Sat 2024.03.23

Network Calculus Bounds for Time-Sensitive Networks: A Revisit

Yuming Jiang

Network calculus (NC), particularly its min-plus branch, has been extensively utilized to construct service models and compute delay bounds for time-sensitive networks (TSNs). This paper provides a revisit to the fundamental results. In particular, counterexamples to the most basic min-plus service models, which have been proposed for TSNs and used for computing delay bounds, indicate that the packetization effect has often been overlooked. To address, the max-plus branch of NC is also considered in this paper, whose models handle packetized traffic more explicitly. It is found that mapping the min-plus models to the max-plus models may bring in an immediate improvement over delay bounds derived from the min-plus analysis. In addition, an integrated analytical approach that combines models from both the min-plus and the max-plus NC branches is introduced. In this approach, the max-plus g-server model is extended and the extended model, called g^{x}-server, is used together with the min-plus arrival curve traffic model. By applying the integrated NC approach, service and delay bounds are derived for several settings that are fundamental in TSNs.

link: http://arxiv.org/abs/2403.13656v1

NELA-PS: A Dataset of Pink Slime News Articles for the Study of Local News Ecosystems

Benjamin D. Horne, Maurício Gruppi

Pink slime news outlets automatically produce low-quality, often partisan content that is framed as authentic local news. Given that local news is trusted by Americans and is increasingly shutting down due to financial distress, pink slime news outlets have the potential to exploit local information voids. Yet, there are gaps in understanding of pink slime production practices and tactics, particularly over time. Hence, to support future research in this area, we built a dataset of over 7.9M articles from 1093 pink slime sources over 2.5 years.

link: http://arxiv.org/abs/2403.13657v1

Multimodal Variational Autoencoder for Low-cost Cardiac Hemodynamics Instability Detection

Mohammod N. I. Suvon, Prasun C. Tripathi, Wenrui Fan, Shuo Zhou, Xianyuan Liu, Samer Alabed, Venet Osmani, Andrew J. Swift, Chen Chen, Haiping Lu

Recent advancements in non-invasive detection of cardiac hemodynamic instability (CHDI) primarily focus on applying machine learning techniques to a single data modality, e.g. cardiac magnetic resonance imaging (MRI). Despite their potential, these approaches often fall short especially when the size of labeled patient data is limited, a common challenge in the medical domain. Furthermore, only a few studies have explored multimodal methods to study CHDI, which mostly rely on costly modalities such as cardiac MRI and echocardiogram. In response to these limitations, we propose a novel multimodal variational autoencoder ($\text{CardioVAE}_\text{X,G}$) to integrate low-cost chest X-ray (CXR) and electrocardiogram (ECG) modalities with pre-training on a large unlabeled dataset. Specifically, $\text{CardioVAE}_\text{X,G}$ introduces a novel tri-stream pre-training strategy to learn both shared and modality-specific features, thus enabling fine-tuning with both unimodal and multimodal datasets. We pre-train $\text{CardioVAE}_\text{X,G}$ on a large, unlabeled dataset of $50,982$ subjects from a subset of MIMIC database and then fine-tune the pre-trained model on a labeled dataset of 795 subjects from the ASPIRE registry. Comprehensive evaluations against existing methods show that $\text{CardioVAE}_\text{X,G}$ offers promising performance (AUROC $=0.79$ and Accuracy $=0.77$), representing a significant step forward in non-invasive prediction of CHDI. Our model also excels in producing fine interpretations of predictions directly associated with clinical features, thereby supporting clinical decision-making.

link: http://arxiv.org/abs/2403.13658v1

Recursive Cross-Modal Attention for Multimodal Fusion in Dimensional Emotion Recognition

