Sat 2024.04.13

Implicit Multi-Spectral Transformer: An Lightweight and Effective Visible to Infrared Image Translation Model

Yijia Chen, Pinghua Chen, Xiangxin Zhou, Yingtie Lei, Ziyang Zhou, Mingxian Li

In the field of computer vision, visible light images often exhibit low contrast in low-light conditions, presenting a significant challenge. While infrared imagery provides a potential solution, its utilization entails high costs and practical limitations. Recent advancements in deep learning, particularly the deployment of Generative Adversarial Networks (GANs), have facilitated the transformation of visible light images to infrared images. However, these methods often experience unstable training phases and may produce suboptimal outputs. To address these issues, we propose a novel end-to-end Transformer-based model that efficiently converts visible light images into high-fidelity infrared images. Initially, the Texture Mapping Module and Color Perception Adapter collaborate to extract texture and color features from the visible light image. The Dynamic Fusion Aggregation Module subsequently integrates these features. Finally, the transformation into an infrared image is refined through the synergistic action of the Color Perception Adapter and the Enhanced Perception Attention mechanism. Comprehensive benchmarking experiments confirm that our model outperforms existing methods, producing infrared images of markedly superior quality, both qualitatively and quantitatively. Furthermore, the proposed model enables more effective downstream applications for infrared images than other methods.

link: http://arxiv.org/abs/2404.07072v1

VLLMs Provide Better Context for Emotion Understanding Through Common Sense Reasoning

Alexandros Xenos, Niki Maria Foteinopoulou, Ioanna Ntinou, Ioannis Patras, Georgios Tzimiropoulos

Recognising emotions in context involves identifying the apparent emotions of an individual, taking into account contextual cues from the surrounding scene. Previous approaches to this task have involved the design of explicit scene-encoding architectures or the incorporation of external scene-related information, such as captions. However, these methods often utilise limited contextual information or rely on intricate training pipelines. In this work, we leverage the groundbreaking capabilities of Vision-and-Large-Language Models (VLLMs) to enhance in-context emotion classification without introducing complexity to the training process in a two-stage approach. In the first stage, we propose prompting VLLMs to generate descriptions in natural language of the subject's apparent emotion relative to the visual context. In the second stage, the descriptions are used as contextual information and, along with the image input, are used to train a transformer-based architecture that fuses text and visual features before the final classification task. Our experimental results show that the text and image features have complementary information, and our fused architecture significantly outperforms the individual modalities without any complex training methods. We evaluate our approach on three different datasets, namely, EMOTIC, CAER-S, and BoLD, and achieve state-of-the-art or comparable accuracy across all datasets and metrics compared to much more complex approaches. The code will be made publicly available on github: https://github.com/NickyFot/EmoCommonSense.git

link: http://arxiv.org/abs/2404.07078v1

Minimizing Chebyshev Prototype Risk Magically Mitigates the Perils of Overfitting

Nathaniel Dean, Dilip Sarkar

Overparameterized deep neural networks (DNNs), if not sufficiently regularized, are susceptible to overfitting their training examples and not generalizing well to test data. To discourage overfitting, researchers have developed multicomponent loss functions that reduce intra-class feature correlation and maximize inter-class feature distance in one or more layers of the network. By analyzing the penultimate feature layer activations output by a DNN's feature extraction section prior to the linear classifier, we find that modified forms of the intra-class feature covariance and inter-class prototype separation are key components of a fundamental Chebyshev upper bound on the probability of misclassification, which we designate the Chebyshev Prototype Risk (CPR). While previous approaches' covariance loss terms scale quadratically with the number of network features, our CPR bound indicates that an approximate covariance loss in log-linear time is sufficient to reduce the bound and is scalable to large architectures. We implement the terms of the CPR bound into our Explicit CPR (exCPR) loss function and observe from empirical results on multiple datasets and network architectures that our training algorithm reduces overfitting and improves upon previous approaches in many settings. Our code is available at https://github.com/Deano1718/Regularization_exCPR .

link: http://arxiv.org/abs/2404.07083v2

Dynamic Generation of Personalities with Large Language Models

Jianzhi Liu, Hexiang Gu, Tianyu Zheng, Liuyu Xiang, Huijia Wu, Jie Fu, Zhaofeng He

In the realm of mimicking human deliberation, large language models (LLMs) show promising performance, thereby amplifying the importance of this research area. Deliberation is influenced by both logic and personality. However, previous studies predominantly focused on the logic of LLMs, neglecting the exploration of personality aspects. In this work, we introduce Dynamic Personality Generation (DPG), a dynamic personality generation method based on Hypernetworks. Initially, we embed the Big Five personality theory into GPT-4 to form a personality assessment machine, enabling it to evaluate characters' personality traits from dialogues automatically. We propose a new metric to assess personality generation capability based on this evaluation method. Then, we use this personality assessment machine to evaluate dialogues in script data, resulting in a personality-dialogue dataset. Finally, we fine-tune DPG on the personality-dialogue dataset. Experiments prove that DPG's personality generation capability is stronger after fine-tuning on this dataset than traditional fine-tuning methods, surpassing prompt-based GPT-4.

