananjay Ram, Yida Wang, Jun Huan

Getting large language models (LLMs) to perform well on the downstream tasks requires pre-training over trillions of tokens. This typically demands a large number of powerful computational devices in addition to a stable distributed training framework to accelerate the training. The growing number of applications leveraging AI/ML had led to a scarcity of the expensive conventional accelerators (such as GPUs), which begs the need for the alternative specialized-accelerators that are scalable and cost-efficient. AWS Trainium is the second-generation machine learning accelerator that has been purposely built for training large deep learning models. Its corresponding instance, Amazon EC2 trn1, is an alternative to GPU instances for LLM training. However, training LLMs with billions of parameters on trn1 is challenging due to its relatively nascent software ecosystem. In this paper, we showcase HLAT: a 7 billion parameter decoder-only LLM pre-trained using trn1 instances over 1.8 trillion tokens. The performance of HLAT is benchmarked against popular open source baseline models including LLaMA and OpenLLaMA, which have been trained on NVIDIA GPUs and Google TPUs, respectively. On various evaluation tasks, we show that HLAT achieves model quality on par with the baselines. We also share the best practice of using the Neuron Distributed Training Library (NDTL), a customized distributed training library for AWS Trainium to achieve efficient training. Our work demonstrates that AWS Trainium powered by the NDTL is able to successfully pre-train state-of-the-art LLM models with high performance and cost-effectiveness.

link: http://arxiv.org/abs/2404.10630v1

Contextrast: Contextual Contrastive Learning for Semantic Segmentation

Changki Sung, Wanhee Kim, Jungho An, Wooju Lee, Hyungtae Lim, Hyun Myung

Despite great improvements in semantic segmentation, challenges persist because of the lack of local/global contexts and the relationship between them. In this paper, we propose Contextrast, a contrastive learning-based semantic segmentation method that allows to capture local/global contexts and comprehend their relationships. Our proposed method comprises two parts: a) contextual contrastive learning (CCL) and b) boundary-aware negative (BANE) sampling. Contextual contrastive learning obtains local/global context from multi-scale feature aggregation and inter/intra-relationship of features for better discrimination capabilities. Meanwhile, BANE sampling selects embedding features along the boundaries of incorrectly predicted regions to employ them as harder negative samples on our contrastive learning, resolving segmentation issues along the boundary region by exploiting fine-grained details. We demonstrate that our Contextrast substantially enhances the performance of semantic segmentation networks, outperforming state-of-the-art contrastive learning approaches on diverse public datasets, e.g. Cityscapes, CamVid, PASCAL-C, COCO-Stuff, and ADE20K, without an increase in computational cost during inference.

link: http://arxiv.org/abs/2404.10633v1

A Cloud Resources Portfolio Optimization Business Model -- From Theory to Practice

Valentin Haag, Maximilian Kiessler, Benedikt Pittl, Erich Schikuta

Cloud resources have become increasingly important, with many businesses using cloud solutions to supplement or outright replace their existing IT infrastructure. However, as there is a plethora of providers with varying products, services, and markets, it has become increasingly more challenging to keep track of the best solutions for each application. Cloud service intermediaries aim to alleviate this problem by offering services that help users meet their requirements. This paper aims to lay the groundwork for developing a cloud portfolio management platform and its business model, defined via a business model canvas. Furthermore, a prototype of a platform is developed offering a cloud portfolio optimization service, using two algorithms developed in previous research to create suitable and well-utilized allocations for a customer's applications.

link: http://arxiv.org/abs/2404.10641v1

Self-playing Adversarial Language Game Enhances LLM Reasoning

Pengyu Cheng, Tianhao Hu, Han Xu, Zhisong Zhang, Yong Dai, Lei Han, Nan Du

We explore the self-play training procedure of large language models (LLMs) in a two-player adversarial language game called Adversarial Taboo. In this game, an attacker and a defender communicate with respect to a target word only visible to the attacker. The attacker aims to induce the defender to utter the target word unconsciously, while the defender tries to infer the target word from the attacker's utterances. To win the game, both players should have sufficient knowledge about the target word and high-level reasoning ability to infer and express in this information-reserved conversation. Hence, we are curious about whether LLMs' reasoning ability can be further enhanced by Self-Play in this Adversarial language Game (SPAG). With this goal, we let LLMs act as the attacker and play with a copy of itself as the defender on an extensive range of target words. Through reinforcement learning on the game outcomes, we observe that the LLMs' performance uniformly improves on a broad range of reasoning benchmarks. Furthermore, iteratively adopting this self-play process can continuously promote LLM's reasoning ability. The code is at https://github.com/Linear95/SPAG.

