Sun 2024.04.28

Formal Specification, Assessment, and Enforcement of Fairness for Generative AIs

Chih-Hong Cheng, Changshun Wu, Harald Ruess, Xingyu Zhao, Saddek Bensalem

The risk of reinforcing or exacerbating societal biases and inequalities is growing as generative AI increasingly produces content that resembles human output, from text to images and beyond. Here we formally characterize the notion of fairness for generative AI as a basis for monitoring and enforcing fairness. We define two levels of fairness utilizing the concept of infinite words. The first is the fairness demonstrated on the generated sequences, which is only evaluated on the outputs while agnostic to the prompts/models used. The second is the inherent fairness of the generative AI model, which requires that fairness be manifested when input prompts are neutral, that is, they do not explicitly instruct the generative AI to produce a particular type of output. We also study relative intersectional fairness to counteract the combinatorial explosion of fairness when considering multiple categories together with lazy fairness enforcement. Our implemented specification monitoring and enforcement tool shows interesting results when tested against several generative AI models.

link: http://arxiv.org/abs/2404.16663v1

PhyRecon: Physically Plausible Neural Scene Reconstruction

Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Yixin Zhu, Song-Chun Zhu, Siyuan Huang

While neural implicit representations have gained popularity in multi-view 3D reconstruction, previous work struggles to yield physically plausible results, thereby limiting their applications in physics-demanding domains like embodied AI and robotics. The lack of plausibility originates from both the absence of physics modeling in the existing pipeline and their inability to recover intricate geometrical structures. In this paper, we introduce PhyRecon, which stands as the first approach to harness both differentiable rendering and differentiable physics simulation to learn implicit surface representations. Our framework proposes a novel differentiable particle-based physical simulator seamlessly integrated with the neural implicit representation. At its core is an efficient transformation between SDF-based implicit representation and explicit surface points by our proposed algorithm, Surface Points Marching Cubes (SP-MC), enabling differentiable learning with both rendering and physical losses. Moreover, we model both rendering and physical uncertainty to identify and compensate for the inconsistent and inaccurate monocular geometric priors. The physical uncertainty additionally enables a physics-guided pixel sampling to enhance the learning of slender structures. By amalgamating these techniques, our model facilitates efficient joint modeling with appearance, geometry, and physics. Extensive experiments demonstrate that PhyRecon significantly outperforms all state-of-the-art methods in terms of reconstruction quality. Our reconstruction results also yield superior physical stability, verified by Isaac Gym, with at least a 40% improvement across all datasets, opening broader avenues for future physics-based applications.

link: http://arxiv.org/abs/2404.16666v1

EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, Wen-Huang Cheng

Visual Instruction Tuning represents a novel learning paradigm involving the fine-tuning of pre-trained language models using task-specific instructions. This paradigm shows promising zero-shot results in various natural language processing tasks but is still unexplored in vision emotion understanding. In this work, we focus on enhancing the model's proficiency in understanding and adhering to instructions related to emotional contexts. Initially, we identify key visual clues critical to visual emotion recognition. Subsequently, we introduce a novel GPT-assisted pipeline for generating emotion visual instruction data, effectively addressing the scarcity of

annotated instruction data in this domain. Expanding on the groundwork established by InstructBLIP, our proposed EmoVIT architecture incorporates emotion-specific instruction data, leveraging the powerful capabilities of Large Language Models to enhance performance. Through extensive experiments, our model showcases its proficiency in emotion classification, adeptness in affective reasoning, and competence in comprehending humor. The comparative analysis provides a robust benchmark for Emotion Visual Instruction Tuning in the era of LLMs, providing valuable insights and opening avenues for future exploration in this domain. Our code is available at \url{https://github.com/aimmemotion/EmoVIT}.

link: http://arxiv.org/abs/2404.16670v1

Multilayer Correlation Clustering

Atsushi Miyauchi, Florian Adriaens, Francesco Bonchi, Nikolaj Tatti

In this paper, we establish Multilayer Correlation Clustering, a novel generalization of Correlation Clustering (Bansal et al., FOCS '02) to the multilayer setting. In this model, we are given a series of inputs of Correlation Clustering (called layers) over the common set V. The goal is then to find a clustering of V that minimizes the ℓ_p-norm ($p\geq 1$) of the disagreements vector, which is defined as the vector (with dimension equal to the number of layers), each element of which represents the disagreements of the clustering on the corresponding layer. For this generalization, we first design an $O(L\log n)$-approximation algorithm, where L is the number of layers, based on the well-known region growing technique. We then study an important special case of our problem, namely the problem with the probability constraint. For this case, we first give an $(\alpha+2)$-approximation algorithm, where α is any possible approximation ratio for the single-layer counterpart. For instance, we can take $\alpha=2.5$ in general (Ailon et al., JACM '08) and $\alpha=1.73+\epsilon$ for the unweighted case (Cohen-Addad et al., FOCS '23). Furthermore, we design a 4-approximation algorithm, which improves the above approximation ratio of $\alpha+2=4.5$ for the general probability-constraint case. Computational experiments using real-world datasets demonstrate the effectiveness of our proposed algorithms.