link: http://arxiv.org/abs/2404.07084v1

LaTiM: Longitudinal representation learning in continuous-time models to predict disease progression

Rachid Zeghlache, Pierre-Henri Conze, Mostafa El Habib Daho, Yihao Li, Hugo Le Boité, Ramin Tadayoni, Pascal Massin, Béatrice Cochener, Alireza Rezaei, Ikram Brahim, Gwenolé Quellec, Mathieu Lamard

This work proposes a novel framework for analyzing disease progression using time-aware neural ordinary differential equations (NODE). We introduce a "time-aware head" in a framework trained through self-supervised learning (SSL) to leverage temporal information in latent space for data augmentation. This approach effectively integrates NODEs with SSL, offering significant performance improvements compared to traditional methods that lack explicit temporal integration. We demonstrate the effectiveness of our strategy for diabetic retinopathy progression prediction using the OPHDIAT database. Compared to the baseline, all NODE architectures achieve statistically significant improvements in area under the ROC curve (AUC) and Kappa metrics, highlighting the efficacy of pre-training with SSL-inspired approaches. Additionally, our framework promotes stable training for NODEs, a commonly encountered challenge in time-aware modeling.

link: http://arxiv.org/abs/2404.07091v1

MoCap-to-Visual Domain Adaptation for Efficient Human Mesh Estimation from 2D Keypoints

Bedirhan Uguz, Ozhan Suat, Batuhan Karagoz, Emre Akbas

This paper presents Key2Mesh, a model that takes a set of 2D human pose keypoints as input and estimates the corresponding body mesh. Since this process does not involve any visual (i.e. RGB image) data, the model can be trained on large-scale motion capture (MoCap) datasets, thereby overcoming the scarcity of image datasets with 3D labels. To enable the model's application on RGB images, we first run an off-the-shelf 2D pose estimator to obtain the 2D keypoints, and then feed these 2D keypoints to Key2Mesh. To improve the performance of our model on RGB images, we apply an adversarial domain adaptation (DA) method to bridge the gap between the MoCap and visual domains. Crucially, our DA method does not require 3D labels for visual data, which enables adaptation to target sets without the need for costly labels. We evaluate Key2Mesh for the task of estimating 3D human meshes from 2D keypoints, in the absence of RGB and mesh label pairs. Our results on widely used H3.6M and 3DPW datasets show that Key2Mesh sets the new state-of-the-art by outperforming other models in PA-MPJPE for both datasets, and in MPJPE and PVE for the 3DPW dataset. Thanks to our model's simple architecture, it operates at least 12x faster than the prior state-of-the-art model, LGD. Additional qualitative samples and code are available on the project website: https://key2mesh.github.io/.

link: http://arxiv.org/abs/2404.07094v1

TransTARec: Time-Adaptive Translating Embedding Model for Next POI Recommendation

Yiping Sun

The rapid growth of location acquisition technologies makes Point-of-Interest(POI) recommendation possible due to redundant user check-in records. In this paper, we focus on next POI recommendation in which next POI is based on previous POI. We observe that time plays an important role in next POI recommendation but is neglected in the recent proposed translating embedding methods. To tackle this shortage, we propose a time-adaptive translating embedding model (TransTARec) for next POI recommendation that naturally incorporates temporal influence, sequential dynamics, and user preference within a single component. Methodologically, we treat a (previous timestamp, user, next timestamp) triplet as a union translation vector and develop a neural-based fusion operation to fuse user preference and temporal influence. The superiority of TransTARec, which is confirmed by extensive experiments on real-world datasets, comes from not only the introduction of temporal influence but also the direct unification with user preference and sequential dynamics.

link: http://arxiv.org/abs/2404.07096v1

Learning Priors for Non Rigid SfM from Casual Videos

Yoni Kasten, Wuyue Lu, Haggai Maron

We tackle the long-standing challenge of reconstructing 3D structures and camera positions from videos. The problem is particularly hard when objects are transformed in a non-rigid way. Current approaches to this problem make unrealistic assumptions or require a long optimization time. We present TracksTo4D, a novel deep learning-based approach that enables inferring 3D structure and camera positions from dynamic content originating from in-the-wild videos using a single feed-forward pass on a sparse point track matrix. To achieve this, we leverage recent advances in 2D point tracking and design an equivariant neural architecture tailored for directly processing 2D point tracks by leveraging their symmetries. TracksTo4D is trained on a dataset of in-the-wild videos utilizing only the 2D point tracks extracted from the videos, without any 3D supervision. Our experiments demonstrate that TracksTo4D generalizes well to unseen videos of unseen semantic categories at inference time, producing equivalent results to state-of-the-art methods while significantly reducing the runtime compared to other baselines.

