Sat 2024.04.20

Unifying Bias and Unfairness in Information Retrieval: A Survey of Challenges and Opportunities with Large Language Models

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, Jun Xu

With the rapid advancement of large language models (LLMs), information retrieval (IR) systems, such as search engines and recommender systems, have undergone a significant paradigm shift. This evolution, while heralding new opportunities, introduces emerging challenges, particularly in terms of biases and unfairness, which may threaten the information ecosystem. In this paper, we present a comprehensive survey of existing works on emerging and pressing bias and unfairness issues in IR systems when the integration of LLMs. We first unify bias and unfairness issues as distribution mismatch problems, providing a groundwork for categorizing various mitigation strategies through distribution alignment. Subsequently, we systematically delve into the specific bias and unfairness issues arising from three critical stages of LLMs integration into IR systems: data collection, model development, and result evaluation. In doing so, we meticulously review and analyze recent literature, focusing on the definitions, characteristics, and corresponding mitigation strategies associated with these issues. Finally, we identify and highlight some open problems and challenges for future work, aiming to inspire researchers and stakeholders in the IR field and beyond to better understand and mitigate bias and unfairness issues of IR in this LLM era. We also consistently maintain a GitHub repository for the relevant papers and resources in this rising direction at https://github.com/KID-22/LLM-IR-Bias-Fairness-Survey.

link: http://arxiv.org/abs/2404.11457v1

Learn to Tour: Operator Design For Solution Feasibility Mapping in Pickup-and-delivery Traveling Salesman Problem

Bowen Fang, Xu Chen, Xuan Di

This paper aims to develop a learning method for a special class of traveling salesman problems (TSP), namely, the pickup-and-delivery TSP (PDTSP), which finds the shortest tour along a sequence of one-to-one pickup-and-delivery nodes. One-to-one here means that the transported people or goods are associated with designated pairs of pickup and delivery nodes, in contrast to that indistinguishable goods can be delivered to any nodes. In PDTSP, precedence constraints need to be satisfied that each pickup node must be visited before its corresponding delivery node. Classic operations research (OR) algorithms for PDTSP are difficult to scale to large-sized problems. Recently, reinforcement learning (RL) has been applied to TSPs. The basic idea is to explore and evaluate visiting sequences in a solution space. However, this approach could be less computationally efficient, as it has to potentially evaluate many infeasible solutions of which precedence constraints are violated. To restrict solution search within a feasible space, we utilize operators that always map one feasible solution to another, without spending time exploring the infeasible solution space. Such operators are evaluated and selected as policies to solve PDTSPs in an RL framework. We make a comparison of our method and baselines, including classic OR algorithms and existing learning methods. Results show that our approach can find tours shorter than baselines.

link: http://arxiv.org/abs/2404.11458v1

Octopus v3: Technical Report for On-device Sub-billion Multimodal AI Agent

Wei Chen, Zhiyuan Li

A multimodal AI agent is characterized by its ability to process and learn from various types of data, including natural language, visual, and audio inputs, to inform its actions. Despite advancements in large language models that incorporate visual data, such as GPT-4V, effectively translating image-based data into actionable outcomes for AI agents continues to be challenging. In this paper, we introduce a multimodal model that incorporates the concept of functional token specifically designed for AI agent applications. To ensure compatibility with edge devices, our model is

optimized to a compact size of less than 1B parameters. Like GPT-4, our model can process both English and Chinese. We demonstrate that this model is capable of operating efficiently on a wide range of edge devices, including as constrained as a Raspberry Pi.

link: http://arxiv.org/abs/2404.11459v2

Using Game Engines and Machine Learning to Create Synthetic Satellite Imagery for a Tabletop Verification Exercise

Johannes Hoster, Sara Al-Sayed, Felix Biessmann, Alexander Glaser, Kristian Hildebrand, Igor Moric, Tuong Vy Nguyen

Satellite imagery is regarded as a great opportunity for citizen-based monitoring of activities of interest. Relevant imagery may however not be available at sufficiently high resolution, quality, or cadence -- let alone be uniformly accessible to open-source analysts. This limits an assessment of the true long-term potential of citizen-based monitoring of nuclear activities using publicly available satellite imagery. In this article, we demonstrate how modern game engines combined with advanced machine-learning techniques can be used to generate synthetic imagery of sites of interest with the ability to choose relevant parameters upon request; these include time of day, cloud cover, season, or level of activity onsite. At the same time, resolution and off-nadir angle can be adjusted to simulate different characteristics of the satellite. While there are several possible use-cases for synthetic imagery, here we focus on its usefulness to support tabletop exercises in which simple monitoring scenarios can be examined to better understand verification capabilities enabled by new satellite constellations and very short revisit times.

