Sat 2024.04.27

Learning Long-form Video Prior via Generative Pre-Training

Jinheng Xie, Jiajun Feng, Zhaoxu Tian, Kevin Qinghong Lin, Yawen Huang, Xi Xia, Nanxu Gong, Xu Zuo, Jiaqi Yang, Yefeng Zheng, Mike Zheng Shou

Concepts involved in long-form videos such as people, objects, and their interactions, can be viewed as following an implicit prior. They are notably complex and continue to pose challenges to be comprehensively learned. In recent years, generative pre-training (GPT) has exhibited versatile capacities in modeling any kind of text content even visual locations. Can this manner work for learning long-form video prior? Instead of operating on pixel space, it is efficient to employ visual locations like bounding boxes and keypoints to represent key information in videos, which can be simply discretized and then tokenized for consumption by GPT. Due to the scarcity of suitable data, we create a new dataset called \textbf{Storyboard20K} from movies to serve as a representative. It includes synopses, shot-by-shot keyframes, and fine-grained annotations of film sets and characters with consistent IDs, bounding boxes, and whole body keypoints. In this way, long-form videos can be represented by a set of tokens and be learned via generative pre-training. Experimental results validate that our approach has great potential for learning long-form video prior. Code and data will be released at \url{https://github.com/showlab/Long-form-Video-Prior}.

link: http://arxiv.org/abs/2404.15909v1

Perception and Localization of Macular Degeneration Applying Convolutional Neural Network, ResNet and Grad-CAM

Tahmim Hossain, Sagor Chandro Bakchy

A well-known retinal disease that feels blurry visions to the affected patients is Macular Degeneration. This research is based on classifying the healthy and macular degeneration fundus with localizing the affected region of the fundus. A CNN architecture and CNN with ResNet architecture (ResNet50, ResNet50v2, ResNet101, ResNet101v2, ResNet152, ResNet152v2) as the backbone are used to classify the two types of fundus. The data are split into three categories including (a) Training set is 90% and Testing set is 10% (b) Training set is 80% and Testing set is 20%, (c) Training set is 50% and Testing set is 50%. After the training, the best model has been selected from the evaluation metrics. Among the models, CNN with backbone of ResNet50 performs best which gives the training accuracy of 98.7\% for 90\% train and 10\% test data split. With this model, we have performed the Grad-CAM visualization to get the region of affected area of fundus.

link: http://arxiv.org/abs/2404.15918v1

An Element-Wise Weights Aggregation Method for Federated Learning

Yi Hu, Hanchi Ren, Chen Hu, Jingjing Deng, Xianghua Xie

Federated learning (FL) is a powerful Machine Learning (ML) paradigm that enables distributed clients to collaboratively learn a shared global model while keeping the data on the original device, thereby preserving privacy. A central challenge in FL is the effective aggregation of local model weights from disparate and potentially unbalanced participating clients. Existing methods often treat each client indiscriminately, applying a single proportion to the entire local model. However, it is empirically advantageous for each weight to be assigned a specific proportion. This paper introduces an innovative Element-Wise Weights Aggregation Method for Federated Learning (EWWA-FL) aimed at optimizing learning performance and accelerating convergence speed. Unlike traditional FL approaches, EWWA-FL aggregates local weights to the global model at the level of individual elements, thereby allowing each participating client to make element-wise contributions to the learning process. By taking into account the unique dataset characteristics of each client, EWWA-FL enhances the robustness of the global model to different datasets while also achieving rapid convergence. The method is flexible enough to employ various weighting strategies. Through comprehensive experiments, we demonstrate the advanced capabilities of EWWA-FL, showing significant improvements in both accuracy and convergence speed across a range of backbones and benchmarks.

link: http://dx.doi.org/10.1109/ICDMW60847.2023.00031

KGValidator: A Framework for Automatic Validation of Knowledge Graph Construction

Jack Boylan, Shashank Mangla, Dominic Thorn, Demian Gholipour Ghalandari, Parsa Ghaffari, Chris Hokamp

This study explores the use of Large Language Models (LLMs) for automatic evaluation of knowledge graph (KG) completion models. Historically, validating information in KGs has been a challenging task, requiring large-scale human annotation at prohibitive cost. With the emergence of general-purpose generative AI and LLMs, it is now plausible that human-in-the-loop validation could be replaced by a generative agent. We introduce a framework for consistency and validation when using generative models to validate knowledge graphs. Our framework is based upon recent open-source developments for structural and semantic validation of LLM outputs, and upon flexible approaches to fact checking and verification, supported by the capacity to reference external knowledge sources of any kind. The design is easy to adapt and extend, and can be used to verify any kind of graph-structured data through a combination of model-intrinsic knowledge, user-supplied context, and agents capable of external knowledge retrieval.