link: http://arxiv.org/abs/2404.16676v1

Multimodal Semantic-Aware Automatic Colorization with Diffusion Prior

Han Wang, Xinning Chai, Yiwen Wang, Yuhong Zhang, Rong Xie, Li Song

Colorizing grayscale images offers an engaging visual experience. Existing automatic colorization methods often fail to generate satisfactory results due to incorrect semantic colors and unsaturated colors. In this work, we propose an automatic colorization pipeline to overcome these challenges. We leverage the extraordinary generative ability of the diffusion prior to synthesize color with plausible semantics. To overcome the artifacts introduced by the diffusion prior, we apply the luminance conditional guidance. Moreover, we adopt multimodal high-level semantic priors to help the model understand the image content and deliver saturated colors. Besides, a luminance-aware decoder is designed to restore details and enhance overall visual quality. The proposed pipeline synthesizes saturated colors while maintaining plausible semantics. Experiments indicate that our proposed method considers both diversity and fidelity, surpassing previous methods in terms of perceptual realism and gain most human preference.

link: http://arxiv.org/abs/2404.16678v1

Multi-scale HSV Color Feature Embedding for High-fidelity NIR-to-RGB Spectrum Translation

Huiyu Zhai, Mo Chen, Xingxing Yang, Gusheng Kang

The NIR-to-RGB spectral domain translation is a formidable task due to the inherent spectral mapping ambiguities within NIR inputs and RGB outputs. Thus, existing methods fail to reconcile the tension between maintaining texture detail fidelity and achieving diverse color variations. In this paper, we propose a Multi-scale HSV Color Feature Embedding Network (MCFNet) that decomposes the mapping process into three sub-tasks, including NIR texture maintenance, coarse geometry reconstruction, and RGB color prediction. Thus, we propose three key modules for each

corresponding sub-task: the Texture Preserving Block (TPB), the HSV Color Feature Embedding Module (HSV-CFEM), and the Geometry Reconstruction Module (GRM). These modules contribute to our MCFNet methodically tackling spectral translation through a series of escalating resolutions, progressively enriching images with color and texture fidelity in a scale-coherent fashion. The proposed MCFNet demonstrates substantial performance gains over the NIR image colorization task. Code is released at: https://github.com/AlexYangxx/MCFNet.

link: http://arxiv.org/abs/2404.16685v1

NTIRE 2024 Quality Assessment of AI-Generated Content Challenge

Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte

This paper reports on the NTIRE 2024 Quality Assessment of AI-Generated Content Challenge, which will be held in conjunction with the New Trends in Image Restoration and Enhancement Workshop (NTIRE) at CVPR 2024. This challenge is to address a major challenge in the field of image and video processing, namely, Image Quality Assessment (IQA) and Video Quality Assessment (VQA) for AI-Generated Content (AIGC). The challenge is divided into the image track and the video track. The image track uses the AIGIQA-20K, which contains 20,000 AI-Generated Images (AIGIs) generated by 15 popular generative models. The image track has a total of 318 registered participants. A total of 1,646 submissions are received in the development phase, and 221 submissions are received in the test phase. Finally, 16 participating teams submitted their models and fact sheets. The video track uses the T2VQA-DB, which contains 10,000 AI-Generated Videos (AIGVs) generated by 9 popular Text-to-Video (T2V) models. A total of 196 participants have registered in the video track. A total of 991 submissions are received in the development phase, and 185 submissions are received in the test phase. Finally, 12 participating teams submitted their models and fact sheets. Some methods have achieved better results than baseline methods, and the winning methods in both tracks have demonstrated superior prediction performance on AIGC.

link: http://arxiv.org/abs/2404.16687v1

Learning to Beat ByteRL: Exploitability of Collectible Card Game Agents

Radovan Haluska, Martin Schmid

While Poker, as a family of games, has been studied extensively in the last decades, collectible card games have seen relatively little attention. Only recently have we seen an agent that can compete with professional human players in Hearthstone, one of the most popular collectible card games. Although artificial agents must be able to work with imperfect information in both of these genres, collectible card games pose another set of distinct challenges. Unlike in many poker variants, agents must deal with state space so vast that even enumerating all states consistent with the agent's beliefs is intractable, rendering the current search methods unusable and requiring the agents to opt for other techniques. In this paper, we investigate the strength of such techniques for this class of games. Namely, we present preliminary analysis results of ByteRL, the state-of-the-art agent in Legends of Code and Magic and Hearthstone. Although ByteRL beat a top-10 Hearthstone player from China, we show that its play in Legends of Code and Magic is highly exploitable.