R. Gnana Praveen, Jahangir Alam

Multi-modal emotion recognition has recently gained a lot of attention since it can leverage diverse and complementary relationships over multiple modalities, such as audio, visual, and text. Most state-of-the-art methods for multimodal fusion rely on recurrent networks or conventional attention mechanisms that do not effectively leverage the complementary nature of the modalities. In this paper, we focus on dimensional emotion recognition based on the fusion of facial, vocal, and text modalities extracted from videos. Specifically, we propose a recursive cross-modal attention (RCMA) to effectively capture the complementary relationships across the modalities in a recursive fashion. The proposed model is able to effectively capture the inter-modal relationships by computing the cross-attention weights across the individual modalities and the joint representation of the other two modalities. To further improve the inter-modal relationships, the obtained attended features of the individual modalities are again fed as input to the cross-modal attention to refine the feature representations of the individual modalities. In addition to that, we have used Temporal convolution networks (TCNs) to capture the temporal modeling (intra-modal relationships) of the individual modalities. By deploying the TCNs as well cross-modal attention in a recursive fashion, we are able to effectively capture both intra- and inter-modal relationships across the audio, visual, and text modalities. Experimental results on validation-set videos from the AffWild2 dataset indicate that our proposed fusion model is able to achieve significant improvement over the baseline for the sixth challenge of Affective Behavior Analysis in-the-Wild 2024 (ABAW6) competition.

link: http://arxiv.org/abs/2403.13659v1

ProMamba: Prompt-Mamba for polyp segmentation

Jianhao Xie, Ruofan Liao, Ziang Zhang, Sida Yi, Yuesheng Zhu, Guibo Luo

Detecting polyps through colonoscopy is an important task in medical image segmentation, which provides significant assistance and reference value for clinical surgery. However, accurate segmentation of polyps is a challenging task due to two main reasons. Firstly, polyps exhibit various shapes and colors. Secondly, the boundaries between polyps and their normal surroundings are often unclear. Additionally, significant differences between different datasets lead to limited generalization capabilities of existing methods. To address these issues, we propose a segmentation model based on Prompt-Mamba, which incorporates the latest Vision-Mamba and prompt technologies. Compared to previous models trained on the same dataset, our model not only maintains high segmentation accuracy on the validation part of the same dataset but also demonstrates superior accuracy on unseen datasets, exhibiting excellent generalization capabilities. Notably, we are the first to apply the Vision-Mamba architecture to polyp segmentation and the first to utilize prompt technology in a polyp segmentation model. Our model efficiently accomplishes segmentation tasks, surpassing previous state-of-the-art methods by an average of 5% across six datasets. Furthermore, we have developed multiple versions of our model with scaled parameter counts, achieving better performance than previous models even with fewer parameters. Our code and trained weights will be released soon.

link: http://arxiv.org/abs/2403.13660v1

T-Pixel2Mesh: Combining Global and Local Transformer for 3D Mesh Generation from a Single Image

Shijie Zhang, Boyan Jiang, Keke He, Junwei Zhu, Ying Tai, Chengjie Wang, Yinda Zhang, Yanwei Fu

Pixel2Mesh (P2M) is a classical approach for reconstructing 3D shapes from a single color image through coarse-to-fine mesh deformation. Although P2M is capable of generating plausible global shapes, its Graph Convolution Network (GCN) often produces overly smooth results, causing the loss of fine-grained geometry details. Moreover, P2M generates non-credible features for occluded regions and struggles with the domain gap from synthetic data to real-world images, which is a common challenge for single-view 3D reconstruction methods. To address these challenges, we propose a novel Transformer-boosted architecture, named T-Pixel2Mesh, inspired by the coarse-to-fine approach of P2M. Specifically, we use a global Transformer to control the holistic shape and a local Transformer to progressively refine the local geometry details with graph-based point upsampling. To enhance real-world reconstruction, we present the simple yet effective Linear Scale Search (LSS), which serves as prompt tuning during the input preprocessing. Our experiments on ShapeNet demonstrate state-of-the-art performance, while results on real-world data show the generalization capability.

