Mon 2024.04.01

Graph Neural Aggregation-diffusion with Metastability

Kaiyuan Cui, Xinyan Wang, Zicheng Zhang, Weichen Zhao

Continuous graph neural models based on differential equations have expanded the architecture of graph neural networks (GNNs). Due to the connection between graph diffusion and message passing, diffusion-based models have been widely studied. However, diffusion naturally drives the system towards an equilibrium state, leading to issues like over-smoothing. To this end, we propose GRADE inspired by graph aggregation-diffusion equations, which includes the delicate balance between nonlinear diffusion and aggregation induced by interaction potentials. The node representations obtained through aggregation-diffusion equations exhibit metastability, indicating that features can aggregate into multiple clusters. In addition, the dynamics within these clusters can persist for long time periods, offering the potential to alleviate over-smoothing effects. This nonlinear diffusion in our model generalizes existing diffusion-based models and establishes a connection with classical GNNs. We prove that GRADE achieves competitive performance across various benchmarks and alleviates the over-smoothing issue in GNNs evidenced by the enhanced Dirichlet energy.

link: http://arxiv.org/abs/2403.20221v1

Shallow Cross-Encoders for Low-Latency Retrieval

Aleksandr V. Petrov, Sean MacAvaney, Craig Macdonald

Transformer-based Cross-Encoders achieve state-of-the-art effectiveness in text retrieval. However, Cross-Encoders based on large transformer models (such as BERT or T5) are computationally expensive and allow for scoring only a small number of documents within a reasonably small latency window. However, keeping search latencies low is important for user satisfaction and energy usage. In this paper, we show that weaker shallow transformer models (i.e., transformers with a limited number of layers) actually perform better than full-scale models when constrained to these practical low-latency settings since they can estimate the relevance of more documents in the same time budget. We further show that shallow transformers may benefit from the generalized Binary Cross-Entropy (gBCE) training scheme, which has recently demonstrated success for recommendation tasks. Our experiments with TREC Deep Learning passage ranking query sets demonstrate significant improvements in shallow and full-scale models in low-latency scenarios. For example, when the latency limit is 25ms per query, MonoBERT-Large (a cross-encoder based on a full-scale BERT model) is only able to achieve NDCG@10 of 0.431 on TREC DL 2019, while TinyBERT-gBCE (a cross-encoder based on TinyBERT trained with gBCE) reaches NDCG@10 of 0.652, a +51% gain over MonoBERT-Large. We also show that shallow Cross-Encoders are effective even when used without a GPU (e.g., with CPU inference, NDCG@10 decreases only by 3% compared to GPU inference with 50ms latency), which makes Cross-Encoders practical to run even without specialized hardware acceleration.

link: http://dx.doi.org/10.1007/978-3-031-56063-7_10

MTMMC: A Large-Scale Real-World Multi-Modal Camera Tracking Benchmark

Sanghyun Woo, Kwanyong Park, Inkyu Shin, Myungchul Kim, In So Kweon

Multi-target multi-camera tracking is a crucial task that involves identifying and tracking individuals over time using video streams from multiple cameras. This task has practical applications in various fields, such as visual surveillance, crowd behavior analysis, and anomaly detection. However, due to the difficulty and cost of collecting and labeling data, existing datasets for this task are either synthetically generated or artificially constructed within a controlled camera network setting, which limits their ability to model real-world dynamics and generalize to diverse camera configurations. To address this issue, we present MTMMC, a real-world, large-scale dataset that includes long video sequences captured by 16 multi-modal cameras in two different environments - campus and factory - across various time, weather, and season conditions. This dataset provides a challenging test-bed

for studying multi-camera tracking under diverse real-world complexities and includes an additional input modality of spatially aligned and temporally synchronized RGB and thermal cameras, which enhances the accuracy of multi-camera tracking. MTMMC is a super-set of existing datasets, benefiting independent fields such as person detection, re-identification, and multiple object tracking. We provide baselines and new learning setups on this dataset and set the reference scores for future studies. The datasets, models, and test server will be made publicly available.

link: http://arxiv.org/abs/2403.20225v1

An FPGA-Based Reconfigurable Accelerator for Convolution-Transformer Hybrid EfficientViT

