Wed 2024.04.10

Dynamic Distinction Learning: Adaptive Pseudo Anomalies for Video Anomaly Detection

Demetris Lappas, Vasileios Argyriou, Dimitrios Makris

We introduce Dynamic Distinction Learning (DDL) for Video Anomaly Detection, a novel video anomaly detection methodology that combines pseudo-anomalies, dynamic anomaly weighting, and a distinction loss function to improve detection accuracy. By training on pseudo-anomalies, our approach adapts to the variability of normal and anomalous behaviors without fixed anomaly thresholds. Our model showcases superior performance on the Ped2, Avenue and ShanghaiTech datasets, where individual models are tailored for each scene. These achievements highlight DDL's effectiveness in advancing anomaly detection, offering a scalable and adaptable solution for video surveillance challenges.

link: http://arxiv.org/abs/2404.04986v1

MLaKE: Multilingual Knowledge Editing Benchmark for Large Language Models

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, Xueqi Cheng

The extensive utilization of large language models (LLMs) underscores the crucial necessity for precise and contemporary knowledge embedded within their intrinsic parameters. Existing research on knowledge editing primarily concentrates on monolingual scenarios, neglecting the complexities presented by multilingual contexts and multi-hop reasoning. To address these challenges, our study introduces MLaKE (Multilingual Language Knowledge Editing), a novel benchmark comprising 4072 multi-hop and 5360 single-hop questions designed to evaluate the adaptability of knowledge editing methods across five languages: English, Chinese, Japanese, French, and German. MLaKE aggregates fact chains from Wikipedia across languages and utilizes LLMs to generate questions in both free-form and multiple-choice. We evaluate the multilingual knowledge editing generalization capabilities of existing methods on MLaKE. Existing knowledge editing methods demonstrate higher success rates in English samples compared to other languages. However, their generalization capabilities are limited in multi-language experiments. Notably, existing knowledge editing methods often show relatively high generalization for languages within the same language family compared to languages from different language families. These results underscore the imperative need for advancements in multilingual knowledge editing and we hope MLaKE can serve as a valuable resource for benchmarking and solution development.

link: http://arxiv.org/abs/2404.04990v1

Efficient Surgical Tool Recognition via HMM-Stabilized Deep Learning

Haifeng Wang, Hao Xu, Jun Wang, Jian Zhou, Ke Deng

Recognizing various surgical tools, actions and phases from surgery videos is an important problem in computer vision with exciting clinical applications. Existing deep-learning-based methods for this problem either process each surgical video as a series of independent images without considering their dependence, or rely on complicated deep learning models to count for dependence of video frames. In this study, we revealed from exploratory data analysis that surgical videos enjoy relatively simple semantic structure, where the presence of surgical phases and tools can be well modeled by a compact hidden Markov model (HMM). Based on this observation, we propose an HMM-stabilized deep learning method for tool presence detection. A wide range of experiments confirm that the proposed approaches achieve better performance with lower training and running costs, and support more flexible ways to construct and utilize training data in scenarios where not all surgery videos of interest are extensively labelled. These results suggest that popular deep learning approaches with over-complicated model structures may suffer from inefficient utilization of data, and integrating ingredients of deep learning and statistical learning wisely may lead to more powerful algorithms that enjoy competitive performance, transparent interpretation and convenient model training simultaneously.

link: http://arxiv.org/abs/2404.04992v1

Fantastic Animals and Where to Find Them: Segment Any Marine Animal with Dual SAM