link: http://arxiv.org/abs/2404.10642v1

A Calibrated and Automated Simulator for Innovations in 5G

Conrado Boeira, Antor Hasan, Khaleda Papry, Yue Ju, Zhongwen Zhu, Israat Haque

The rise of 5G deployments has created the environment for many emerging technologies to flourish. Self-driving vehicles, Augmented and Virtual Reality, and remote operations are examples of applications that leverage 5G networks' support for extremely low latency, high bandwidth, and increased throughput. However, the complex architecture of 5G hinders innovation due to the lack of accessibility to testbeds or realistic simulators with adequate 5G functionalities. Also, configuring and managing simulators are complex and time consuming. Finally, the lack of adequate representative data hinders the data-driven designs in 5G campaigns. Thus, we calibrated a system-level open-source simulator, Simu5G, following 3GPP guidelines to enable faster innovation in the 5G domain. Furthermore, we developed an API for automatic simulator configuration without knowing the underlying architectural details. Finally, we demonstrate the usage of the calibrated and automated simulator by developing an ML-based anomaly detection in a 5G Radio Access Network (RAN).

link: http://arxiv.org/abs/2404.10643v1

Continuous Control Reinforcement Learning: Distributed Distributional DrQ Algorithms

Zehao Zhou

Distributed Distributional DrQ is a model-free and off-policy RL algorithm for continuous control tasks based on the state and observation of the agent, which is an actor-critic method with the data-augmentation and the distributional perspective of critic value function. Aim to learn to control the agent and master some tasks in a high-dimensional continuous space. DrQ-v2 uses DDPG as the backbone and achieves out-performance in various continuous control tasks. Here Distributed Distributional DrQ uses Distributed Distributional DDPG as the backbone, and this modification aims to achieve better performance in some hard continuous control tasks through the better expression ability of distributional value function and distributed actor policies.

link: http://arxiv.org/abs/2404.10645v1

Efficient Parking Search using Shared Fleet Data

Niklas Strauß, Lukas Rottkamp, Sebatian Schmoll, Matthias Schubert

Finding an available on-street parking spot is a relevant problem of day-to-day life. In recent years, cities such as Melbourne and San Francisco deployed sensors that provide real-time information about the occupation of parking spots. Finding a free parking spot in such a smart environment can be modeled and solved as a Markov decision process (MDP). The problem has to consider uncertainty as available parking spots might not remain available until arrival due to other vehicles also claiming spots in the meantime. Knowing the parking intention of every vehicle in the environment would eliminate this uncertainty. Unfortunately, it does currently not seem realistic to have such data from all vehicles. In contrast, acquiring data from a subset of vehicles or a vehicle fleet appears feasible and has the potential to reduce uncertainty. In this paper, we examine the question of how useful sharing data within a vehicle fleet might be for the search times of particular drivers. We use fleet data to better estimate the availability of parking spots at arrival. Since optimal solutions for large scenarios are infeasible, we base our method on approximate solutions, which have been shown to perform well in single-agent settings. Our experiments are conducted on a simulation using real-world and synthetic data from the city of Melbourne. The results indicate that fleet data can significantly reduce search times for an available parking spot.

link: http://dx.doi.org/10.1109/MDM52706.2021.00026

ViTextVQA: A Large-Scale Visual Question Answering Dataset for Evaluating Vietnamese Text Comprehension in Images

Quan Van Nguyen, Dan Quang Tran, Huy Quang Pham, Thang Kien-Bao Nguyen, Nghia Hieu Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen

Visual Question Answering (VQA) is a complicated task that requires the capability of simultaneously processing natural language and images. Initially, this task was researched, focusing on methods to help machines understand objects and scene contexts in images. However, some text appearing in the image that carries explicit information about the full content of the image is not mentioned. Along with the continuous development of the AI era, there have been many studies on the reading comprehension ability of VQA models in the world. As a developing country, conditions are still limited, and this task is still open in Vietnam. Therefore, we introduce the first large-scale dataset in Vietnamese specializing in the ability to understand text appearing in images, we call it ViTextVQA (\textbf{Vi}etnamese \textbf{Text}-based \textbf{V}isual \textbf{Q}uestion \textbf{A}nswering dataset) which contains \textbf{over 16,000} images and \textbf{over 50,000} questions with answers. Through meticulous experiments with various state-of-the-art models, we uncover the significance of the order in which tokens in OCR text are processed and selected to formulate answers. This finding helped us significantly improve the performance of the baseline models on the ViTextVQA dataset. Our dataset is available at this \href{https://github.com/minhquan6203/ViTextVQA-Dataset}{link} for research purposes.