link: http://arxiv.org/abs/2404.11461v1

X-posing Free Speech: Examining the Impact of Moderation Relaxation on Online Social Networks

Arvindh Arun, Saurav Chhatani, Jisun An, Ponnurangam Kumaraguru

We investigate the impact of free speech and the relaxation of moderation on online social media platforms using Elon Musk's takeover of Twitter as a case study. By curating a dataset of over 10 million tweets, our study employs a novel framework combining content and network analysis. Our findings reveal a significant increase in the distribution of certain forms of hate content, particularly targeting the LGBTQ+ community and liberals. Network analysis reveals the formation of cohesive hate communities facilitated by influential bridge users, with substantial growth in interactions hinting at increased hate production and diffusion. By tracking the temporal evolution of PageRank, we identify key influencers, primarily self-identified far-right supporters disseminating hate against liberals and woke culture. Ironically, embracing free speech principles appears to have enabled hate speech against the very concept of freedom of expression and free speech itself. Our findings underscore the delicate balance platforms must strike between open expression and robust moderation to curb the proliferation of hate online.

link: http://arxiv.org/abs/2404.11465v1

A Federated Learning Approach to Privacy Preserving Offensive Language Identification

Marcos Zampieri, Damith Premasiri, Tharindu Ranasinghe

The spread of various forms of offensive speech online is an important concern in social media. While platforms have been investing heavily in ways of coping with this problem, the question of privacy remains largely unaddressed. Models trained to detect offensive language on social media are trained and/or fine-tuned using large amounts of data often stored in centralized servers. Since most social media data originates from end users, we propose a privacy preserving decentralized architecture for identifying offensive language online by introducing Federated Learning (FL) in the context of offensive language identification. FL is a decentralized architecture that allows multiple models to be trained locally without the need for data sharing hence preserving users' privacy. We propose a model fusion approach to perform FL. We trained multiple deep learning models on four publicly available English benchmark datasets (AHSD, HASOC, HateXplain, OLID) and evaluated

their performance in detail. We also present initial cross-lingual experiments in English and Spanish. We show that the proposed model fusion approach outperforms baselines in all the datasets while preserving privacy.

link: http://arxiv.org/abs/2404.11470v1

Towards Highly Realistic Artistic Style Transfer via Stable Diffusion with Step-aware and Layer-aware Prompt

Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, Wei Xing, Juncheng Mo, Shuaicheng Huang, Jinheng Xie, Guangyuan Li, Junsheng Luan, Lei Zhao, Dalong Zhang, Lixia Chen

Artistic style transfer aims to transfer the learned artistic style onto an arbitrary content image, generating artistic stylized images. Existing generative adversarial network-based methods fail to generate highly realistic stylized images and always introduce obvious artifacts and disharmonious patterns. Recently, large-scale pre-trained diffusion models opened up a new way for generating highly realistic artistic stylized images. However, diffusion model-based methods generally fail to preserve the content structure of input content images well, introducing some undesired content structure and style patterns. To address the above problems, we propose a novel pre-trained diffusion-based artistic style transfer method, called LSAST, which can generate highly realistic artistic stylized images while preserving the content structure of input content images well, without bringing obvious artifacts and disharmonious style patterns. Specifically, we introduce a Step-aware and Layer-aware Prompt Space, a set of learnable prompts, which can learn the style information from the collection of artworks and dynamically adjusts the input images' content structure and style pattern. To train our prompt space, we propose a novel inversion method, called Step-ware and Layer-aware Prompt Inversion, which allows the prompt space to learn the style information of the artworks collection. In addition, we inject a pre-trained conditional branch of ControlNet into our LSAST, which further improved our framework's ability to maintain content structure. Extensive experiments demonstrate that our proposed method can generate more highly realistic artistic stylized images than the state-of-the-art artistic style transfer methods.

link: http://arxiv.org/abs/2404.11474v1

AdaIR: Exploiting Underlying Similarities of Image Restoration Tasks with Adapters

Hao-Wei Chen, Yu-Syuan Xu, Kelvin C. K. Chan, Hsien-Kai Kuo, Chun-Yi Lee, Ming-Hsuan Yang

Existing image restoration approaches typically employ extensive networks specifically trained for designated degradations. Despite being effective, such methods inevitably entail considerable storage costs and computational overheads due to the reliance on task-specific networks. In this work, we go beyond this well-established framework and exploit the inherent commonalities among image restoration tasks. The primary objective is to identify components that are shareable across restoration tasks and augment the shared components with modules specifically trained for individual tasks. Towards this goal, we propose AdaIR, a novel framework that enables low storage cost and efficient training without sacrificing performance. Specifically, a generic restoration network is first constructed through self-supervised pre-training using synthetic degradations. Subsequent to the pre-training phase, adapters are trained to adapt the pre-trained network to specific degradations. AdaIR requires solely the training of lightweight, task-specific modules, ensuring a more efficient storage and training regimen. We have conducted extensive experiments to validate the effectiveness of AdaIR and analyze the influence of the pre-training strategy on discovering shareable components. Extensive experimental results show that AdaIR achieves outstanding results on multi-task restoration while utilizing significantly fewer parameters (1.9 MB) and less training time (7 hours) for each restoration task. The source codes and trained models will be released.

