Thu 2024.04.04

SyncMask: Synchronized Attentional Masking for Fashion-centric Vision-Language Pretraining

Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, Yeong Hyeon Gu

Vision-language models (VLMs) have made significant strides in cross-modal understanding through large-scale paired datasets. However, in fashion domain, datasets often exhibit a disparity between the information conveyed in image and text. This issue stems from datasets containing multiple images of a single fashion item all paired with one text, leading to cases where some textual details are not visible in individual images. This mismatch, particularly when non-co-occurring elements are masked, undermines the training of conventional VLM objectives like Masked Language Modeling and Masked Image Modeling, thereby hindering the model's ability to accurately align fine-grained visual and textual features. Addressing this problem, we propose Synchronized attentional Masking (SyncMask), which generate masks that pinpoint the image patches and word tokens where the information co-occur in both image and text. This synchronization is accomplished by harnessing cross-attentional features obtained from a momentum model, ensuring a precise alignment between the two modalities. Additionally, we enhance grouped batch sampling with semi-hard negatives, effectively mitigating false negative issues in Image-Text Matching and Image-Text Contrastive learning objectives within fashion datasets. Our experiments demonstrate the effectiveness of the proposed approach, outperforming existing methods in three downstream tasks.

link: http://arxiv.org/abs/2404.01156v1

Green AI: Exploring Carbon Footprints, Mitigation Strategies, and Trade Offs in Large Language Model Training

Vivian Liu, Yiqiao Yin

Prominent works in the field of Natural Language Processing have long attempted to create new innovative models by improving upon previous model training approaches, altering model architecture, and developing more in-depth datasets to better their performance. However, with the quickly advancing field of NLP comes increased greenhouse gas emissions, posing concerns over the environmental damage caused by training LLMs. Gaining a comprehensive understanding of the various costs, particularly those pertaining to environmental aspects, that are associated with artificial intelligence serves as the foundational basis for ensuring safe AI models. Currently, investigations into the CO2 emissions of AI models remain an emerging area of research, and as such, in this paper, we evaluate the CO2 emissions of well-known large language models, which have an especially high carbon footprint due to their significant amount of model parameters. We argue for the training of LLMs in a way that is responsible and sustainable by suggesting measures for reducing carbon emissions. Furthermore, we discuss how the choice of hardware affects CO2 emissions by contrasting the CO2 emissions during model training for two widely used GPUs. Based on our results, we present the benefits and drawbacks of our proposed solutions and make the argument for the possibility of training more environmentally safe AI models without sacrificing their robustness and performance.

link: http://arxiv.org/abs/2404.01157v1

Dialogue with Robots: Proposals for Broadening Participation and Research in the SLIVAR Community

Casey Kennington, Malihe Alikhani, Heather Pon-Barry, Katherine Atwell, Yonatan Bisk, Daniel Fried, Felix Gervits, Zhao Han, Mert Inan, Michael Johnston, Raj Korpan, Diane Litman, Matthew Marge, Cynthia Matuszek, Ross Mead, Shiwali Mohan, Raymond Mooney, Natalie Parde, Jivko Sinapov, Angela Stewart, Matthew Stone, Stefanie Tellex, Tom Williams

The ability to interact with machines using natural human language is becoming not just commonplace, but expected. The next step is not just text interfaces, but speech interfaces and not

just with computers, but with all machines including robots. In this paper, we chronicle the recent history of this growing field of spoken dialogue with robots and offer the community three proposals, the first focused on education, the second on benchmarks, and the third on the modeling of language when it comes to spoken interaction with robots. The three proposals should act as white papers for any researcher to take and build upon.