Haikuo Shao, Huihong Shi, Wendong Mao, Zhongfeng Wang

Vision Transformers (ViTs) have achieved significant success in computer vision. However, their intensive computations and massive memory footprint challenge ViTs' deployment on embedded devices, calling for efficient ViTs. Among them, EfficientViT, the state-of-the-art one, features a Convolution-Transformer hybrid architecture, enhancing both accuracy and hardware efficiency. Unfortunately, existing accelerators cannot fully exploit the hardware benefits of EfficientViT due to its unique architecture. In this paper, we propose an FPGA-based accelerator for EfficientViT to advance the hardware efficiency frontier of ViTs. Specifically, we design a reconfigurable architecture to efficiently support various operation types, including lightweight convolutions and attention, boosting hardware utilization. Additionally, we present a time-multiplexed and pipelined dataflow to facilitate both intra- and inter-layer fusions, reducing off-chip data access costs. Experimental results show that our accelerator achieves up to 780.2 GOPS in throughput and 105.1 GOPS/W in energy efficiency at 200MHz on the Xilinx ZCU102 FPGA, which significantly outperforms prior works.

link: http://arxiv.org/abs/2403.20230v1

U-VAP: User-specified Visual Appearance Personalization via Decoupled Self Augmentation

You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, Jintao Li

Concept personalization methods enable large text-to-image models to learn specific subjects (e.g., objects/poses/3D models) and synthesize renditions in new contexts. Given that the image references are highly biased towards visual attributes, state-of-the-art personalization models tend to overfit the whole subject and cannot disentangle visual characteristics in pixel space. In this study, we proposed a more challenging setting, namely fine-grained visual appearance personalization. Different from existing methods, we allow users to provide a sentence describing the desired attributes. A novel decoupled self-augmentation strategy is proposed to generate target-related and non-target samples to learn user-specified visual attributes. These augmented data allow for refining the model's understanding of the target attribute while mitigating the impact of unrelated attributes. At the inference stage, adjustments are conducted on semantic space through the learned target and non-target embeddings to further enhance the disentanglement of target attributes. Extensive experiments on various kinds of visual attributes with SOTA personalization methods show the ability of the proposed method to mimic target visual appearance in novel contexts, thus improving the controllability and flexibility of personalization.

link: http://arxiv.org/abs/2403.20231v1

Functional Bilevel Optimization for Machine Learning

Ieva Petrulionyte, Julien Mairal, Michael Arbel

In this paper, we introduce a new functional point of view on bilevel optimization problems for machine learning, where the inner objective is minimized over a function space. These types of problems are most often solved by using methods developed in the parametric setting, where the inner objective is strongly convex with respect to the parameters of the prediction function. The functional point of view does not rely on this assumption and notably allows using over-parameterized neural networks as the inner prediction function. We propose scalable and

efficient algorithms for the functional bilevel optimization problem and illustrate the benefits of our approach on instrumental regression and reinforcement learning tasks, which admit natural functional bilevel structures.

link: http://arxiv.org/abs/2403.20233v1

Artificial Neural Networks-based Real-time Classification of ENG Signals for Implanted Nerve Interfaces

ntonio Coviello, Francesco Linsalata, Umberto Spagnolini, Maurizio Magarini

Neuropathies are gaining higher relevance in clinical settings, as they risk permanently jeopardizing a person's life. To support the recovery of patients, the use of fully implanted devices is emerging as one of the most promising solutions. However, these devices, even if becoming an integral part of a fully complex neural nanonetwork system, pose numerous challenges. In this article, we address one of them, which consists of the classification of motor/sensory stimuli. The task is performed by exploring four different types of artificial neural networks (ANNs) to extract various sensory stimuli from the electroneurographic (ENG) signal measured in the sciatic nerve of rats. Different sizes of the data sets are considered to analyze the feasibility of the investigated ANNs for real-time classification through a comparison of their performance in terms of accuracy, F1-score, and prediction time. The design of the ANNs takes advantage of the modelling of the ENG signal as a multiple-input multiple-output (MIMO) system to describe the measures taken by state-of-the-art implanted nerve interfaces. These are based on the use of multi-contact cuff electrodes to achieve nanoscale spatial discrimination of the nerve activity. The MIMO ENG signal model is another contribution of this paper. Our results show that some ANNs are more suitable for real-time applications, being capable of achieving accuracies over 90% for signal windows of 100 and $200\,$ms with a low enough processing time to be effective for pathology recovery.