link: http://arxiv.org/abs/2404.15923v1

Inside the echo chamber: Linguistic underpinnings of misinformation on Twitter

Xinyu Wang, Jiayi Li, Sarah Rajtmajer

Social media users drive the spread of misinformation online by sharing posts that include erroneous information or commenting on controversial topics with unsubstantiated arguments often in earnest. Work on echo chambers has suggested that users' perspectives are reinforced through repeated interactions with like-minded peers, promoted by homophily and bias in information diffusion. Building on long-standing interest in the social bases of language and linguistic underpinnings of social behavior, this work explores how conversations around misinformation are mediated through language use. We compare a number of linguistic measures, e.g., in-/out-group cues, readability, and discourse connectives, within and across topics of conversation and user communities. Our findings reveal increased presence of group identity signals and processing fluency within echo chambers during discussions of misinformation. We discuss the specific character of these broader trends across topics and examine contextual influences.

link: http://arxiv.org/abs/2404.15925v1

Generalization Measures for Zero-Shot Cross-Lingual Transfer

Saksham Bassi, Duygu Ataman, Kyunghyun Cho

A model's capacity to generalize its knowledge to interpret unseen inputs with different characteristics is crucial to build robust and reliable machine learning systems. Language model evaluation tasks lack information metrics about model generalization and their applicability in a new setting is measured using task and language-specific downstream performance, which is often lacking in many languages and tasks. In this paper, we explore a set of efficient and reliable measures that could aid in computing more information related to the generalization capability of language models in cross-lingual zero-shot settings. In addition to traditional measures such as variance in parameters after training and distance from initialization, we also measure the effectiveness of sharpness in loss landscape in capturing the success in cross-lingual transfer and propose a novel and stable algorithm to reliably compute the sharpness of a model optimum that correlates to generalization.

link: http://arxiv.org/abs/2404.15928v1

Decentralized Personalized Federated Learning based on a Conditional Sparse-to-Sparser Scheme

Qianyu Long, Qiyuan Wang, Christos Anagnostopoulos, Daning Bi

Decentralized Federated Learning (DFL) has become popular due to its robustness and avoidance of centralized coordination. In this paradigm, clients actively engage in training by exchanging models with their networked neighbors. However, DFL introduces increased costs in terms of training and communication. Existing methods focus on minimizing communication often overlooking training efficiency and data heterogeneity. To address this gap, we propose a novel \textit{sparse-to-sparser} training scheme: DA-DPFL. DA-DPFL initializes with a subset of model parameters, which progressively reduces during training via \textit{dynamic aggregation} and leads to substantial energy savings while retaining adequate information during critical learning periods. Our experiments showcase that DA-DPFL substantially outperforms DFL baselines in test accuracy, while achieving up to 5 times reduction in energy costs. We provide a theoretical analysis of DA-DPFL's convergence by solidifying its applicability in decentralized and personalized learning. The code is available at:https://github.com/EricLoong/da-dpfl

link: http://arxiv.org/abs/2404.15943v2

Mammo-CLIP: Leveraging Contrastive Language-Image Pre-training (CLIP) for Enhanced Breast Cancer Diagnosis with Multi-view Mammography

Xuxin Chen, Yuheng Li, Mingzhe Hu, Ella Salari, Xiaoqian Chen, Richard L. J. Qiu, Bin Zheng, Xiaofeng Yang

Although fusion of information from multiple views of mammograms plays an important role to increase accuracy of breast cancer detection, developing multi-view mammograms-based computer-aided diagnosis (CAD) schemes still faces challenges and no such CAD schemes have been used in clinical practice. To overcome the challenges, we investigate a new approach based on Contrastive Language-Image Pre-training (CLIP), which has sparked interest across various medical imaging tasks. By solving the challenges in (1) effectively adapting the single-view CLIP for multi-view feature fusion and (2) efficiently fine-tuning this parameter-dense model with limited samples and computational resources, we introduce Mammo-CLIP, the first multi-modal framework to process multi-view mammograms and corresponding simple texts. Mammo-CLIP uses an early feature fusion strategy to learn multi-view relationships in four mammograms acquired from the CC and MLO views of the left and right breasts. To enhance learning efficiency, plug-and-play adapters are added into CLIP image and text encoders for fine-tuning parameters and limiting updates to about 1% of the parameters. For framework evaluation, we assembled two datasets retrospectively. The first dataset, comprising 470 malignant and 479 benign cases, was used for few-shot fine-tuning and internal evaluation of the proposed Mammo-CLIP via 5-fold cross-validation. The second dataset, including 60 malignant and 294 benign cases, was used to test generalizability of Mammo-CLIP. Study results show that Mammo-CLIP outperforms the state-of-art cross-view transformer in AUC (0.841 vs. 0.817, 0.837 vs. 0.807) on both datasets. It also surpasses previous two CLIP-based methods by 20.3% and 14.3%. This study highlights the potential of applying the finetuned vision-language models for developing next-generation, image-text-based CAD schemes of breast cancer.