link: http://arxiv.org/abs/2404.01158v1

GPU-accelerated Evolutionary Multiobjective Optimization Using Tensorized RVEA

Zhenyu Liang, Tao Jiang, Kebin Sun, Ran Cheng

Evolutionary multiobjective optimization has witnessed remarkable progress during the past decades. However, existing algorithms often encounter computational challenges in large-scale scenarios, primarily attributed to the absence of hardware acceleration. In response, we introduce a Tensorized Reference Vector Guided Evolutionary Algorithm (TensorRVEA) for harnessing the advancements of GPU acceleration. In TensorRVEA, the key data structures and operators are fully transformed into tensor forms for leveraging GPU-based parallel computing. In numerical benchmark tests involving large-scale populations and problem dimensions, TensorRVEA consistently demonstrates high computational performance, achieving up to over 1000$\times$ speedups. Then, we applied TensorRVEA to the domain of multiobjective neuroevolution for addressing complex challenges in robotic control tasks. Furthermore, we assessed TensorRVEA's extensibility by altering several tensorized reproduction operators. Experimental results demonstrate promising scalability and robustness of TensorRVEA. Source codes are available at https://github.com/EMI-Group/tensorrvea.

link: http://arxiv.org/abs/2404.01159v1

Diagnosis of Skin Cancer Using VGG16 and VGG19 Based Transfer Learning Models

Amir Faghihi, Mohammadreza Fathollahi, Roozbeh Rajabi

Today, skin cancer is considered as one of the most dangerous and common cancers in the world which demands special attention. Skin cancer may be developed in different types; including melanoma, actinic keratosis, basal cell carcinoma, squamous cell carcinoma, and Merkel cell carcinoma. Among them, melanoma is more unpredictable. Melanoma cancer can be diagnosed at early stages increasing the possibility of disease treatment. Automatic classification of skin lesions is a challenging task due to diverse forms and grades of the disease, demanding the requirement of novel methods implementation. Deep convolution neural networks (CNN) have shown an excellent potential for data and image classification. In this article, we inspect skin lesion classification problem using CNN techniques. Remarkably, we present that prominent classification accuracy of lesion detection can be obtained by proper designing and applying of transfer learning framework on pre-trained neural networks, without any requirement for data enlargement procedures i.e. merging VGG16 and VGG19 architectures pre-trained by a generic dataset with modified AlexNet network, and then, fine-tuned by a subject-specific dataset containing dermatology images. The convolution neural network was trained using 2541 images and, in particular, dropout was used to prevent the network from overfitting. Finally, the validity of the model was checked by applying the K-fold cross validation method. The proposed model increased classification accuracy by 3% (from 94.2% to 98.18%) in comparison with other methods.

link: http://arxiv.org/abs/2404.01160v1

Capturing Shock Waves by Relaxation Neural Networks

Nan Zhou, Zheng Ma

In this paper, we put forward a neural network framework to solve the nonlinear hyperbolic systems. This framework, named relaxation neural networks(RelaxNN), is a simple and scalable extension of physics-informed neural networks(PINN). It is shown later that a typical PINN framework struggles to handle shock waves that arise in hyperbolic systems' solutions. This ultimately results in the failure of optimization that is based on gradient descent in the training process. Relaxation systems

provide a smooth asymptotic to the discontinuity solution, under the expectation that macroscopic problems can be solved from a microscopic perspective. Based on relaxation systems, the RelaxNN framework alleviates the conflict of losses in the training process of the PINN framework. In addition to the remarkable results demonstrated in numerical simulations, most of the acceleration techniques and improvement strategies aimed at the standard PINN framework can also be applied to the RelaxNN framework.

link: http://arxiv.org/abs/2404.01163v1

LITE: Modeling Environmental Ecosystems with Multimodal Large Language Models

Haoran Li, Junqi Liu, Zexian Wang, Shiyuan Luo, Xiaowei Jia, Huaxiu Yao

The modeling of environmental ecosystems plays a pivotal role in the sustainable management of our planet. Accurate prediction of key environmental variables over space and time can aid in informed policy and decision-making, thus improving people's livelihood. Recently, deep learning-based methods have shown promise in modeling the spatial-temporal relationships for predicting environmental variables. However, these approaches often fall short in handling incomplete features and distribution shifts, which are commonly observed in environmental data due to the substantial cost of data collection and malfunctions in measuring instruments. To address these issues, we propose LITE -- a multimodal large language model for environmental ecosystems modeling. Specifically, LITE unifies different environmental variables by transforming them into natural language descriptions and line graph images. Then, LITE utilizes unified encoders to capture spatial-temporal dynamics and correlations in different modalities. During this step, the incomplete features are imputed by a sparse Mixture-of-Experts framework, and the distribution shift is handled by incorporating multi-granularity information from past observations. Finally, guided by domain instructions, a language model is employed to fuse the multimodal representations for the prediction. Our experiments demonstrate that LITE significantly enhances performance in environmental spatial-temporal prediction across different domains compared to the best baseline, with a 41.25% reduction in prediction error. This justifies its effectiveness. Our data and code are available at https://github.com/hrlics/LITE.