link: http://arxiv.org/abs/2404.10652v1

Continual Offline Reinforcement Learning via Diffusion-based Dual Generative Replay

Jinmei Liu, Wenbin Li, Xiangyu Yue, Shilin Zhang, Chunlin Chen, Zhi Wang

We study continual offline reinforcement learning, a practical paradigm that facilitates forward transfer and mitigates catastrophic forgetting to tackle sequential offline tasks. We propose a dual generative replay framework that retains previous knowledge by concurrent replay of generated pseudo-data. First, we decouple the continual learning policy into a diffusion-based generative behavior model and a multi-head action evaluation model, allowing the policy to inherit distributional expressivity for encompassing a progressive range of diverse behaviors. Second, we train a task-conditioned diffusion model to mimic state distributions of past tasks. Generated states are paired with corresponding responses from the behavior generator to represent old tasks with high-fidelity replayed samples. Finally, by interleaving pseudo samples with real ones of the new task, we continually update the state and behavior generators to model progressively diverse behaviors, and regularize the multi-head critic via behavior cloning to mitigate forgetting. Experiments demonstrate that our method achieves better forward transfer with less forgetting, and closely approximates the results of using previous ground-truth data due to its high-fidelity replay of the sample space. Our code is available at \href{https://github.com/NJU-RL/CuGRO}{https://github.com/NJU-RL/CuGRO}.

link: http://arxiv.org/abs/2404.10662v2

Assessing The Impact of CNN Auto Encoder-Based Image Denoising on Image Classification Tasks

Mohsen Hami, Mahdi JameBozorg

Images captured from the real world are often affected by different types of noise, which can significantly impact the performance of Computer Vision systems and the quality of visual data. This study presents a novel approach for defect detection in casting product noisy images, specifically focusing on submersible pump impellers. The methodology involves utilizing deep learning models such as VGG16, InceptionV3, and other models in both the spatial and frequency domains to identify noise types and defect status. The research process begins with preprocessing images, followed by applying denoising techniques tailored to specific noise categories. The goal is to enhance the accuracy and robustness of defect detection by integrating noise detection and denoising into the classification pipeline. The study achieved remarkable results using VGG16 for noise type classification in the frequency domain, achieving an accuracy of over 99%. Removal of salt and pepper noise resulted in an average SSIM of 87.9, while Gaussian noise removal had an average SSIM of 64.0, and periodic noise removal yielded an average SSIM of 81.6. This comprehensive approach showcases the effectiveness of the deep AutoEncoder model and median filter, for denoising strategies in real-world industrial applications. Finally, our study reports significant improvements in binary classification accuracy for defect detection compared to previous methods. For the VGG16 classifier, accuracy increased from 94.6% to 97.0%, demonstrating the effectiveness of the proposed noise detection and denoising approach. Similarly, for the InceptionV3 classifier, accuracy improved from 84.7% to 90.0%, further validating the benefits of integrating noise analysis into the classification pipeline.

link: http://arxiv.org/abs/2404.10664v1

VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, Baining Guo

We introduce VASA, a framework for generating lifelike talking faces with appealing visual affective skills (VAS) given a single static image and a speech audio clip. Our premiere model, VASA-1, is capable of not only producing lip movements that are exquisitely synchronized with the audio, but also capturing a large spectrum of facial nuances and natural head motions that contribute to the perception of authenticity and liveliness. The core innovations include a holistic facial dynamics and head movement generation model that works in a face latent space, and the development of such an expressive and disentangled face latent space using videos. Through extensive experiments including evaluation on a set of new metrics, we show that our method significantly outperforms previous methods along various dimensions comprehensively. Our method not only delivers high video quality with realistic facial and head dynamics but also supports the online generation of 512x512 videos at up to 40 FPS with negligible starting latency. It paves the way for real-time engagements with lifelike avatars that emulate human conversational behaviors.

link: http://arxiv.org/abs/2404.10667v1

Automating REST API Postman Test Cases Using LLM

S Deepika Sri, Mohammed Aadil S, Sanjjushri Varshini R, Raja CSP Raman, Gopinath Rajagopal, S Taranath Chan