link: http://arxiv.org/abs/2403.20234v1

Long-Tailed Anomaly Detection with Learnable Class Names

Chih-Hui Ho, Kuan-Chuan Peng, Nuno Vasconcelos

Anomaly detection (AD) aims to identify defective images and localize their defects (if any). Ideally, AD models should be able to detect defects over many image classes; without relying on hard-coded class names that can be uninformative or inconsistent across datasets; learn without anomaly supervision; and be robust to the long-tailed distributions of real-world applications. To address these challenges, we formulate the problem of long-tailed AD by introducing several datasets with different levels of class imbalance and metrics for performance evaluation. We then propose a novel method, LTAD, to detect defects from multiple and long-tailed classes, without relying on dataset class names. LTAD combines AD by reconstruction and semantic AD modules. AD by reconstruction is implemented with a transformer-based reconstruction module. Semantic AD is implemented with a binary classifier, which relies on learned pseudo class names and a pretrained foundation model. These modules are learned over two phases. Phase 1 learns the pseudo-class names and a variational autoencoder (VAE) for feature synthesis that augments the training data to combat long-tails. Phase 2 then learns the parameters of the reconstruction and classification modules of LTAD. Extensive experiments using the proposed long-tailed datasets show that LTAD substantially outperforms the state-of-the-art methods for most forms of dataset imbalance. The long-tailed dataset split is available at https://zenodo.org/records/10854201 .

link: http://arxiv.org/abs/2403.20236v1

Enhancing Dimension-Reduced Scatter Plots with Class and Feature Centroids

Daniel B. Hier, Tayo Obafemi-Ajayi, Gayla R. Olbricht, Devin M. Burns, Sasha Petrenko, Donald C. Wunsch II

Dimension reduction is increasingly applied to high-dimensional biomedical data to improve its interpretability. When datasets are reduced to two dimensions, each observation is assigned an x and y coordinates and is represented as a point on a scatter plot. A significant challenge lies in interpreting the meaning of the x and y axes due to the complexities inherent in dimension

reduction. This study addresses this challenge by using the x and y coordinates derived from dimension reduction to calculate class and feature centroids, which can be overlaid onto the scatter plots. This method connects the low-dimension space to the original high-dimensional space. We illustrate the utility of this approach with data derived from the phenotypes of three neurogenetic diseases and demonstrate how the addition of class and feature centroids increases the interpretability of scatter plots.

link: http://arxiv.org/abs/2403.20246v1

Relation Rectification in Diffusion Model

Yinwei Wu, Xingyi Yang, Xinchao Wang

Despite their exceptional generative abilities, large text-to-image diffusion models, much like skilled but careless artists, often struggle with accurately depicting visual relationships between objects. This issue, as we uncover through careful analysis, arises from a misaligned text encoder that struggles to interpret specific relationships and differentiate the logical order of associated objects. To resolve this, we introduce a novel task termed Relation Rectification, aiming to refine the model to accurately represent a given relationship it initially fails to generate. To address this, we propose an innovative solution utilizing a Heterogeneous Graph Convolutional Network (HGCN). It models the directional relationships between relation terms and corresponding objects within the input prompts. Specifically, we optimize the HGCN on a pair of prompts with identical relational words but reversed object orders, supplemented by a few reference images. The lightweight HGCN adjusts the text embeddings generated by the text encoder, ensuring the accurate reflection of the textual relation in the embedding space. Crucially, our method retains the parameters of the text encoder and diffusion model, preserving the model's robust performance on unrelated descriptions. We validated our approach on a newly curated dataset of diverse relational data, demonstrating both quantitative and qualitative enhancements in generating images with precise visual relations. Project page: https://wuyinwei-hah.github.io/rrnet.github.io/.

link: http://arxiv.org/abs/2403.20249v1

Optimal Policy Learning with Observational Data in Multi-Action Scenarios: Estimation, Risk Preference, and Potential Failures

Giovanni Cerulli

This paper deals with optimal policy learning (OPL) with observational data, i.e. data-driven optimal decision-making, in multi-action (or multi-arm) settings, where a finite set of decision options is available. It is organized in three parts, where I discuss respectively: estimation, risk preference, and potential failures. The first part provides a brief review of the key approaches to estimating the reward (or value) function and optimal policy within this context of analysis. Here, I delineate the identification assumptions and statistical properties related to offline optimal policy learning estimators. In the second part, I delve into the analysis of decision risk. This analysis reveals that the optimal choice can be influenced by the decision maker's attitude towards risks, specifically in terms of the trade-off between reward conditional mean and conditional variance. Here, I present an application of the proposed model to real data, illustrating that the average regret of a policy with multi-valued treatment is contingent on the decision-maker's attitude towards risk. The third part of the paper discusses the limitations of optimal data-driven decision-making by highlighting conditions under which decision-making can falter. This aspect is linked to the failure of the two fundamental assumptions essential for identifying the optimal choice: (i) overlapping, and (ii) unconfoundedness. Some conclusions end the paper.