link: http://arxiv.org/abs/2404.01165v1

Mirror-3DGS: Incorporating Mirror Reflections into 3D Gaussian Splatting

Jiarui Meng, Haijie Li, Yanmin Wu, Qiankun Gao, Shuzhou Yang, Jian Zhang, Siwei Ma

3D Gaussian Splatting (3DGS) has marked a significant breakthrough in the realm of 3D scene reconstruction and novel view synthesis. However, 3DGS, much like its predecessor Neural Radiance Fields (NeRF), struggles to accurately model physical reflections, particularly in mirrors that are ubiquitous in real-world scenes. This oversight mistakenly perceives reflections as separate entities that physically exist, resulting in inaccurate reconstructions and inconsistent reflective properties across varied viewpoints. To address this pivotal challenge, we introduce Mirror-3DGS, an innovative rendering framework devised to master the intricacies of mirror geometries and reflections, paving the way for the generation of realistically depicted mirror reflections. By ingeniously incorporating mirror attributes into the 3DGS and leveraging the principle of plane mirror imaging, Mirror-3DGS crafts a mirrored viewpoint to observe from behind the mirror, enriching the realism of scene renderings. Extensive assessments, spanning both synthetic and real-world scenes, showcase our method's ability to render novel views with enhanced fidelity in real-time, surpassing the state-of-the-art Mirror-NeRF specifically within the challenging mirror regions. Our code will be made publicly available for reproducible research.

link: http://arxiv.org/abs/2404.01168v1

SpikeMba: Multi-Modal Spiking Saliency Mamba for Temporal Video Grounding

Wenrui Li, Xiaopeng Hong, Xiaopeng Fan

Temporal video grounding (TVG) is a critical task in video content understanding. Despite significant advancements, existing methods often limit in capturing the fine-grained relationships

between multimodal inputs and the high computational costs with processing long video sequences. To address these limitations, we introduce a novel SpikeMba: multi-modal spiking saliency mamba for temporal video grounding. In our work, we integrate the Spiking Neural Networks (SNNs) and state space models (SSMs) to capture the fine-grained relationships of multimodal features effectively. Specifically, we introduce the relevant slots to enhance the model's memory capabilities, enabling a deeper contextual understanding of video sequences. The contextual moment reasoner leverages these slots to maintain a balance between contextual information preservation and semantic relevance exploration. Simultaneously, the spiking saliency detector capitalizes on the unique properties of SNNs to accurately locate salient proposals. Our experiments demonstrate the effectiveness of SpikeMba, which consistently outperforms state-of-the-art methods across mainstream benchmarks.

link: http://arxiv.org/abs/2404.01174v1

BEM: Balanced and Entropy-based Mix for Long-Tailed Semi-Supervised Learning

Hongwei Zheng, Linyuan Zhou, Han Li, Jinming Su, Xiaoming Wei, Xiaoming Xu

Data mixing methods play a crucial role in semi-supervised learning (SSL), but their application is unexplored in long-tailed semi-supervised learning (LTSSL). The primary reason is that the in-batch mixing manner fails to address class imbalance. Furthermore, existing LTSSL methods mainly focus on re-balancing data quantity but ignore class-wise uncertainty, which is also vital for class balance. For instance, some classes with sufficient samples might still exhibit high uncertainty due to indistinguishable features. To this end, this paper introduces the Balanced and Entropy-based Mix (BEM), a pioneering mixing approach to re-balance the class distribution of both data quantity and uncertainty. Specifically, we first propose a class balanced mix bank to store data of each class for mixing. This bank samples data based on the estimated quantity distribution, thus re-balancing data quantity. Then, we present an entropy-based learning approach to re-balance class-wise uncertainty, including entropy-based sampling strategy, entropy-based selection module, and entropy-based class balanced loss. Our BEM first leverages data mixing for improving LTSSL, and it can also serve as a complement to the existing re-balancing methods. Experimental results show that BEM significantly enhances various LTSSL frameworks and achieves state-of-the-art performances across multiple benchmarks.