link: http://arxiv.org/abs/2404.11475v1

Taxonomy to Regulation: A (Geo)Political Taxonomy for AI Risks and Regulatory Measures in the EU AI Act

Sinan Arda

Technological innovations have shown remarkable capabilities to benefit and harm society alike. AI constitutes a democratized sophisticated technology accessible to large parts of society, including malicious actors. This work proposes a taxonomy focusing on on (geo)political risks associated with AI. It identifies 12 risks in total divided into four categories: (1) Geopolitical Pressures, (2) Malicious Usage, (3) Environmental, Social, and Ethical Risks, and (4) Privacy and Trust Violations. Incorporating a regulatory side, this paper conducts a policy assessment of the EU AI Act. Adopted in March 2023, the landmark regulation has the potential to have a positive top-down impact concerning AI risk reduction but needs regulatory adjustments to mitigate risks more comprehensively. Regulatory exceptions for open-source models, excessively high parameters for the classification of GPAI models as a systemic risk, and the exclusion of systems designed exclusively for military purposes from the regulation's obligations leave room for future action.

link: http://arxiv.org/abs/2404.11476v1

IoTSim-Osmosis-RES: Towards autonomic renewable energy-aware osmotic computing

Tomasz Szydlo, Amadeusz Szabala, Nazar Kordiumov, Konrad Siuzdak, Lukasz Wolski, Khaled Alwasel, Fawzy Habeeb, Rajiv Ranjan

Internet of Things systems exists in various areas of our everyday life. For example, sensors installed in smart cities and homes are processed in edge and cloud computing centres providing several benefits that improve our lives. The place of data processing is related to the required system response times -- processing data closer to its source results in a shorter system response time. The Osmotic Computing concept enables flexible deployment of data processing services and their possible movement, just like particles in the osmosis phenomenon move between regions of different densities. At the same time, the impact of complex computer architecture on the environment is increasingly being compensated by the use of renewable and low-carbon energy sources. However, the uncertainty of supplying green energy makes the management of Osmotic Computing demanding, and therefore their autonomy is desirable. In the paper, we present a framework enabling osmotic computing simulation based on renewable energy sources and autonomic osmotic agents, allowing the analysis of distributed management algorithms. We discuss the challenges posed to the framework and analyze various management algorithms for cooperating osmotic agents. In the evaluation we show that changing the adaptation logic of the osmotic agents, it is possible to increase the self-consumption of renewable energy sources or increase the usage of low emission ones.

link: http://dx.doi.org/10.1002/spe.3084

AgentKit: Flow Engineering with Graphs, not Coding

Yue Wu, Yewen Fan, So Yeon Min, Shrimai Prabhumoye, Stephen McAleer, Yonatan Bisk, Ruslan Salakhutdinov, Yuanzhi Li, Tom Mitchell

We propose an intuitive LLM prompting framework (AgentKit) for multifunctional agents. AgentKit offers a unified framework for explicitly constructing a complex "thought process" from simple natural language prompts. The basic building block in AgentKit is a node, containing a natural language prompt for a specific subtask. The user then puts together chains of nodes, like stacking LEGO pieces. The chains of nodes can be designed to explicitly enforce a naturally structured "thought process". For example, for the task of writing a paper, one may start with the thought process of 1) identify a core message, 2) identify prior research gaps, etc. The nodes in AgentKit can be designed and combined in different ways to implement multiple advanced capabilities including on-the-fly hierarchical planning, reflection, and learning from interactions. In addition, due to the modular nature and the intuitive design to simulate explicit human thought process, a basic agent could be implemented as simple as a list of prompts for the subtasks and therefore could be designed and tuned by someone without any programming experience. Quantitatively, we show that agents designed through AgentKit achieve SOTA performance on WebShop and Crafter. These advances underscore AgentKit's potential in making LLM agents effective and accessible for a wider range of applications. https://github.com/holmeswww/AgentKit

link: http://arxiv.org/abs/2404.11483v1

Multi-resolution Rescored ByteTrack for Video Object Detection on Ultra-low-power Embedded Systems