Pingping Zhang, Tianyu Yan, Yang Liu, Huchuan Lu

As an important pillar of underwater intelligence, Marine Animal Segmentation (MAS) involves segmenting animals within marine environments. Previous methods don't excel in extracting long-range contextual features and overlook the connectivity between discrete pixels. Recently, Segment Anything Model (SAM) offers a universal framework for general segmentation tasks. Unfortunately, trained with natural images, SAM does not obtain the prior knowledge from marine images. In addition, the single-position prompt of SAM is very insufficient for prior guidance. To address these issues, we propose a novel feature learning framework, named Dual-SAM for high-performance MAS. To this end, we first introduce a dual structure with SAM's paradigm to enhance feature learning of marine images. Then, we propose a Multi-level Coupled Prompt (MCP) strategy to instruct comprehensive underwater prior information, and enhance the multi-level features of SAM's encoder with adapters. Subsequently, we design a Dilated Fusion Attention Module (DFAM) to progressively integrate multi-level features from SAM's encoder. Finally, instead of directly predicting the masks of marine animals, we propose a Criss-Cross Connectivity Prediction (C^{3}P) paradigm to capture the inter-connectivity between discrete pixels. With dual decoders, it generates pseudo-labels and achieves mutual supervision for complementary feature representations, resulting in considerable improvements over previous techniques. Extensive experiments verify that our proposed method achieves state-of-the-art performances on five widely-used MAS datasets. The code is available at https://github.com/Drchip61/Dual_SAM.

link: http://arxiv.org/abs/2404.04996v1

Adapting LLMs for Efficient Context Processing through Soft Prompt Compression

Cangqing Wang, Yutian Yang, Ruisi Li, Dan Sun, Ruicong Cai, Yuzhu Zhang, Chengqian Fu, Lillian Floyd

The rapid advancement of Large Language Models (LLMs) has inaugurated a transformative epoch in natural language processing, fostering unprecedented proficiency in text generation, comprehension, and contextual scrutiny. Nevertheless, effectively handling extensive contexts, crucial for myriad applications, poses a formidable obstacle owing to the intrinsic constraints of the models' context window sizes and the computational burdens entailed by their operations. This investigation presents an innovative framework that strategically tailors LLMs for streamlined context processing by harnessing the synergies among natural language summarization, soft prompt compression, and augmented utility preservation mechanisms. Our methodology, dubbed SoftPromptComp, amalgamates natural language prompts extracted from summarization methodologies with dynamically generated soft prompts to forge a concise yet semantically robust depiction of protracted contexts. This depiction undergoes further refinement via a weighting mechanism optimizing information retention and utility for subsequent tasks. We substantiate that our framework markedly diminishes computational overhead and enhances LLMs' efficacy across various benchmarks, while upholding or even augmenting the caliber of the produced content. By amalgamating soft prompt compression with sophisticated summarization, SoftPromptComp confronts the dual challenges of managing lengthy contexts and ensuring model scalability. Our findings point towards a propitious trajectory for augmenting LLMs' applicability and efficiency, rendering them more versatile and pragmatic for real-world applications. This research enriches the ongoing discourse on optimizing language models, providing insights into the potency of soft prompts and summarization techniques as pivotal instruments for the forthcoming generation of NLP solutions.

link: http://arxiv.org/abs/2404.04997v1

Weakly Supervised Deep Hyperspherical Quantization for Image Retrieval

Jinpeng Wang, Bin Chen, Qiang Zhang, Zaiqiao Meng, Shangsong Liang, Shu-Tao Xia

Deep quantization methods have shown high efficiency on large-scale image retrieval. However, current models heavily rely on ground-truth information, hindering the application of quantization in label-hungry scenarios. A more realistic demand is to learn from inexhaustible uploaded images that are associated with informal tags provided by amateur users. Though such sketchy tags do not obviously reveal the labels, they actually contain useful semantic information for supervising deep quantization. To this end, we propose Weakly-Supervised Deep Hyperspherical Quantization (WSDHQ), which is the first work to learn deep quantization from weakly tagged images. Specifically, 1) we use word embeddings to represent the tags and enhance their semantic information based on a tag correlation graph. 2) To better preserve semantic information in quantization codes and reduce quantization error, we jointly learn semantics-preserving embeddings and supervised quantizer on hypersphere by employing a well-designed fusion layer and tailor-made loss functions. Extensive experiments show that WSDHQ can achieve state-of-art performance on weakly-supervised compact coding. Code is available at https://github.com/gimpong/AAAI21-WSDHQ.