link: http://arxiv.org/abs/2404.01179v1

A Neuro-Symbolic Approach to Monitoring Salt Content in Food

Anuja Tayal, Barbara Di Eugenio, Devika Salunke, Andrew D. Boyd, Carolyn A Dickens, Eulalia P Abril, Olga Garcia-Bedoya, Paula G Allen-Meares

We propose a dialogue system that enables heart failure patients to inquire about salt content in foods and help them monitor and reduce salt intake. Addressing the lack of specific datasets for food-based salt content inquiries, we develop a template-based conversational dataset. The dataset is structured to ask clarification questions to identify food items and their salt content. Our findings indicate that while fine-tuning transformer-based models on the dataset yields limited performance, the integration of Neuro-Symbolic Rules significantly enhances the system's performance. Our experiments show that by integrating neuro-symbolic rules, our system achieves an improvement in joint goal accuracy of over 20% across different data sizes compared to naively fine-tuning transformer-based models.

link: http://arxiv.org/abs/2404.01182v1

Efficient Motion Planning for Manipulators with Control Barrier Function-Induced Neural Controller

Mingxin Yu, Chenning Yu, M-Mahdi Naddaf-Sh, Devesh Upadhyay, Sicun Gao, Chuchu Fan

Sampling-based motion planning methods for manipulators in crowded environments often suffer from expensive collision checking and high sampling complexity, which make them difficult to use in real time. To address this issue, we propose a new generalizable control barrier function (CBF)-based steering controller to reduce the number of samples needed in a sampling-based

motion planner RRT. Our method combines the strength of CBF for real-time collision-avoidance control and RRT for long-horizon motion planning, by using CBF-induced neural controller (CBF-INC) to generate control signals that steer the system towards sampled configurations by RRT. CBF-INC is learned as Neural Networks and has two variants handling different inputs, respectively: state (signed distance) input and point-cloud input from LiDAR. In the latter case, we also study two different settings: fully and partially observed environmental information. Compared to manually crafted CBF which suffers from over-approximating robot geometry, CBF-INC can balance safety and goal-reaching better without being over-conservative. Given state-based input, our neural CBF-induced neural controller-enhanced RRT (CBF-INC-RRT) can increase the success rate by 14% while reducing the number of nodes explored by 30%, compared with vanilla RRT on hard test cases. Given LiDAR input where vanilla RRT is not directly applicable, we demonstrate that our CBF-INC-RRT can improve the success rate by 10%, compared with planning with other steering controllers. Our project page with supplementary material is at https://mit-realm.github.io/CBF-INC-RRT-website/.

link: http://arxiv.org/abs/2404.01184v1

MonoBox: Tightness-free Box-supervised Polyp Segmentation using Monotonicity Constraint

Qiang Hu, Zhenyu Yi, Ying Zhou, Ting Li, Fan Huang, Mei Liu, Qiang Li, Zhiwei Wang

We propose MonoBox, an innovative box-supervised segmentation method constrained by monotonicity to liberate its training from the user-unfriendly box-tightness assumption. In contrast to conventional box-supervised segmentation, where the box edges must precisely touch the target boundaries, MonoBox leverages imprecisely-annotated boxes to achieve robust pixel-wise segmentation. The 'linchpin' is that, within the noisy zones around box edges, MonoBox discards the traditional misguiding multiple-instance learning loss, and instead optimizes a carefully-designed objective, termed monotonicity constraint. Along directions transitioning from the foreground to background, this new constraint steers responses to adhere to a trend of monotonically decreasing values. Consequently, the originally unreliable learning within the noisy zones is transformed into a correct and effective monotonicity optimization. Moreover, an adaptive label correction is introduced, enabling MonoBox to enhance the tightness of box annotations using predicted masks from the previous epoch and dynamically shrink the noisy zones as training progresses. We verify MonoBox in the box-supervised segmentation task of polyps, where satisfying box-tightness is challenging due to the vague boundaries between the polyp and normal tissues. Experiments on both public synthetic and in-house real noisy datasets demonstrate that MonoBox exceeds other anti-noise state-of-the-arts by improving Dice by at least 5.5% and 3.3%, respectively. Codes are at https://github.com/Huster-Hq/MonoBox.