link: http://arxiv.org/abs/2404.15946v1

Sequence can Secretly Tell You What to Discard

Jincheng Dai, Zhuowei Huang, Haiyun Jiang, Chen Chen, Deng Cai, Wei Bi, Shuming Shi

Large Language Models (LLMs), despite their impressive performance on a wide range of tasks, require significant GPU memory and consume substantial computational resources. In addition to model weights, the memory occupied by KV cache increases linearly with sequence length, becoming a main bottleneck for inference. In this paper, we introduce a novel approach for optimizing the KV cache which significantly reduces its memory footprint. Through a comprehensive investigation, we find that on LLaMA2 series models, (i) the similarity between adjacent tokens' query vectors is remarkably high, and (ii) current query's attention calculation can rely solely on the attention information of a small portion of the preceding queries. Based on these observations, we propose CORM, a KV cache eviction policy that dynamically retains important key-value pairs for inference without finetuning the model. We validate that CORM reduces the inference memory usage of KV cache by up to 70% without noticeable performance degradation across six tasks in LongBench.

link: http://arxiv.org/abs/2404.15949v1

Mixed Supervised Graph Contrastive Learning for Recommendation

Weizhi Zhang, Liangwei Yang, Zihe Song, Henry Peng Zou, Ke Xu, Yuanjie Zhu, Philip S. Yu

Recommender systems (RecSys) play a vital role in online platforms, offering users personalized suggestions amidst vast information. Graph contrastive learning aims to learn from high-order collaborative filtering signals with unsupervised augmentation on the user-item bipartite graph, which predominantly relies on the multi-task learning framework involving both the pair-wise recommendation loss and the contrastive loss. This decoupled design can cause inconsistent optimization direction from different losses, which leads to longer convergence time and even sub-optimal performance. Besides, the self-supervised contrastive loss falls short in alleviating the data sparsity issue in RecSys as it learns to differentiate users/items from different views without providing extra supervised collaborative filtering signals during augmentations. In this paper, we propose Mixed Supervised Graph Contrastive Learning for Recommendation (MixSGCL) to address these concerns. MixSGCL originally integrates the training of recommendation and unsupervised contrastive losses into a supervised contrastive learning loss to align the two tasks within one optimization direction. To cope with the data sparsity issue, instead unsupervised augmentation, we further propose node-wise and edge-wise mixup to mine more direct supervised collaborative filtering signals based on existing user-item interactions. Extensive experiments on three real-world datasets demonstrate that MixSGCL surpasses state-of-the-art methods, achieving top performance on both accuracy and efficiency. It validates the effectiveness of MixSGCL with our coupled design on supervised graph contrastive learning.

link: http://arxiv.org/abs/2404.15954v1

Beyond Deepfake Images: Detecting AI-Generated Videos

Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, Matthew C. Stamm

Recent advances in generative AI have led to the development of techniques to generate visually realistic synthetic video. While a number of techniques have been developed to detect AI-generated synthetic images, in this paper we show that synthetic image detectors are unable to detect synthetic videos. We demonstrate that this is because synthetic video generators introduce substantially different traces than those left by image generators. Despite this, we show that synthetic video traces can be learned, and used to perform reliable synthetic video detection or generator source attribution even after H.264 re-compression. Furthermore, we demonstrate that while detecting videos from new generators through zero-shot transferability is challenging, accurate detection of videos from a new generator can be achieved through few-shot learning.