In the contemporary landscape of technological advancements, the automation of manual processes is crucial, compelling the demand for huge datasets to effectively train and test machines. This research paper is dedicated to the exploration and implementation of an automated approach to generate test cases specifically using Large Language Models. The methodology integrates the use of Open AI to enhance the efficiency and effectiveness of test case generation for training and evaluating Large Language Models. This formalized approach with LLMs simplifies the testing process, making it more efficient and comprehensive. Leveraging natural language understanding, LLMs can intelligently formulate test cases that cover a broad range of REST API properties, ensuring comprehensive testing. The model that is developed during the research is trained using manually collected postman test cases or instances for various Rest APIs. LLMs enhance the creation of Postman test cases by automating the generation of varied and intricate test scenarios. Postman test cases offer streamlined automation, collaboration, and dynamic data handling, providing a user-friendly and efficient approach to API testing compared to traditional test cases. Thus, the model developed not only conforms to current technological standards but also holds the promise of evolving into an idea of substantial importance in future technological advancements.

link: http://arxiv.org/abs/2404.10678v1

HSVI-based Online Minimax Strategies for Partially Observable Stochastic Games with Neural Perception Mechanisms

Rui Yan, Gabriel Santos, Gethin Norman, David Parker, Marta Kwiatkowska

We consider a variant of continuous-state partially-observable stochastic games with neural perception mechanisms and an asymmetric information structure. One agent has partial information, with the observation function implemented as a neural network, while the other agent is assumed to have full knowledge of the state. We present, for the first time, an efficient online method to compute an ε-minimax strategy profile, which requires only one linear program to be solved for each agent at every stage, instead of a complex estimation of opponent counterfactual values. For the partially-informed agent, we propose a continual resolving approach which uses lower bounds, pre-computed offline with heuristic search value iteration (HSVI), instead of opponent counterfactual values. This inherits the soundness of continual resolving at the cost of pre-computing the bound. For the fully-informed agent, we propose an inferred-belief strategy, where the agent maintains an inferred belief about the belief of the partially-informed agent based on (offline) upper bounds from HSVI, guaranteeing ε-distance to the value of the game at the initial belief known to both agents.

link: http://arxiv.org/abs/2404.10679v1

StyleCity: Large-Scale 3D Urban Scenes Stylization with Vision-and-Text Reference via Progressive Optimization

Yingshu Chen, Huajian Huang, Tuan-Anh Vu, Ka Chun Shum, Sai-Kit Yeung

Creating large-scale virtual urban scenes with variant styles is inherently challenging. To facilitate prototypes of virtual production and bypass the need for complex materials and lighting setups, we introduce the first vision-and-text-driven texture stylization system for large-scale urban scenes, StyleCity. Taking an image and text as references, StyleCity stylizes a 3D textured mesh of a large-scale urban scene in a semantics-aware fashion and generates a harmonic omnidirectional sky background. To achieve that, we propose to stylize a neural texture field by transferring 2D vision-and-text priors to 3D globally and locally. During 3D stylization, we progressively scale the planned training views of the input 3D scene at different levels in order to preserve high-quality scene content. We then optimize the scene style globally by adapting the scale of the style image with the scale of the training views. Moreover, we enhance local semantics consistency by the semantics-aware style loss which is crucial for photo-realistic stylization. Besides texture stylization, we further adopt a generative diffusion model to synthesize a style-consistent omnidirectional sky image, which offers a more immersive atmosphere and assists the semantic stylization process. The stylized neural texture field can be baked into an arbitrary-resolution texture, enabling seamless integration into conventional rendering pipelines and significantly easing the virtual production prototyping process. Extensive experiments demonstrate our stylized scenes' superiority in qualitative and quantitative performance and user preferences.

link: http://arxiv.org/abs/2404.10681v1

Simplex Decomposition for Portfolio Allocation Constraints in Reinforcement Learning

David Winkel, Niklas Strauß, Matthias Schubert, Thomas Seidl

Portfolio optimization tasks describe sequential decision problems in which the investor's wealth is distributed across a set of assets. Allocation constraints are used to enforce minimal or maximal investments into particular subsets of assets to control for objectives such as limiting the portfolio's exposure to a certain sector due to environmental concerns. Although methods for constrained Reinforcement Learning (CRL) can optimize policies while considering allocation constraints, it can be observed that these general methods yield suboptimal results. In this paper, we propose a novel approach to handle allocation constraints based on a decomposition of the constraint action space into a set of unconstrained allocation problems. In particular, we examine this approach for the case of two constraints. For example, an investor may wish to invest at least a certain percentage of the portfolio into green technologies while limiting the investment in the fossil energy sector. We show that the action space of the task is equivalent to the decomposed action space, and introduce a new reinforcement learning (RL) approach CAOSD, which is built on top of the decomposition. The experimental evaluation on real-world Nasdaq-100 data demonstrates that our approach consistently outperforms state-of-the-art CRL benchmarks for portfolio optimization.

link: http://dx.doi.org/10.3233/FAIA230573