link: http://arxiv.org/abs/2403.13663v1

Grounding Spatial Relations in Text-Only Language Models

Gorka Azkune, Ander Salaberria, Eneko Agirre

This paper shows that text-only Language Models (LM) can learn to ground spatial relations like "left of" or "below" if they are provided with explicit location information of objects and they are properly trained to leverage those locations. We perform experiments on a verbalized version of the Visual Spatial Reasoning (VSR) dataset, where images are coupled with textual statements which contain real or fake spatial relations between two objects of the image. We verbalize the images using an off-the-shelf object detector, adding location tokens to every object label to represent their bounding boxes in textual form. Given the small size of VSR, we do not observe any improvement when using locations, but pretraining the LM over a synthetic dataset automatically derived by us improves results significantly when using location tokens. We thus show that locations allow LMs to ground spatial relations, with our text-only LMs outperforming Vision-and-Language Models and setting the new state-of-the-art for the VSR dataset. Our analysis show that our text-only LMs can generalize beyond the relations seen in the synthetic dataset to some extent, learning also more useful information than that encoded in the spatial rules we used to create the synthetic dataset itself.

link: http://dx.doi.org/10.1016/j.neunet.2023.11.031

DanceCamera3D: 3D Camera Movement Synthesis with Music and Dance

Zixuan Wang, Jia Jia, Shikun Sun, Haozhe Wu, Rong Han, Zhenyu Li, Di Tang, Jiaqing Zhou, Jiebo Luo

Choreographers determine what the dances look like, while cameramen determine the final presentation of dances. Recently, various methods and datasets have showcased the feasibility of dance synthesis. However, camera movement synthesis with music and dance remains an unsolved challenging problem due to the scarcity of paired data. Thus, we present DCM, a new multi-modal 3D dataset, which for the first time combines camera movement with dance motion and music audio. This dataset encompasses 108 dance sequences (3.2 hours) of paired dance-camera-music data from the anime community, covering 4 music genres. With this dataset, we uncover that dance camera movement is multifaceted and human-centric, and possesses multiple influencing factors, making dance camera synthesis a more challenging task compared to camera or dance synthesis alone. To overcome these difficulties, we propose DanceCamera3D, a transformer-based diffusion model that incorporates a novel body attention loss and a condition separation strategy. For evaluation, we devise new metrics measuring camera movement quality, diversity, and dancer fidelity. Utilizing these metrics, we conduct extensive experiments on our DCM dataset, providing both quantitative and qualitative evidence showcasing the effectiveness of our DanceCamera3D model. Code and video demos are available at https://github.com/Carmenw1203/DanceCamera3D-Official.

link: http://arxiv.org/abs/2403.13667v1

Retina Vision Transformer (RetinaViT): Introducing Scaled Patches into Vision Transformers

Yuyang Shu, Michael E. Bain

Humans see low and high spatial frequency components at the same time, and combine the information from both to form a visual scene. Drawing on this neuroscientific inspiration, we propose an altered Vision Transformer architecture where patches from scaled down versions of the input image are added to the input of the first Transformer Encoder layer. We name this model Retina Vision Transformer (RetinaViT) due to its inspiration from the human visual system. Our experiments show that when trained on the ImageNet-1K dataset with a moderate configuration, RetinaViT achieves a 3.3% performance improvement over the original ViT. We hypothesize that this improvement can be attributed to the inclusion of low spatial frequency components in the input, which improves the ability to capture structural features, and to select and forward important features to deeper layers. RetinaViT thereby opens doors to further investigations into vertical pathways and attention patterns.

link: http://arxiv.org/abs/2403.13677v1

AUD-TGN: Advancing Action Unit Detection with Temporal Convolution and GPT-2 in Wild Audiovisual Contexts

Jun Yu, Zerui Zhang, Zhihong Wei, Gongpeng Zhao, Zhongpeng Cai, Yongqi Wang, Guochen Xie, Jichao Zhu, Wangyuan Zhu