link: http://arxiv.org/abs/2404.15955v1

A Survey on Visual Mamba

Hanwei Zhang, Ying Zhu, Dan Wang, Lijun Zhang, Tianxiang Chen, Zi Ye

State space models (SSMs) with selection mechanisms and hardware-aware architectures, namely Mamba, have recently demonstrated significant promise in long-sequence modeling. Since the self-attention mechanism in transformers has quadratic complexity with image size and increasing computational demands, the researchers are now exploring how to adapt Mamba for computer vision tasks. This paper is the first comprehensive survey aiming to provide an in-depth analysis of Mamba models in the field of computer vision. It begins by exploring the foundational concepts contributing to Mamba's success, including the state space model framework, selection mechanisms, and hardware-aware design. Next, we review these vision mamba models by categorizing them into foundational ones and enhancing them with techniques such as convolution, recurrence, and attention to improve their sophistication. We further delve into the widespread applications of Mamba in vision tasks, which include their use as a backbone in various levels of vision processing. This encompasses general visual tasks, Medical visual tasks (e.g., 2D / 3D segmentation, classification, and image registration, etc.), and Remote Sensing visual tasks. We specially introduce general visual tasks from two levels: High/Mid-level vision (e.g., Object detection, Segmentation, Video classification, etc.) and Low-level vision (e.g., Image super-resolution, Image restoration, Visual generation, etc.). We hope this endeavor will spark additional interest within the community to address current challenges and further apply Mamba models in computer vision.

link: http://arxiv.org/abs/2404.15956v1

Soil analysis with machine-learning-based processing of stepped-frequency GPR field measurements: Preliminary study

Chunlei Xu, Michael Pregesbauer, Naga Sravani Chilukuri, Daniel Windhager, Mahsa Yousefi, Pedro Julian, Lothar Ratschbacher

Ground Penetrating Radar (GPR) has been widely studied as a tool for extracting soil parameters relevant to agriculture and horticulture. When combined with Machine-Learning-based (ML) methods, high-resolution Stepped Frequency Countinuous Wave Radar (SFCW) measurements hold the promise to give cost effective access to depth resolved soil parameters, including at root-level depth. In a first step in this direction, we perform an extensive field survey with a tractor mounted SFCW GPR instrument. Using ML data processing we test the GPR instrument's capabilities to predict the apparent electrical conductivity (ECaR) as measured by a simultaneously recording Electromagnetic Induction (EMI) instrument. The large-scale field measurement campaign with 3472 co-registered and geo-located GPR and EMI data samples distributed over ~6600 square meters was performed on a golf course. The selected terrain benefits from a high surface homogeneity, but also features the challenge of only small, and hence hard to discern, variations in the measured soil parameter. Based on the quantitative results we suggest the use of nugget-to-sill ratio as a performance metric for the evaluation of end-to-end ML performance in the agricultural setting and discuss the limiting factors in the multi-sensor regression setting. The code is released as open source and available at https://opensource.silicon-austria.com/xuc/soil-analysis-machine-learning-stepped-frequency-gpr.

link: http://arxiv.org/abs/2404.15961v1

Interpretable Clustering with the Distinguishability Criterion

Ali Turfah, Xiaoquan Wen

Cluster analysis is a popular unsupervised learning tool used in many disciplines to identify heterogeneous sub-populations within a sample. However, validating cluster analysis results and determining the number of clusters in a data set remains an outstanding problem. In this work, we present a global criterion called the Distinguishability criterion to quantify the separability of identified clusters and validate inferred cluster configurations. Our computational implementation of the Distinguishability criterion corresponds to the Bayes risk of a randomized classifier under the 0-1 loss. We propose a combined loss function-based computational framework that integrates the Distinguishability criterion with many commonly used clustering procedures, such as hierarchical clustering, k-means, and finite mixture models. We present these new algorithms as well as the results from comprehensive data analysis based on simulation studies and real data applications.

link: http://arxiv.org/abs/2404.15967v2

On the Fourier analysis in the SO(3) space : EquiLoPO Network

Dmitrii Zhemchuzhnikov, Sergei Grudinin

Analyzing volumetric data with rotational invariance or equivariance is an active topic in current research. Existing deep-learning approaches utilize either group convolutional networks limited to discrete rotations or steerable convolutional networks with constrained filter structures. This work proposes a novel equivariant neural network architecture that achieves analytical Equivariance to Local Pattern Orientation on the continuous SO(3) group while allowing unconstrained trainable filters - EquiLoPO Network. Our key innovations are a group convolutional operation leveraging irreducible representations as the Fourier basis and a local activation function in the SO(3) space that provides a well-defined mapping from input to output functions, preserving equivariance. By integrating these operations into a ResNet-style architecture, we propose a model that overcomes the limitations of prior methods. A comprehensive evaluation on diverse 3D medical imaging datasets from MedMNIST3D demonstrates the effectiveness of our approach, which consistently outperforms state of the art. This work suggests the benefits of true rotational equivariance on SO(3) and flexible unconstrained filters enabled by the local activation function, providing a flexible framework for equivariant deep learning on volumetric data with potential applications across domains. Our code is publicly available at \url{https://gricad-gitlab.univ-grenoble-alpes.fr/GruLab/ILPO/-/tree/main/EquiLoPO}.

link: http://arxiv.org/abs/2404.15979v1