Fri 2024.04.26

Compete and Compose: Learning Independent Mechanisms for Modular World Models

Anson Lei, Frederik Nolte, Bernhard Schölkopf, Ingmar Posner

We present COmpetitive Mechanisms for Efficient Transfer (COMET), a modular world model which leverages reusable, independent mechanisms across different environments. COMET is trained on multiple environments with varying dynamics via a two-step process: competition and composition. This enables the model to recognise and learn transferable mechanisms. Specifically, in the competition phase, COMET is trained with a winner-takes-all gradient allocation, encouraging the emergence of independent mechanisms. These are then re-used in the composition phase, where COMET learns to re-compose learnt mechanisms in ways that capture the dynamics of intervened environments. In so doing, COMET explicitly reuses prior knowledge, enabling efficient and interpretable adaptation. We evaluate COMET on environments with image-based observations. In contrast to competitive baselines, we demonstrate that COMET captures recognisable mechanisms without supervision. Moreover, we show that COMET is able to adapt to new environments with varying numbers of objects with improved sample efficiency compared to more conventional finetuning approaches.

link: http://arxiv.org/abs/2404.15109v1

Taming Diffusion Probabilistic Models for Character Control

Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, Xuelin Chen

We present a novel character control framework that effectively utilizes motion diffusion probabilistic models to generate high-quality and diverse character animations, responding in real-time to a variety of dynamic user-supplied control signals. At the heart of our method lies a transformer-based Conditional Autoregressive Motion Diffusion Model (CAMDM), which takes as input the character's historical motion and can generate a range of diverse potential future motions conditioned on high-level, coarse user control. To meet the demands for diversity, controllability, and computational efficiency required by a real-time controller, we incorporate several key algorithmic designs. These include separate condition tokenization, classifier-free guidance on past motion, and heuristic future trajectory extension, all designed to address the challenges associated with taming motion diffusion probabilistic models for character control. As a result, our work represents the first model that enables real-time generation of high-quality, diverse character animations based on user interactive control, supporting animating the character in multiple styles with a single unified model. We evaluate our method on a diverse set of locomotion skills, demonstrating the merits of our method over existing character controllers. Project page and source codes: https://aiganimation.github.io/CAMDM/

link: http://arxiv.org/abs/2404.15121v1

MedDr: Diagnosis-Guided Bootstrapping for Large-Scale Medical Vision-Language Learning

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, Hao Chen

The rapid advancement of large-scale vision-language models has showcased remarkable capabilities across various tasks. However, the lack of extensive and high-quality image-text data in medicine has greatly hindered the development of large-scale medical vision-language models. In this work, we present a diagnosis-guided bootstrapping strategy that exploits both image and label information to construct vision-language datasets. Based on the constructed dataset, we developed MedDr, a generalist foundation model for healthcare capable of handling diverse medical data modalities, including radiology, pathology, dermatology, retinography, and endoscopy. Moreover, during inference, we propose a simple but effective retrieval-augmented medical diagnosis strategy, which enhances the model's generalization ability. Extensive experiments on visual question answering, medical report generation, and medical image diagnosis demonstrate the superiority of

our method.

link: http://arxiv.org/abs/2404.15127v1

Gallbladder Cancer Detection in Ultrasound Images based on YOLO and Faster R-CNN

Sara Dadjouy, Hedieh Sajedi

Medical image analysis is a significant application of artificial intelligence for disease diagnosis. A crucial step in this process is the identification of regions of interest within the images. This task can be automated using object detection algorithms. YOLO and Faster R-CNN are renowned for such algorithms, each with its own strengths and weaknesses. This study aims to explore the advantages of both techniques to select more accurate bounding boxes for gallbladder detection from ultrasound images, thereby enhancing gallbladder cancer classification. A fusion method that leverages the benefits of both techniques is presented in this study. The proposed method demonstrated superior classification performance, with an accuracy of 92.62%, compared to the individual use of Faster R-CNN and YOLOv8, which yielded accuracies of 90.16% and 82.79%, respectively.