link: http://arxiv.org/abs/2404.16689v1

Influence of Solution Efficiency and Valence of Instruction on Additive and Subtractive Solution Strategies in Humans and GPT-4

Lydia Uhler, Verena Jordan, Jürgen Buder, Markus Huff, Frank Papenmeier

We explored the addition bias, a cognitive tendency to prefer adding elements over removing them to alter an initial state or structure, by conducting four preregistered experiments examining the problem-solving behavior of both humans and OpenAI's GPT-4 large language model. The experiments involved 588 participants from the U.S. and 680 iterations of the GPT-4 model. The problem-solving task was either to create symmetry within a grid (Experiments 1 and 3) or to edit a summary (Experiments 2 and 4). As hypothesized, we found that overall, the addition bias was

present. Solution efficiency (Experiments 1 and 2) and valence of the instruction (Experiments 3 and 4) played important roles. Human participants were less likely to use additive strategies when subtraction was relatively more efficient than when addition and subtraction were equally efficient. GPT-4 exhibited the opposite behavior, with a strong addition bias when subtraction was more efficient. In terms of instruction valence, GPT-4 was more likely to add words when asked to "improve" compared to "edit", whereas humans did not show this effect. When we looked at the addition bias under different conditions, we found more biased responses for GPT-4 compared to humans. Our findings highlight the importance of considering comparable and sometimes superior subtractive alternatives, as well as reevaluating one's own and particularly the language models' problem-solving behavior.

link: http://arxiv.org/abs/2404.16692v1

Cooperate or Collapse: Emergence of Sustainability Behaviors in a Society of LLM Agents

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, Rada Mihalcea

In the rapidly evolving field of artificial intelligence, ensuring safe decision-making of Large Language Models (LLMs) is a significant challenge. This paper introduces Governance of the Commons Simulation (GovSim), a simulation platform designed to study strategic interactions and cooperative decision-making in LLMs. Through this simulation environment, we explore the dynamics of resource sharing among AI agents, highlighting the importance of ethical considerations, strategic planning, and negotiation skills. GovSim is versatile and supports any text-based agent, including LLMs agents. Using the Generative Agent framework, we create a standard agent that facilitates the integration of different LLMs. Our findings reveal that within GovSim, only two out of 15 tested LLMs managed to achieve a sustainable outcome, indicating a significant gap in the ability of models to manage shared resources. Furthermore, we find that by removing the ability of agents to communicate, they overuse the shared resource, highlighting the importance of communication for cooperation. Interestingly, most LLMs lack the ability to make universalized hypotheses, which highlights a significant weakness in their reasoning skills. We open source the full suite of our research results, including the simulation environment, agent prompts, and a comprehensive web interface.

link: http://arxiv.org/abs/2404.16698v1

Efficient and Near-Optimal Noise Generation for Streaming Differential Privacy

Krishnamurthy, Dvijotham, H. Brendan McMahan, Krishna Pillutla, Thomas Steinke, Abhradeep Thakurta

In the task of differentially private (DP) continual counting, we receive a stream of increments and our goal is to output an approximate running total of these increments, without revealing too much about any specific increment. Despite its simplicity, differentially private continual counting has attracted significant attention both in theory and in practice. Existing algorithms for differentially private continual counting are either inefficient in terms of their space usage or add an excessive amount of noise, inducing suboptimal utility. The most practical DP continual counting algorithms add carefully correlated Gaussian noise to the values. The task of choosing the covariance for this noise can be expressed in terms of factoring the lower-triangular matrix of ones (which computes prefix sums). We present two approaches from this class (for different parameter regimes) that achieve near-optimal utility for DP continual counting and only require logarithmic or polylogarithmic space (and time). Our first approach is based on a space-efficient streaming matrix multiplication algorithm for a class of Toeplitz matrices. We show that to instantiate this algorithm for DP continual counting, it is sufficient to find a low-degree rational function that approximates the square root on a circle in the complex plane. We then apply and extend tools from approximation theory to achieve this. We also derive efficient closed-forms for the objective function for arbitrarily many steps, and show direct numerical optimization yields a highly practical solution to the problem. Our second approach combines our first approach with a recursive construction similar to the binary tree mechanism.