link: http://arxiv.org/abs/2403.20250v1

Latent Embedding Clustering for Occlusion Robust Head Pose Estimation

José Celestino, Manuel Marques, Jacinto C. Nascimento

Head pose estimation has become a crucial area of research in computer vision given its usefulness in a wide range of applications, including robotics, surveillance, or driver attention monitoring. One of the most difficult challenges in this field is managing head occlusions that

frequently take place in real-world scenarios. In this paper, we propose a novel and efficient framework that is robust in real world head occlusion scenarios. In particular, we propose an unsupervised latent embedding clustering with regression and classification components for each pose angle. The model optimizes latent feature representations for occluded and non-occluded images through a clustering term while improving fine-grained angle predictions. Experimental evaluation on in-the-wild head pose benchmark datasets reveal competitive performance in comparison to state-of-the-art methodologies with the advantage of having a significant data reduction. We observe a substantial improvement in occluded head pose estimation. Also, an ablation study is conducted to ascertain the impact of the clustering term within our proposed framework.

link: http://arxiv.org/abs/2403.20251v1

Using LLMs to Model the Beliefs and Preferences of Targeted Populations

Keiichi Namikoshi, Alex Filipowicz, David A. Shamma, Rumen Iliev, Candice L. Hogan, Nikos Arechiga

We consider the problem of aligning a large language model (LLM) to model the preferences of a human population. Modeling the beliefs, preferences, and behaviors of a specific population can be useful for a variety of different applications, such as conducting simulated focus groups for new products, conducting virtual surveys, and testing behavioral interventions, especially for interventions that are expensive, impractical, or unethical. Existing work has had mixed success using LLMs to accurately model human behavior in different contexts. We benchmark and evaluate two well-known fine-tuning approaches and evaluate the resulting populations on their ability to match the preferences of real human respondents on a survey of preferences for battery electric vehicles (BEVs). We evaluate our models against their ability to match population-wide statistics as well as their ability to match individual responses, and we investigate the role of temperature in controlling the trade-offs between these two. Additionally, we propose and evaluate a novel loss term to improve model performance on responses that require a numeric response.

link: http://arxiv.org/abs/2403.20252v1

MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, Yiming Xiao

Medical image segmentation of anatomical structures and pathology is crucial in modern clinical diagnosis, disease study, and treatment planning. To date, great progress has been made in deep learning-based segmentation techniques, but most methods still lack data efficiency, generalizability, and interactability. Consequently, the development of new, precise segmentation methods that demand fewer labeled datasets is of utmost importance in medical image analysis. Recently, the emergence of foundation models, such as CLIP and Segment-Anything-Model (SAM), with comprehensive cross-domain representation opened the door for interactive and universal image segmentation. However, exploration of these models for data-efficient medical image segmentation is still limited, but is highly necessary. In this paper, we propose a novel framework, called MedCLIP-SAM that combines CLIP and SAM models to generate segmentation of clinical scans using text prompts in both zero-shot and weakly supervised settings. To achieve this, we employed a new Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss to fine-tune the BiomedCLIP model and the recent gScoreCAM to generate prompts to obtain segmentation masks from SAM in a zero-shot setting. Additionally, we explored the use of zero-shot segmentation labels in a weakly supervised paradigm to improve the segmentation quality further. By extensively testing three diverse segmentation tasks and medical image modalities (breast tumor ultrasound, brain tumor MRI, and lung X-ray), our proposed framework has demonstrated excellent accuracy.

link: http://arxiv.org/abs/2403.20253v1

Benchmarking the Robustness of Temporal Action Detection Models Against Temporal Corruptions

Runhao Zeng, Xiaoyong Chen, Jiaming Liang, Huisi Wu, Guangzhong Cao, Yong Guo

Temporal action detection (TAD) aims to locate action positions and recognize action categories in long-term untrimmed videos. Although many methods have achieved promising results, their robustness has not been thoroughly studied. In practice, we observe that temporal information in videos can be occasionally corrupted, such as missing or blurred frames. Interestingly, existing methods often incur a significant performance drop even if only one frame is affected. To formally evaluate the robustness, we establish two temporal corruption robustness benchmarks, namely THUMOS14-C and ActivityNet-v1.3-C. In this paper, we extensively analyze the robustness of seven leading TAD methods and obtain some interesting findings: 1) Existing methods are particularly vulnerable to temporal corruptions, and end-to-end methods are often more susceptible than those with a pre-trained feature extractor; 2) Vulnerability mainly comes from localization error rather than classification error; 3) When corruptions occur in the middle of an action instance, TAD models tend to yield the largest performance drop. Besides building a benchmark, we further develop a simple but effective robust training method to defend against temporal corruptions, through the FrameDrop augmentation and Temporal-Robust Consistency loss. Remarkably, our approach not only improves robustness but also yields promising improvements on clean data. We believe that this study will serve as a benchmark for future research in robust video analysis. Source code and models are available at https://github.com/Alvin-Zeng/temporal-robustness-benchmark.

link: http://arxiv.org/abs/2403.20254v1