Luca Bompani, Manuele Rusci, Daniele Palossi, Francesco Conti, Luca Benini

This paper introduces Multi-Resolution Rescored Byte-Track (MR2-ByteTrack), a novel video object detection framework for ultra-low-power embedded processors. This method reduces the average compute load of an off-the-shelf Deep Neural Network (DNN) based object detector by up to 2.25$\times$ by alternating the processing of high-resolution images (320$\times$320 pixels) with multiple down-sized frames (192$\times$192 pixels). To tackle the accuracy degradation due to the reduced image input size, MR2-ByteTrack correlates the output detections over time using the ByteTrack tracker and corrects potential misclassification using a novel probabilistic Rescore algorithm. By interleaving two down-sized images for every high-resolution one as the input of different state-of-the-art DNN object detectors with our MR2-ByteTrack, we demonstrate an average accuracy increase of 2.16% and a latency reduction of 43% on the GAP9 microcontroller compared to a baseline frame-by-frame inference scheme using exclusively full-resolution images. Code available at: https://github.com/Bomps4/Multi_Resolution_Rescored_ByteTrack

link: http://arxiv.org/abs/2404.11488v1

arcjetCV: an open-source software to analyze material ablation

Alexandre Quintart, Magnus Haw, Federico Semeraro

arcjetCV is an open-source Python software designed to automate time-resolved measurements of heatshield material recession and recession rates from arcjet test video footage. This new automated and accessible capability greatly exceeds previous manual extraction methods, enabling rapid and detailed characterization of material recession for any sample with a profile video. arcjetCV automates the video segmentation process using machine learning models, including a one-dimensional (1D) Convolutional Neural Network (CNN) to infer the time-window of interest, a two-dimensional (2D) CNN for image and edge segmentation, and a Local Outlier Factor (LOF) for outlier filtering. A graphical user interface (GUI) simplifies the user experience and an application programming interface (API) allows users to call the core functions from scripts, enabling video batch processing. arcjetCV's capability to measure time-resolved recession in turn enables characterization of non-linear processes (shrinkage, swelling, melt flows, etc.), contributing to higher fidelity validation and improved modeling of heatshield material performance. The source code associated with this article can be found at https://github.com/magnus-haw/arcjetCV.

link: http://arxiv.org/abs/2404.11492v1

Runtime Analysis of Evolutionary Diversity Optimization on the Multi-objective (LeadingOnes, TrailingZeros) Problem

Denis Antipov, Aneta Neumann, Frank Neumann. Andrew M. Sutton

The diversity optimization is the class of optimization problems, in which we aim at finding a diverse set of good solutions. One of the frequently used approaches to solve such problems is to use evolutionary algorithms which evolve a desired diverse population. This approach is called evolutionary diversity optimization (EDO). In this paper, we analyse EDO on a 3-objective function LOTZ$_k$, which is a modification of the 2-objective benchmark function (LeadingOnes, TrailingZeros). We prove that the GSEMO computes a set of all Pareto-optimal solutions in $O(kn^3)$ expected iterations. We also analyze the runtime of the GSEMO$_D$ (a modification of the GSEMO for diversity optimization) until it finds a population with the best possible diversity for two different diversity measures, the total imbalance and the sorted imbalances vector. For the first measure we show that the GSEMO$_D$ optimizes it asymptotically faster than it finds a Pareto-optimal population, in $O(kn^2\log(n))$ expected iterations, and for the second measure we show an upper bound of $O(k^2n^3\log(n))$ expected iterations. We complement our theoretical analysis with an empirical study, which shows a very similar behavior for both diversity measures that is close to the theory predictions.

link: http://arxiv.org/abs/2404.11496v1

A Data-Driven Representation for Sign Language Production

Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, Richard Bowden

Phonetic representations are used when recording spoken languages, but no equivalent exists for recording signed languages. As a result, linguists have proposed several annotation systems that operate on the gloss or sub-unit level; however, these resources are notably irregular and scarce. Sign Language Production (SLP) aims to automatically translate spoken language sentences into continuous sequences of sign language. However, current state-of-the-art approaches rely on scarce linguistic resources to work. This has limited progress in the field. This paper introduces an innovative solution by transforming the continuous pose generation problem into a discrete sequence generation problem. Thus, overcoming the need for costly annotation. Although, if available, we leverage the additional information to enhance our approach. By applying Vector Quantisation (VQ) to sign language data, we first learn a codebook of short motions that can be combined to create a natural sequence of sign. Where each token in the codebook can be thought of as the lexicon of our representation. Then using a transformer we perform a translation from spoken language text to a sequence of codebook tokens. Each token can be directly mapped to a sequence of poses allowing the translation to be performed by a single network. Furthermore, we present a sign stitching method to effectively join tokens together. We evaluate on the RWTH-PHOENIX-Weather-2014T (PHOENIX14T) and the more challenging Meine DGS Annotated (mDGS) datasets. An extensive evaluation shows our approach outperforms previous methods, increasing the BLEU-1 back translation score by up to 72%.

link: http://arxiv.org/abs/2404.11499v1