link: http://dx.doi.org/10.1609/aaai.v35i4.16380

Dual-Scale Transformer for Large-Scale Single-Pixel Imaging

Gang Qu, Ping Wang, Xin Yuan

Single-pixel imaging (SPI) is a potential computational imaging technique which produces image by solving an illposed reconstruction problem from few measurements captured by a single-pixel detector. Deep learning has achieved impressive success on SPI reconstruction. However, previous poor reconstruction performance and impractical imaging model limit its real-world applications. In this paper, we propose a deep unfolding network with hybrid-attention Transformer on Kronecker SPI model, dubbed HATNet, to improve the imaging quality of real SPI cameras. Specifically, we unfold the computation graph of the iterative shrinkagethresholding algorithm (ISTA) into two alternative modules: efficient tensor gradient descent and hybrid-attention multiscale denoising. By virtue of Kronecker SPI, the gradient descent module can avoid high computational overheads rooted in previous gradient descent modules based on vectorized SPI. The denoising module is an encoder-decoder architecture powered by dual-scale spatial attention for high- and low-frequency aggregation and channel attention for global information recalibration. Moreover, we build a SPI prototype to verify the effectiveness of the proposed method. Extensive experiments on synthetic and real data demonstrate that our method achieves the state-of-the-art performance. The source code and pre-trained models are available at https://github.com/Gang-Qu/HATNet-SPI.

link: http://arxiv.org/abs/2404.05001v1

Camera-Based Remote Physiology Sensing for Hundreds of Subjects Across Skin Tones

Jiankai Tang, Xinyi Li, Jiacheng Liu, Xiyuxing Zhang, Zeyu Wang, Yuntao Wang

Remote photoplethysmography (rPPG) emerges as a promising method for non-invasive, convenient measurement of vital signs, utilizing the widespread presence of cameras. Despite advancements, existing datasets fall short in terms of size and diversity, limiting comprehensive evaluation under diverse conditions. This paper presents an in-depth analysis of the VitalVideo dataset, the largest real-world rPPG dataset to date, encompassing 893 subjects and 6 Fitzpatrick skin tones. Our experimentation with six unsupervised methods and three supervised models demonstrates that datasets comprising a few hundred subjects(i.e., 300 for UBFC-rPPG, 500 for PURE, and 700 for MMPD-Simple) are sufficient for effective rPPG model training. Our findings highlight the importance of diversity and consistency in skin tones for precise performance evaluation across different datasets.

link: http://arxiv.org/abs/2404.05003v1

Towards Reliable and Empathetic Depression-Diagnosis-Oriented Chats

Kunyao Lan, Cong Ming, Binwei Yao, Lu Chen, Mengyue Wu

Chatbots can serve as a viable tool for preliminary depression diagnosis via interactive conversations with potential patients. Nevertheless, the blend of task-oriented and chit-chat in diagnosis-related dialogues necessitates professional expertise and empathy. Such unique requirements challenge traditional dialogue frameworks geared towards single optimization goals. To address this, we propose an innovative ontology definition and generation framework tailored explicitly for depression diagnosis dialogues, combining the reliability of task-oriented conversations with the appeal of empathy-related chit-chat. We further apply the framework to D^4, the only existing public dialogue dataset on depression diagnosis-oriented chats. Exhaustive experimental results indicate significant improvements in task completion and emotional support generation in depression diagnosis, fostering a more comprehensive approach to task-oriented chat dialogue system development and its applications in digital mental health.