link: http://arxiv.org/abs/2404.07097v1

Rethinking Out-of-Distribution Detection for Reinforcement Learning: Advancing Methods for Evaluation and Detection

Linas Nasvytis, Kai Sandbrink, Jakob Foerster, Tim Franzmeyer, Christian Schroeder de Witt

While reinforcement learning (RL) algorithms have been successfully applied across numerous sequential decision-making problems, their generalization to unforeseen testing environments remains a significant concern. In this paper, we study the problem of out-of-distribution (OOD) detection in RL, which focuses on identifying situations at test time that RL agents have not encountered in their training environments. We first propose a clarification of terminology for OOD detection in RL, which aligns it with the literature from other machine learning domains. We then present new benchmark scenarios for OOD detection, which introduce anomalies with temporal autocorrelation into different components of the agent-environment loop. We argue that such scenarios have been understudied in the current literature, despite their relevance to real-world situations. Confirming our theoretical predictions, our experimental results suggest that state-of-the-art OOD detectors are not able to identify such anomalies. To address this problem, we propose a novel method for OOD detection, which we call DEXTER (Detection via Extraction of Time Series Representations). By treating environment observations as time series data, DEXTER extracts salient time series features, and then leverages an ensemble of isolation forest algorithms to detect anomalies. We find that DEXTER can reliably identify anomalies across benchmark scenarios, exhibiting superior performance compared to both state-of-the-art OOD detectors and high-dimensional changepoint detectors adopted from statistics.

link: http://arxiv.org/abs/2404.07099v1

Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Suhang Wang, Yu Meng, Jiawei Han

Large language models (LLMs), while exhibiting exceptional performance, suffer from hallucinations, especially on knowledge-intensive tasks. Existing works propose to augment LLMs with individual text units retrieved from external knowledge corpora to alleviate the issue. However, in many domains, texts are interconnected (e.g., academic papers in a bibliographic graph are linked by citations and co-authorships) which form a (text-attributed) graph. The knowledge in such graphs is encoded not only in single texts/nodes but also in their associated connections. To facilitate the research of augmenting LLMs with graphs, we manually construct a Graph Reasoning Benchmark dataset called GRBench, containing 1,740 questions that can be answered with the knowledge from 10 domain graphs. Then, we propose a simple and effective framework called Graph Chain-of-thought (Graph-CoT) to augment LLMs with graphs by encouraging LLMs to reason on the graph iteratively. Each Graph-CoT iteration consists of three sub-steps: LLM reasoning, LLM-graph interaction, and graph execution. We conduct systematic experiments with three LLM backbones on GRBench, where Graph-CoT outperforms the baselines consistently. The code is available at https://github.com/PeterGriffinJin/Graph-CoT.

link: http://arxiv.org/abs/2404.07103v1

3DMambaComplete: Exploring Structured State Space Model for Point Cloud Completion

Yixuan Li, Weidong Yang, Ben Fei

Point cloud completion aims to generate a complete and high-fidelity point cloud from an initially incomplete and low-quality input. A prevalent strategy involves leveraging Transformer-based models to encode global features and facilitate the reconstruction process. However, the adoption of pooling operations to obtain global feature representations often results in the loss of local details within the point cloud. Moreover, the attention mechanism inherent in Transformers introduces additional computational complexity, rendering it challenging to handle long sequences effectively. To address these issues, we propose 3DMambaComplete, a point cloud completion network built on the novel Mamba framework. It comprises three modules: HyperPoint Generation encodes point cloud features using Mamba's selection mechanism and predicts a set of Hyperpoints. A specific offset is estimated, and the down-sampled points become HyperPoints. The HyperPoint Spread module disperses these HyperPoints across different spatial locations to avoid concentration. Finally, a deformation method transforms the 2D mesh representation of HyperPoints into a fine-grained 3D structure for point cloud reconstruction. Extensive experiments conducted on various established benchmarks demonstrate that 3DMambaComplete surpasses state-of-the-art point cloud completion methods, as confirmed by qualitative and quantitative analyses.

link: http://arxiv.org/abs/2404.07106v1

From Model-centered to Human-Centered: Revision Distance as a Metric for Text Evaluation in LLMs-based Applications

Yongqiang Ma, Lizhi Qing, Jiawei Liu, Yangyang Kang, Yue Zhang, Wei Lu, Xiaozhong Liu, Qikai Cheng