link: http://dx.doi.org/10.1109/QICAR61538.2024.10496645

Black Hole Search by a Set of Scattered Agents in Dynamic Rings

Giuseppe Antonio Di Luna, Paola Flocchini, Giuseppe Prencipe, Nicola Santoro

In this paper we investigate the problem of searching for a black hole in a dynamic graph by a set of scattered agents (i.e., the agents start from arbitrary locations of the graph). The black hole is a node that silently destroys any agent visiting it. This kind of malicious node nicely models network failures such as a crashed host or a virus that erases the visiting agents. The black hole search problem is solved when at least one agent survives, and it has the entire map of the graph with the location of the black hole. We consider the case in which the underlining graph is a dynamic 1-interval connected ring: a ring graph in which at each round at most one edge can be missing. We first show that the problem cannot be solved if the agents can only communicate by using a face-to-face mechanism: this holds for any set of agents of constant size, with respect to the size n of the ring. To circumvent this impossibility we consider agents equipped with movable pebbles that can be left on nodes as a form of communication with other agents. When pebbles are available, three agents can localize the black hole in $O(n^2)$ moves. We show that such a number of agents is optimal. We also show that the complexity is tight, that is $\Omega(n^2)$ moves are required for any algorithm solving the problem with three agents, even with stronger communication mechanisms (e.g., a whiteboard on each node on which agents can write messages of unlimited size). To the best of our knowledge this is the first paper examining the problem of searching a black hole in a dynamic environment with scattered agents.

link: http://arxiv.org/abs/2404.15132v1

From Space-Time to Space-Order: Directly Planning a Temporal Planning Graph by Redefining CBS

Yu Wu, Rishi Veerapaneni, Jiaoyang Li, Maxim Likhachev

The majority of multi-agent path finding (MAPF) methods compute collision-free space-time paths which require agents to be at a specific location at a specific discretized timestep. However, executing these space-time paths directly on robotic systems is infeasible due to real-time execution differences (e.g. delays) which can lead to collisions. To combat this, current methods translate the space-time paths into a temporal plan graph (TPG) that only requires that agents observe the order in which they navigate through locations where their paths cross. However, planning space-time paths and then post-processing them into a TPG does not reduce the required agent-to-agent coordination, which is fixed once the space-time paths are computed. To that end, we propose a novel algorithm Space-Order CBS that can directly plan a TPG and explicitly minimize coordination. Our main theoretical insight is our novel perspective on viewing a TPG as a set of space-visitation order paths where agents visit locations in relative orders (e.g. 1st vs 2nd) as

opposed to specific timesteps. We redefine unique conflicts and constraints for adapting CBS for space-order planning. We experimentally validate how Space-Order CBS can return TPGs which significantly reduce coordination, thus subsequently reducing the amount of agent-agent communication and leading to more robustness to delays during execution.

link: http://arxiv.org/abs/2404.15137v1

CutDiffusion: A Simple, Fast, Cheap, and Strong Diffusion Extrapolation Method

Mingbao Lin, Zhihang Lin, Wengyi Zhan, Liujuan Cao, Rongrong Ji

Transforming large pre-trained low-resolution diffusion models to cater to higher-resolution demands, i.e., diffusion extrapolation, significantly improves diffusion adaptability. We propose tuning-free CutDiffusion, aimed at simplifying and accelerating the diffusion extrapolation process, making it more affordable and improving performance. CutDiffusion abides by the existing patch-wise extrapolation but cuts a standard patch diffusion process into an initial phase focused on comprehensive structure denoising and a subsequent phase dedicated to specific detail refinement. Comprehensive experiments highlight the numerous almighty advantages of CutDiffusion: (1) simple method construction that enables a concise higher-resolution diffusion process without third-party engagement; (2) fast inference speed achieved through a single-step higher-resolution diffusion process, and fewer inference patches required; (3) cheap GPU cost resulting from patch-wise inference and fewer patches during the comprehensive structure denoising; (4) strong generation performance, stemming from the emphasis on specific detail refinement.

link: http://arxiv.org/abs/2404.15141v1

Rethinking LLM Memorization through the Lens of Adversarial Compression

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C. Lipton, J. Zico Kolter

Large language models (LLMs) trained on web-scale datasets raise substantial concerns regarding permissible data usage. One major question is whether these models "memorize" all their training data or they integrate many data sources in some way more akin to how a human would learn and synthesize information. The answer hinges, to a large degree, on $\textit{how we define memorization}$. In this work, we propose the Adversarial Compression Ratio (ACR) as a metric for assessing memorization in LLMs -- a given string from the training data is considered memorized if it can be elicited by a prompt shorter than the string itself. In other words, these strings can be "compressed" with the model by computing adversarial prompts of fewer tokens. We outline the limitations of existing notions of memorization and show how the ACR overcomes these challenges by (i) offering an adversarial view to measuring memorization, especially for monitoring unlearning and compliance; and (ii) allowing for the flexibility to measure memorization for arbitrary strings at a reasonably low compute. Our definition serves as a valuable and practical tool for determining when model owners may be violating terms around data usage, providing a potential legal tool and a critical lens through which to address such scenarios. Project page: https://locuslab.github.io/acr-memorization.