link: http://arxiv.org/abs/2404.05012v1

MagicTime: Time-lapse Video Generation Models as Metamorphic Simulators

Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, Jiebo Luo

Recent advances in Text-to-Video generation (T2V) have achieved remarkable success in synthesizing high-quality general videos from textual descriptions. A largely overlooked problem in T2V is that existing models have not adequately encoded physical knowledge of the real world, thus generated videos tend to have limited motion and poor variations. In this paper, we propose \textbf{MagicTime}, a metamorphic time-lapse video generation model, which learns real-world physics knowledge from time-lapse videos and implements metamorphic generation. First, we design a MagicAdapter scheme to decouple spatial and temporal training, encode more physical knowledge from metamorphic videos, and transform pre-trained T2V models to generate metamorphic videos. Second, we introduce a Dynamic Frames Extraction strategy to adapt to metamorphic time-lapse videos, which have a wider variation range and cover dramatic object metamorphic processes, thus embodying more physical knowledge than general videos. Finally, we introduce a Magic Text-Encoder to improve the understanding of metamorphic video prompts. Furthermore, we create a time-lapse video-text dataset called \textbf{ChronoMagic}, specifically curated to unlock the metamorphic video generation ability. Extensive experiments demonstrate the superiority and effectiveness of MagicTime for generating high-quality and dynamic metamorphic videos, suggesting time-lapse video generation is a promising path toward building metamorphic simulators of the physical world.

link: http://arxiv.org/abs/2404.05014v1

Hyperbolic Learning with Synthetic Captions for Open-World Detection

Fanjie Kong, Yanbei Chen, Jiarui Cai, Davide Modolo

Open-world detection poses significant challenges, as it requires the detection of any object using either object class labels or free-form texts. Existing related works often use large-scale manual annotated caption datasets for training, which are extremely expensive to collect. Instead, we propose to transfer knowledge from vision-language models (VLMs) to enrich the open-vocabulary descriptions automatically. Specifically, we bootstrap dense synthetic captions using pre-trained VLMs to provide rich descriptions on different regions in images, and incorporate these captions to train a novel detector that generalizes to novel concepts. To mitigate the noise caused by hallucination in synthetic captions, we also propose a novel hyperbolic vision-language learning approach to impose a hierarchy between visual and caption embeddings. We call our detector ``HyperLearner''. We conduct extensive experiments on a wide variety of open-world detection benchmarks (COCO, LVIS, Object Detection in the Wild, RefCOCO) and our results show that our model consistently outperforms existing state-of-the-art methods, such as GLIP, GLIPv2 and Grounding DINO, when using the same backbone.

link: http://arxiv.org/abs/2404.05016v1

Shortcut-connected Expert Parallelism for Accelerating Mixture-of-Experts

Weilin Cai, Juyong Jiang, Le Qin, Junwei Cui, Sunghun Kim, Jiayi Huang

Expert parallelism has been introduced as a strategy to distribute the computational workload of sparsely-gated mixture-of-experts (MoE) models across multiple computing devices, facilitating the execution of these increasingly large-scale models. However, the All-to-All communication intrinsic to expert parallelism constitutes a significant overhead, diminishing the MoE models' efficiency. Current optimization approaches offer some relief, yet they are constrained by the sequential interdependence of communication and computation operations. To address this limitation, we present a novel shortcut-connected MoE architecture with overlapping parallel strategy, designated as ScMoE, which effectively decouples communication from its conventional sequence, allowing for a substantial overlap of 70% to 100% with computation. When compared with the prevalent top-2 MoE architecture, ScMoE demonstrates training speed improvements of 30% and 11%, and inference improvements of 40% and 15%, in our PCIe and NVLink hardware environments, respectively, where communication constitutes 60% and 15% of the total MoE time consumption. On the other hand, extensive experiments and theoretical analyses indicate that ScMoE not only achieves comparable but in some instances surpasses the model quality of existing approaches in vision and language tasks.

link: http://arxiv.org/abs/2404.05019v1

Context-dependent Causality (the Non-Nonotonic Case)