link: http://arxiv.org/abs/2404.01188v2

Generating Faithful and Complete Hospital-Course Summaries from the Electronic Health Record

Griffin Adams

The rapid adoption of Electronic Health Records (EHRs) has been instrumental in streamlining administrative tasks, increasing transparency, and enabling continuity of care across providers. An unintended consequence of the increased documentation burden, however, has been reduced face-time with patients and, concomitantly, a dramatic rise in clinician burnout. In this thesis, we pinpoint a particularly time-intensive, yet critical, documentation task: generating a summary of a patient's hospital admissions, and propose and evaluate automated solutions. In Chapter 2, we construct a dataset based on 109,000 hospitalizations (2M source notes) and perform exploratory analyses to motivate future work on modeling and evaluation [NAACL 2021]. In Chapter 3, we address faithfulness from a modeling perspective by revising noisy references [EMNLP 2022] and, to reduce the reliance on references, directly calibrating model outputs to metrics [ACL 2023]. These works relied heavily on automatic metrics as human annotations were limited. To fill this gap, in Chapter 4, we conduct a fine-grained expert annotation of system errors in order to meta-evaluate existing metrics and better understand task-specific issues of domain adaptation and

source-summary alignments. To learn a metric less correlated to extractiveness (copy-and-paste), we derive noisy faithfulness labels from an ensemble of existing metrics and train a faithfulness classifier on these pseudo labels [MLHC 2023]. Finally, in Chapter 5, we demonstrate that fine-tuned LLMs (Mistral and Zephyr) are highly prone to entity hallucinations and cover fewer salient entities. We improve both coverage and faithfulness by performing sentence-level entity planning based on a set of pre-computed salient entities from the source text, which extends our work on entity-guided news summarization [ACL, 2023], [EMNLP, 2023].

link: http://arxiv.org/abs/2404.01189v1

iMD4GC: Incomplete Multimodal Data Integration to Advance Precise Treatment Response Prediction and Survival Analysis for Gastric Cancer

Fengtao Zhou, Yingxue Xu, Yanfen Cui, Shenyan Zhang, Yun Zhu, Weiyang He, Jiguang Wang, Xin Wang, Ronald Chan, Louis Ho Shing Lau, Chu Han, Dafu Zhang, Zhenhui Li, Hao Chen

Gastric cancer (GC) is a prevalent malignancy worldwide, ranking as the fifth most common cancer with over 1 million new cases and 700 thousand deaths in 2020. Locally advanced gastric cancer (LAGC) accounts for approximately two-thirds of GC diagnoses, and neoadjuvant chemotherapy (NACT) has emerged as the standard treatment for LAGC. However, the effectiveness of NACT varies significantly among patients, with a considerable subset displaying treatment resistance. Ineffective NACT not only leads to adverse effects but also misses the optimal therapeutic window, resulting in lower survival rate. However, existing multimodal learning methods assume the availability of all modalities for each patient, which does not align with the reality of clinical practice. The limited availability of modalities for each patient would cause information loss, adversely affecting predictive accuracy. In this study, we propose an incomplete multimodal data integration framework for GC (iMD4GC) to address the challenges posed by incomplete multimodal data, enabling precise response prediction and survival analysis. Specifically, iMD4GC incorporates unimodal attention layers for each modality to capture intra-modal information. Subsequently, the cross-modal interaction layers explore potential inter-modal interactions and capture complementary information across modalities, thereby enabling information compensation for missing modalities. To evaluate iMD4GC, we collected three multimodal datasets for GC study: GastricRes (698 cases) for response prediction, GastricSur (801 cases) for survival analysis, and TCGA-STAD (400 cases) for survival analysis. The scale of our datasets is significantly larger than previous studies. The iMD4GC achieved impressive performance with an 80.2% AUC on GastricRes, 71.4% C-index on GastricSur, and 66.1% C-index on TCGA-STAD, significantly surpassing other compared methods.

link: http://arxiv.org/abs/2404.01192v1