Leveraging the synergy of both audio data and visual data is essential for understanding human emotions and behaviors, especially in in-the-wild setting. Traditional methods for integrating such multimodal information often stumble, leading to less-than-ideal outcomes in the task of facial action unit detection. To overcome these shortcomings, we propose a novel approach utilizing audio-visual multimodal data. This method enhances audio feature extraction by leveraging Mel Frequency Cepstral Coefficients (MFCC) and Log-Mel spectrogram features alongside a pre-trained VGGish network. Moreover, this paper adaptively captures fusion features across modalities by modeling the temporal relationships, and ultilizes a pre-trained GPT-2 model for sophisticated context-aware fusion of multimodal information. Our method notably improves the accuracy of AU detection by understanding the temporal and contextual nuances of the data, showcasing significant advancements in the comprehension of intricate scenarios. These findings underscore the potential of integrating temporal dynamics and contextual interpretation, paving the way for future research endeavors.

link: http://arxiv.org/abs/2403.13678v1

RoleInteract: Evaluating the Social Interaction of Role-Playing Agents

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, Jingren Zhou

Large language models (LLMs) have advanced the development of various AI conversational agents, including role-playing conversational agents that mimic diverse characters and human behaviors. While prior research has predominantly focused on enhancing the conversational capability, role-specific knowledge, and stylistic attributes of these agents, there has been a noticeable gap in assessing their social intelligence. In this paper, we introduce RoleInteract, the first benchmark designed to systematically evaluate the sociality of role-playing conversational agents at both individual and group levels of social interactions. The benchmark is constructed from a variety of sources and covers a wide range of 500 characters and over 6,000 question prompts and 30,800 multi-turn role-playing utterances. We conduct comprehensive evaluations on this benchmark using mainstream open-source and closed-source LLMs. We find that agents excelling in individual level does not imply their proficiency in group level. Moreover, the behavior of individuals may drift as a result of the influence exerted by other agents within the group. Experimental results on RoleInteract confirm its significance as a testbed for assessing the social interaction of role-playing conversational agents. The benchmark is publicly accessible at https://github.com/X-PLUG/RoleInteract.

link: http://arxiv.org/abs/2403.13679v2

Step-Calibrated Diffusion for Biomedical Optical Image Restoration

Yiwei Lyu, Sung Jik Cha, Cheng Jiang, Asadur Chowdury, Xinhai Hou, Edward Harake, Akhil Kondepudi, Christian Freudiger, Honglak Lee, Todd C. Hollon

High-quality, high-resolution medical imaging is essential for clinical care. Raman-based biomedical optical imaging uses non-ionizing infrared radiation to evaluate human tissues in real time and is used for early cancer detection, brain tumor diagnosis, and intraoperative tissue analysis. Unfortunately, optical imaging is vulnerable to image degradation due to laser scattering and absorption, which can result in diagnostic errors and misguided treatment. Restoration of optical images is a challenging computer vision task because the sources of image degradation are multi-factorial, stochastic, and tissue-dependent, preventing a straightforward method to obtain paired low-quality/high-quality data. Here, we present Restorative Step-Calibrated Diffusion (RSCD), an unpaired image restoration method that views the image restoration problem as completing the finishing steps of a diffusion-based image generation task. RSCD uses a step calibrator model to dynamically determine the severity of image degradation and the number of steps required to complete the reverse diffusion process for image restoration. RSCD outperforms other widely used unpaired image restoration methods on both image quality and perceptual evaluation metrics for restoring optical images. Medical imaging experts consistently prefer images restored using RSCD in blinded comparison experiments and report minimal to no hallucinations. Finally, we show that RSCD improves performance on downstream clinical imaging tasks, including automated brain tumor diagnosis and deep tissue imaging. Our code is available at https://github.com/MLNeurosurg/restorative_step-calibrated_diffusion.

link: http://arxiv.org/abs/2403.13680v1

PARAMANU-AYN: An Efficient Novel Generative and Instruction-tuned Language Model for Indian Legal Case Documents