link: http://arxiv.org/abs/2404.15146v1

Bias patterns in the application of LLMs for clinical decision support: A comprehensive study

Raphael Poulain, Hamed Fayyaz, Rahmatollah Beheshti

Large Language Models (LLMs) have emerged as powerful candidates to inform clinical decision-making processes. While these models play an increasingly prominent role in shaping the digital landscape, two growing concerns emerge in healthcare applications: 1) to what extent do LLMs exhibit social bias based on patients' protected attributes (like race), and 2) how do design choices (like architecture design and prompting strategies) influence the observed biases? To answer these questions rigorously, we evaluated eight popular LLMs across three question-answering (QA) datasets using clinical vignettes (patient descriptions) standardized for bias evaluations. We employ red-teaming strategies to analyze how demographics affect LLM outputs, comparing both general-purpose and clinically-trained models. Our extensive experiments

reveal various disparities (some significant) across protected groups. We also observe several counter-intuitive patterns such as larger models not being necessarily less biased and fined-tuned models on medical data not being necessarily better than the general-purpose models. Furthermore, our study demonstrates the impact of prompt design on bias patterns and shows that specific phrasing can influence bias patterns and reflection-type approaches (like Chain of Thought) can reduce biased outcomes effectively. Consistent with prior studies, we call on additional evaluations, scrutiny, and enhancement of LLMs used in clinical decision support applications.

link: http://arxiv.org/abs/2404.15149v1

Regressive Side Effects of Training Language Models to Mimic Student Misconceptions

Shashank Sonkar, Naiming Liu, Richard G. Baraniuk

This paper presents a novel exploration into the regressive side effects of training Large Language Models (LLMs) to mimic student misconceptions for personalized education. We highlight the problem that as LLMs are trained to more accurately mimic student misconceptions, there is a compromise in the factual integrity and reasoning ability of the models. Our work involved training an LLM on a student-tutor dialogue dataset to predict student responses. The results demonstrated a decrease in the model's performance across multiple benchmark datasets, including the ARC reasoning challenge and TruthfulQA, which evaluates the truthfulness of model's generated responses. Furthermore, the HaluEval Dial dataset, used for hallucination detection, and MemoTrap, a memory-based task dataset, also reported a decline in the model accuracy. To combat these side effects, we introduced a "hallucination token" technique. This token, appended at the beginning of each student response during training, instructs the model to switch between mimicking student misconceptions and providing factually accurate responses. Despite the significant improvement across all datasets, the technique does not completely restore the LLM's baseline performance, indicating the need for further research in this area. This paper contributes to the ongoing discussion on the use of LLMs for student modeling, emphasizing the need for a balance between personalized education and factual accuracy.

link: http://arxiv.org/abs/2404.15156v1

Combating Missing Modalities in Egocentric Videos at Test Time

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, Motasem Alfarra

Understanding videos that contain multiple modalities is crucial, especially in egocentric videos, where combining various sensory inputs significantly improves tasks like action recognition and moment localization. However, real-world applications often face challenges with incomplete modalities due to privacy concerns, efficiency needs, or hardware issues. Current methods, while effective, often necessitate retraining the model entirely to handle missing modalities, making them computationally intensive, particularly with large training datasets. In this study, we propose a novel approach to address this issue at test time without requiring retraining. We frame the problem as a test-time adaptation task, where the model adjusts to the available unlabeled data at test time. Our method, MiDl~(Mutual information with self-Distillation), encourages the model to be insensitive to the specific modality source present during testing by minimizing the mutual information between the prediction and the available modality. Additionally, we incorporate self-distillation to maintain the model's original performance when both modalities are available. MiDl represents the first self-supervised, online solution for handling missing modalities exclusively at test time. Through experiments with various pretrained models and datasets, MiDl demonstrates substantial performance improvement without the need for retraining.

link: http://arxiv.org/abs/2404.15161v1

Adaptive Mixed-Scale Feature Fusion Network for Blind AI-Generated Image Quality Assessment