Nir Billfeld, Moshe Kim

We develop a novel identification strategy as well as a new estimator for context-dependent causal inference in non-parametric triangular models with non-separable disturbances. Departing from the common practice, our analysis does not rely on the strict monotonicity assumption. Our key contribution lies in leveraging on diffusion models to formulate the structural equations as a system evolving from noise accumulation to account for the influence of the latent context (confounder) on the outcome. Our identifiability strategy involves a system of Fredholm integral equations expressing the distributional relationship between a latent context variable and a vector of observables. These integral equations involve an unknown kernel and are governed by a set of structural form functions, inducing a non-monotonic inverse problem. We prove that if the kernel density can be represented as an infinite mixture of Gaussians, then there exists a unique solution for the unknown function. This is a significant result, as it shows that it is possible to solve a non-monotonic inverse problem even when the kernel is unknown. On the methodological front we leverage on a novel and enriched Contaminated Generative Adversarial (Neural) Networks (CONGAN) which we provide as a solution to the non-monotonic inverse problem.

link: http://arxiv.org/abs/2404.05021v1

DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology

Valentin Koch, Sophia J. Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia Schnabel, Tingying Peng, Carsten Marr

In hematology, computational models offer significant potential to improve diagnostic accuracy, streamline workflows, and reduce the tedious work of analyzing single cells in peripheral blood or bone marrow smears. However, clinical adoption of computational models has been hampered by the lack of generalization due to large batch effects, small dataset sizes, and poor performance in transfer learning from natural images. To address these challenges, we introduce DinoBloom, the first foundation model for single cell images in hematology, utilizing a tailored DINOv2 pipeline. Our model is built upon an extensive collection of 13 diverse, publicly available datasets of peripheral blood and bone marrow smears, the most substantial open-source cohort in hematology so far, comprising over 380,000 white blood cell images. To assess its generalization capability, we evaluate it on an external dataset with a challenging domain shift. We show that our model outperforms existing medical and non-medical vision models in (i) linear probing and k-nearest neighbor evaluations for cell-type classification on blood and bone marrow smears and (ii) weakly supervised multiple instance learning for acute myeloid leukemia subtyping by a large margin. A family of four DinoBloom models (small, base, large, and giant) can be adapted for a wide range of downstream applications, be a strong baseline for classification problems, and facilitate the assessment of batch effects in new datasets. All models are available at github.com/marrlab/DinoBloom.

link: http://arxiv.org/abs/2404.05022v1

Scalable and Efficient Hierarchical Visual Topological Mapping

Saravanabalagi Ramachandran, Jonathan Horgan, Ganesh Sistu, John McDonald

Hierarchical topological representations can significantly reduce search times within mapping and localization algorithms. Although recent research has shown the potential for such approaches, limited consideration has been given to the suitability and comparative performance of different global feature representations within this context. In this work, we evaluate state-of-the-art hand-crafted and learned global descriptors using a hierarchical topological mapping technique on benchmark datasets and present results of a comprehensive evaluation of the impact of the global descriptor used. Although learned descriptors have been incorporated into place recognition methods to improve retrieval accuracy and enhance overall recall, the problem of scalability and efficiency when applied to longer trajectories has not been adequately addressed in a majority of research studies. Based on our empirical analysis of multiple runs, we identify that continuity and distinctiveness are crucial characteristics for an optimal global descriptor that enable efficient and scalable hierarchical mapping, and present a methodology for quantifying and contrasting these characteristics across different global descriptors. Our study demonstrates that the use of global descriptors based on an unsupervised learned Variational Autoencoder (VAE) excels in these characteristics and achieves significantly lower runtime. It runs on a consumer grade desktop, up to 2.3x faster than the second best global descriptor, NetVLAD, and up to 9.5x faster than the hand-crafted descriptor, PHOG, on the longest track evaluated (St Lucia, 17.6 km), without sacrificing overall recall performance.

link: http://dx.doi.org/10.1109/ICAR58858.2023.10406394