link: http://arxiv.org/abs/2404.16706v1

Multi-view Cardiac Image Segmentation via Trans-Dimensional Priors

Abbas Khan, Muhammad Asad, Martin Benning, Caroline Roney, Gregory Slabaugh

We propose a novel multi-stage trans-dimensional architecture for multi-view cardiac image segmentation. Our method exploits the relationship between long-axis (2D) and short-axis (3D) magnetic resonance (MR) images to perform a sequential 3D-to-2D-to-3D segmentation, segmenting the long-axis and short-axis images. In the first stage, 3D segmentation is performed using the short-axis image, and the prediction is transformed to the long-axis view and used as a segmentation prior in the next stage. In the second step, the heart region is localized and cropped around the segmentation prior using a Heart Localization and Cropping (HLC) module, focusing the subsequent model on the heart region of the image, where a 2D segmentation is performed. Similarly, we transform the long-axis prediction to the short-axis view, localize and crop the heart region and again perform a 3D segmentation to refine the initial short-axis segmentation. We evaluate our proposed method on the Multi-Disease, Multi-View & Multi-Center Right Ventricular Segmentation in Cardiac MRI (M&Ms-2;) dataset, where our method outperforms state-of-the-art methods in segmenting cardiac regions of interest in both short-axis and long-axis images. The pre-trained models, source code, and implementation details will be publicly available.

link: http://arxiv.org/abs/2404.16708v1

Layer Skip: Enabling Early Exit Inference and Self-Speculative Decoding

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed A Aly, Beidi Chen, Carole-Jean Wu

We present LayerSkip, an end-to-end solution to speed-up inference of large language models (LLMs). First, during training we apply layer dropout, with low dropout rates for earlier layers and higher dropout rates for later layers, and an early exit loss where all transformer layers share the same exit. Second, during inference, we show that this training recipe increases the accuracy of early exit at earlier layers, without adding any auxiliary layers or modules to the model. Third, we present a novel self-speculative decoding solution where we exit at early layers and verify and correct with remaining layers of the model. Our proposed self-speculative decoding approach has less memory footprint than other speculative decoding approaches and benefits from shared compute and activations of the draft and verification stages. We run experiments on different Llama model sizes on different types of training: pretraining from scratch, continual pretraining, finetuning on specific data domain, and finetuning on specific task. We implement our inference solution and show speedups of up to 2.16x on summarization for CNN/DM documents, 1.82x on coding, and 2.0x on TOPv2 semantic parsing task.

link: http://arxiv.org/abs/2404.16710v1

Embracing Diversity: Interpretable Zero-shot classification beyond one vector per class

Mazda Moayeri, Michael Rabbat, Mark Ibrahim, Diane Bouchacourt

Vision-language models enable open-world classification of objects without the need for any retraining. While this zero-shot paradigm marks a significant advance, even today's best models exhibit skewed performance when objects are dissimilar from their typical depiction. Real world objects such as pears appear in a variety of forms -- from diced to whole, on a table or in a bowl -- yet standard VLM classifiers map all instances of a class to a \it{single vector based on the class label}. We argue that to represent this rich diversity within a class, zero-shot classification should move beyond a single vector. We propose a method to encode and account for diversity within a class using inferred attributes, still in the zero-shot setting without retraining. We find our method consistently outperforms standard zero-shot classification over a large suite of datasets encompassing hierarchies, diverse object states, and real-world geographic diversity, as well finer-grained datasets where intra-class diversity may be less prevalent. Importantly, our method is

inherently interpretable, offering faithful explanations for each inference to facilitate model debugging and enhance transparency. We also find our method scales efficiently to a large number of attributes to account for diversity -- leading to more accurate predictions for atypical instances. Finally, we characterize a principled trade-off between overall and worst class accuracy, which can be tuned via a hyperparameter of our method. We hope this work spurs further research into the promise of zero-shot classification beyond a single class vector for capturing diversity in the world, and building transparent AI systems without compromising performance.

link: http://dx.doi.org/10.1145/3630106.3659039

Features Fusion for Dual-View Mammography Mass Detection

Arina Varlamova, Valery Belotsky, Grigory Novikov, Anton Konushin, Evgeny Sidorov

Detection of malignant lesions on mammography images is extremely important for early breast cancer diagnosis. In clinical practice, images are acquired from two different angles, and radiologists can fully utilize information from both views, simultaneously locating the same lesion. However, for automatic detection approaches such information fusion remains a challenge. In this paper, we propose a new model called MAMM-Net, which allows the processing of both mammography views simultaneously by sharing information not only on an object level, as seen in existing works, but also on a feature level. MAMM-Net's key component is the Fusion Layer, based on deformable attention and designed to increase detection precision while keeping high recall. Our experiments show superior performance on the public DDSM dataset compared to the previous state-of-the-art model, while introducing new helpful features such as lesion annotation on pixel-level and classification of lesions malignancy.

link: http://arxiv.org/abs/2404.16718v1