Tianwei Zhou, Songbai Tan, Wei Zhou, Yu Luo, Yuan-Gen Wang, Guanghui Yue

With the increasing maturity of the text-to-image and image-to-image generative models, AI-generated images (AGIs) have shown great application potential in advertisement, entertainment, education, social media, etc. Although remarkable advancements have been achieved in generative models, very few efforts have been paid to design relevant quality assessment models. In this paper, we propose a novel blind image quality assessment (IQA) network, named AMFF-Net, for AGIs. AMFF-Net evaluates AGI quality from three dimensions, i.e., "visual quality", "authenticity", and "consistency". Specifically, inspired by the characteristics of the human visual system and motivated by the observation that "visual quality" and "authenticity" are characterized by both local and global aspects, AMFF-Net scales the image up and down and takes the scaled images and original-sized image as the inputs to obtain multi-scale features. After that, an Adaptive Feature Fusion (AFF) block is used to adaptively fuse the multi-scale features with learnable weights. In addition, considering the correlation between the image and prompt, AMFF-Net compares the semantic features from text encoder and image encoder to evaluate the text-to-image alignment. We carry out extensive experiments on three AGI quality assessment databases, and the experimental results show that our AMFF-Net obtains better performance than nine state-of-the-art blind IQA methods. The results of ablation experiments further demonstrate the effectiveness of the proposed multi-scale input strategy and AFF block.

link: http://arxiv.org/abs/2404.15163v1

Fourier-enhanced Implicit Neural Fusion Network for Multispectral and Hyperspectral Image Fusion

Yu-Jie Liang, Zihan Cao, Liang-Jian Deng, Xiao Wu

Recently, implicit neural representations (INR) have made significant strides in various vision-related domains, providing a novel solution for Multispectral and Hyperspectral Image Fusion (MHIF) tasks. However, INR is prone to losing high-frequency information and is confined to the lack of global perceptual capabilities. To address these issues, this paper introduces a Fourier-enhanced Implicit Neural Fusion Network (FeINFN) specifically designed for MHIF task, targeting the following phenomena: The Fourier amplitudes of the HR-HSI latent code and LR-HSI are remarkably similar; however, their phases exhibit different patterns. In FeINFN, we innovatively propose a spatial and frequency implicit fusion function (Spa-Fre IFF), helping INR capture high-frequency information and expanding the receptive field. Besides, a new decoder employing a complex Gabor wavelet activation function, called Spatial-Frequency Interactive Decoder (SFID), is invented to enhance the interaction of INR features. Especially, we further theoretically prove that the Gabor wavelet activation possesses a time-frequency tightness property that favors learning the optimal bandwidths in the decoder. Experiments on two benchmark MHIF datasets verify the state-of-the-art (SOTA) performance of the proposed method, both visually and quantitatively. Also, ablation studies demonstrate the mentioned contributions. The code will be available on Anonymous GitHub (https://anonymous.4open.science/r/FeINFN-15C9/) after possible acceptance.

link: http://arxiv.org/abs/2404.15174v1

Voice Passing : a Non-Binary Voice Gender Prediction System for evaluating Transgender voice transition

David Doukhan, Simon Devauchelle, Lucile Girard-Monneron, Mía Chávez Ruz, V. Chaddouk, Isabelle Wagner, Albert Rilliard

This paper presents a software allowing to describe voices using a continuous Voice Femininity Percentage (VFP). This system is intended for transgender speakers during their voice transition and for voice therapists supporting them in this process. A corpus of 41 French cis- and transgender speakers was recorded. A perceptual evaluation allowed 57 participants to estimate the VFP for each voice. Binary gender classification models were trained on external gender-balanced data and used on overlapping windows to obtain average gender prediction estimates, which were calibrated to predict VFP and obtained higher accuracy than F_0 or vocal track length-based models. Training data speaking style and DNN architecture were shown to impact VFP estimation. Accuracy of the models was affected by speakers' age. This highlights the importance of style, age, and the conception of gender as binary or not, to build adequate statistical

representations of cultural concepts.

link: http://dx.doi.org/10.21437/Interspeech.2023-1835

Closed Loop Interactive Embodied Reasoning for Robot Manipulation

Michal Nazarczuk, Jan Kristof Behrens, Karla Stepanova, Matej Hoffmann, Krystian Mikolajczyk

Embodied reasoning systems integrate robotic hardware and cognitive processes to perform complex tasks typically in response to a natural language query about a specific physical environment. This usually involves changing the belief about the scene or physically interacting and changing the scene (e.g. 'Sort the objects from lightest to heaviest'). In order to facilitate the development of such systems we introduce a new simulating environment that makes use of MuJoCo physics engine and high-quality renderer Blender to provide realistic visual observations that are also accurate to the physical state of the scene. Together with the simulator we propose a new benchmark composed of 10 classes of multi-step reasoning scenarios that require simultaneous visual and physical measurements. Finally, we develop a new modular Closed Loop Interactive Reasoning (CLIER) approach that takes into account the measurements of non-visual object properties, changes in the scene caused by external disturbances as well as uncertain outcomes of robotic actions. We extensively evaluate our reasoning approach in simulation and in the real world manipulation tasks with a success rate above 76% and 64%, respectively.

link: http://arxiv.org/abs/2404.15194v1