Evaluating large language models (LLMs) is fundamental, particularly in the context of practical applications. Conventional evaluation methods, typically designed primarily for LLM development, yield numerical scores that ignore the user experience. Therefore, our study shifts the focus from model-centered to human-centered evaluation in the context of AI-powered writing assistance applications. Our proposed metric, termed ``Revision Distance,'' utilizes LLMs to suggest revision edits that mimic the human writing process. It is determined by counting the revision edits generated by LLMs. Benefiting from the generated revision edit details, our metric can provide a self-explained text evaluation result in a human-understandable manner beyond the context-independent score. Our results show that for the easy-writing task, ``Revision Distance'' is consistent with established metrics (ROUGE, Bert-score, and GPT-score), but offers more insightful, detailed feedback and better distinguishes between texts. Moreover, in the context of challenging academic writing tasks, our metric still delivers reliable evaluations where other metrics tend to struggle. Furthermore, our metric also holds significant potential for scenarios lacking reference texts.

link: http://arxiv.org/abs/2404.07108v2

Wild Visual Navigation: Fast Traversability Learning via Pre-Trained Models and Online Self-Supervision

Matías Mattamala, Jonas Frey, Piotr Libera, Nived Chebrolu, Georg Martius, Cesar Cadena, Marco Hutter, Maurice Fallon

Natural environments such as forests and grasslands are challenging for robotic navigation because of the false perception of rigid obstacles from high grass, twigs, or bushes. In this work, we present Wild Visual Navigation (WVN), an online self-supervised learning system for visual traversability estimation. The system is able to continuously adapt from a short human demonstration in the field, only using onboard sensing and computing. One of the key ideas to achieve this is the use of high-dimensional features from pre-trained self-supervised models, which implicitly encode semantic information that massively simplifies the learning task. Further, the development of an online scheme for supervision generator enables concurrent training and inference of the learned model in the wild. We demonstrate our approach through diverse real-world deployments in forests, parks, and grasslands. Our system is able to bootstrap the traversable terrain segmentation in less than 5 min of in-field training time, enabling the robot to navigate in complex, previously unseen outdoor terrains. Code: https://bit.ly/498b0CV - Project page:https://bit.ly/3M6nMHH

link: http://arxiv.org/abs/2404.07110v1

Unfolding ADMM for Enhanced Subspace Clustering of Hyperspectral Images

Xianlu Li, Nicolas Nadisic, Shaoguang Huang, Aleksandra Pižurica

Deep subspace clustering methods are now prominent in clustering, typically using fully connected networks and a self-representation loss function. However, these methods often struggle with overfitting and lack interpretability. In this paper, we explore an alternative clustering approach based on deep unfolding. By unfolding iterative optimization methods into neural networks, this approach offers enhanced interpretability and reliability compared to data-driven deep learning methods, and greater adaptability and generalization than model-based approaches. Hence, unfolding has become widely used in inverse imaging problems, such as image restoration, reconstruction, and super-resolution, but has not been sufficiently explored yet in the context of clustering. In this work, we introduce an innovative clustering architecture for hyperspectral images (HSI) by unfolding an iterative solver based on the Alternating Direction Method of Multipliers (ADMM) for sparse subspace clustering. To our knowledge, this is the first attempt to apply unfolding ADMM for computing the self-representation matrix in subspace clustering. Moreover, our approach captures well the structural characteristics of HSI data by employing the K nearest neighbors algorithm as part of a structure preservation module. Experimental evaluation of three established HSI datasets shows clearly the potential of the unfolding approach in HSI clustering and even demonstrates superior performance compared to state-of-the-art techniques.

link: http://arxiv.org/abs/2404.07112v1

Continuous Language Model Interpolation for Dynamic and Controllable Text Generation

Sara Kangaslahti, David Alvarez-Melis

As large language models (LLMs) have gained popularity for a variety of use cases, making them adaptable and controllable has become increasingly important, especially for user-facing applications. While the existing literature on LLM adaptation primarily focuses on finding a model (or models) that optimizes a single predefined objective, here we focus on the challenging case where the model must dynamically adapt to diverse -- and often changing -- user preferences. For this, we leverage adaptation methods based on linear weight interpolation, casting them as continuous multi-domain interpolators that produce models with specific prescribed generation characteristics on-the-fly. Specifically, we use low-rank updates to fine-tune a base model to various different domains, yielding a set of anchor models with distinct generation profiles. Then, we use the weight updates of these anchor models to parametrize the entire (infinite) class of models contained within their convex hull. We empirically show that varying the interpolation weights yields predictable and consistent change in the model outputs with respect to all of the controlled attributes. We find that there is little entanglement between most attributes and identify and discuss the pairs of attributes for which this is not the case. Our results suggest that linearly interpolating between the weights of fine-tuned models facilitates predictable, fine-grained control of model outputs with respect to multiple stylistic characteristics simultaneously.

link: http://arxiv.org/abs/2404.07117v1