Mitodru Niyogi, Arnab Bhattacharya

In this paper, we present PARAMANU-AYN, a language model based exclusively on case documents of the Supreme Court of India, the Constitution of India, and the Indian Penal Code. The novel Auto Regressive (AR) decoder based model is pretrained from scratch at a context size of 8192. We evaluated our pretrained legal model on perplexity metrics. We also instruction-tuned our pretrained model on a set of 10,763 instructions covering various legal tasks such as legal reasoning, judgement explanation, legal clause generation, legal drafting, legal contract drafting, case summarization, constitutional question-answering, etc. We also evaluated the responses of prompts for instruction-tuned models by GPT-3.5-Turbo on clarity, relevance, completeness, and legal reasoning metrics in a scale of 10. Our model can be run on CPU and achieved 42.46 tokens/sec CPU inference speed. We found that our models, despite not being pretrained on legal books, various legal contracts, and legal documents, were able to learn the domain knowledge required for drafting various legal contracts and legal clauses, and generalize to draft legal contracts and legal clauses with limited instruction tuning. Hence, we conclude that for a strong domain-specialized generative language model (such as legal), very large amounts of data are not required to develop models from scratch. We believe that this work is the first attempt to make a dedicated generative legal language model from scratch for Indian Supreme Court jurisdiction or in legal NLP overall. We plan to release our Paramanu-Ayn model at https://www.bharatgpts.com.

link: http://arxiv.org/abs/2403.13681v1

Threats, Attacks, and Defenses in Machine Unlearning: A Survey

Ziyao Liu, Huanyi Ye, Chen Chen, Kwok-Yan Lam

Recently, Machine Unlearning (MU) has gained considerable attention for its potential to improve AI safety by removing the influence of specific data from trained Machine Learning (ML) models. This process, known as knowledge removal, addresses concerns about data such as sensitivity, copyright restrictions, obsolescence, or low quality. This capability is also crucial for ensuring compliance with privacy regulations such as the Right To Be Forgotten (RTBF). Therefore, strategic knowledge removal mitigates the risk of harmful outcomes, safeguarding against biases, misinformation, and unauthorized data exploitation, thereby enhancing the ethical use and reliability of AI systems. Efforts have been made to design efficient unlearning approaches, with MU services being examined for integration with existing machine learning as a service (MLaaS), allowing users to submit requests to erase data. However, recent research highlights vulnerabilities in machine unlearning systems, such as information leakage and malicious unlearning requests, that can lead to significant security and privacy concerns. Moreover, extensive research indicates that unlearning methods and prevalent attacks fulfill diverse roles within MU systems. For instance, unlearning can act as a mechanism to recover models from backdoor attacks, while backdoor attacks themselves can serve as an evaluation metric for unlearning effectiveness. This underscores the intricate relationship and complex interplay between these elements in maintaining system functionality and safety. Therefore, this survey seeks to bridge the gap between the extensive number of studies on threats, attacks, and defenses in machine unlearning and the absence of a comprehensive review that categorizes their taxonomy, methods, and solutions, thus offering valuable insights for future research directions and practical implementations.

link: http://arxiv.org/abs/2403.13682v1

DVMNet: Computing Relative Pose for Unseen Objects Beyond Hypotheses

Chen Zhao, Tong Zhang, Zheng Dang, Mathieu Salzmann

Determining the relative pose of an object between two images is pivotal to the success of generalizable object pose estimation. Existing approaches typically approximate the continuous pose representation with a large number of discrete pose hypotheses, which incurs a computationally expensive process of scoring each hypothesis at test time. By contrast, we present a Deep Voxel Matching Network (DVMNet) that eliminates the need for pose hypotheses and computes the relative object pose in a single pass. To this end, we map the two input RGB images, reference and query, to their respective voxelized 3D representations. We then pass the resulting voxels through a pose estimation module, where the voxels are aligned and the pose is computed in an end-to-end fashion by solving a least-squares problem. To enhance robustness, we introduce a weighted closest voxel algorithm capable of mitigating the impact of noisy voxels. We conduct extensive experiments on the CO3D, LINEMOD, and Objaverse datasets, demonstrating that our method delivers more accurate relative pose estimates for novel objects at a lower computational cost compared to state-of-the-art methods. Our code is released at: https://github.com/sailor-z/DVMNet/.

link: http://arxiv.org/abs/2403.13683v1