Wed 2024.05.08

JOSENet: A Joint Stream Embedding Network for Violence Detection in Surveillance Videos

Pietro Nardelli, Danilo Comminiello

Due to the ever-increasing availability of video surveillance cameras and the growing need for crime prevention, the violence detection task is attracting greater attention from the research community. With respect to other action recognition tasks, violence detection in surveillance videos shows additional issues, such as the presence of a significant variety of real fight scenes. Unfortunately, available datasets seem to be very small compared with other action recognition datasets. Moreover, in surveillance applications, people in the scenes always differ for each video and the background of the footage differs for each camera. Also, violent actions in real-life surveillance videos must be detected quickly to prevent unwanted consequences, thus models would definitely benefit from a reduction in memory usage and computational costs. Such problems make classical action recognition methods difficult to be adopted. To tackle all these issues, we introduce JOSENet, a novel self-supervised framework that provides outstanding performance for violence detection in surveillance videos. The proposed model receives two spatiotemporal video streams, i.e., RGB frames and optical flows, and involves a new regularized self-supervised learning approach for videos. JOSENet provides improved performance compared to self-supervised state-of-the-art methods, while requiring one-fourth of the number of frames per video segment and a reduced frame rate. The source code and the instructions to reproduce our experiments are available at https://github.com/ispamm/JOSENet.

link: http://arxiv.org/abs/2405.02961v1

VectorPainter: A Novel Approach to Stylized Vector Graphics Synthesis with Vectorized Strokes

Juncheng Hu, Ximing Xing, Zhengqi Zhang, Jing Zhang, Qian Yu

We propose a novel method, VectorPainter, for the task of stylized vector graphics synthesis. Given a text prompt and a reference style image, VectorPainter generates a vector graphic that aligns in content with the text prompt and remains faithful in style to the reference image. We recognize that the key to this task lies in fully leveraging the intrinsic properties of vector graphics. Innovatively, we conceptualize the stylization process as the rearrangement of vectorized strokes extracted from the reference image. VectorPainter employs an optimization-based pipeline. It begins by extracting vectorized strokes from the reference image, which are then used to initialize the synthesis process. To ensure fidelity to the reference style, a novel style preservation loss is introduced. Extensive experiments have been conducted to demonstrate that our method is capable of aligning with the text description while remaining faithful to the reference image.

link: http://arxiv.org/abs/2405.02962v1

Robust Collaborative Perception without External Localization and Clock Devices

Zixing Lei, Zhenyang Ni, Ruize Han, Shuo Tang, Chen Feng, Siheng Chen, Yanfeng Wang

A consistent spatial-temporal coordination across multiple agents is fundamental for collaborative perception, which seeks to improve perception abilities through information exchange among agents. To achieve this spatial-temporal alignment, traditional methods depend on external devices to provide localization and clock signals. However, hardware-generated signals could be vulnerable to noise and potentially malicious attack, jeopardizing the precision of spatial-temporal alignment. Rather than relying on external hardwares, this work proposes a novel approach: aligning by recognizing the inherent geometric patterns within the perceptual data of various agents. Following this spirit, we propose a robust collaborative perception system that operates independently of external localization and clock devices. The key module of our system,~\emph{FreeAlign}, constructs a salient object graph for each agent based on its detected boxes and uses a graph neural network to identify common subgraphs between agents, leading to accurate relative pose

and time. We validate \emph{FreeAlign} on both real-world and simulated datasets. The results show that, the ~\emph{FreeAlign} empowered robust collaborative perception system perform comparably to systems relying on precise localization and clock devices.

link: http://arxiv.org/abs/2405.02965v1

CoverLib: Classifiers-equipped Experience Library by Iterative Problem Distribution Coverage Maximization for Domain-tuned Motion Planning

Hirokazu Ishida, Naoki Hiraoka, Kei Okada, Masayuki Inaba

Library-based methods are known to be very effective for fast motion planning by adapting an experience retrieved from a precomputed library. This article presents CoverLib, a principled approach for constructing and utilizing such a library. CoverLib iteratively adds an experience-classifier-pair to the library, where each classifier corresponds to an adaptable region of the experience within the problem space. This iterative process is an active procedure, as it selects the next experience based on its ability to effectively cover the uncovered region. During the query phase, these classifiers are utilized to select an experience that is expected to be adaptable for a given problem. Experimental results demonstrate that CoverLib effectively mitigates the trade-off between plannability and speed observed in global (e.g. sampling-based) and local (e.g. optimization-based) methods. As a result, it achieves both fast planning and high success rates over the problem domain. Moreover, due to its adaptation-algorithm-agnostic nature, CoverLib seamlessly integrates with various adaptation methods, including nonlinear programming-based and sampling-based algorithms.

link: http://arxiv.org/abs/2405.02968v2

Towards a Flexible and High-Fidelity Approach to Distributed DNN Training Emulation

Banruo Liu, Mubarak Adetunji Ojewale, Yuhan Ding, Marco Canini

We propose NeuronaBox, a flexible, user-friendly, and high-fidelity approach to emulate DNN training workloads. We argue that to accurately observe performance, it is possible to execute the training workload on a subset of real nodes and emulate the networked execution environment along with the collective communication operations. Initial results from a proof-of-concept implementation show that NeuronaBox replicates the behavior of actual systems with high accuracy, with an error margin of less than 1% between the emulated measurements and the real system.

link: http://arxiv.org/abs/2405.02969v1

Multi-Agent RL-Based Industrial AIGC Service Offloading over Wireless Edge Networks

Siyuan Li, Xi Lin, Hansong Xu, Kun Hua, Xiaomin Jin, Gaolei Li, Jianhua Li

Currently, the generative model has garnered considerable attention due to its application in addressing the challenge of scarcity of abnormal samples in the industrial Internet of Things (IoT). However, challenges persist regarding the edge deployment of generative models and the optimization of joint edge AI-generated content (AIGC) tasks. In this paper, we focus on the edge optimization of AIGC task execution and propose GMEL, a generative model-driven industrial AIGC collaborative edge learning framework. This framework aims to facilitate efficient few-shot learning by leveraging realistic sample synthesis and edge-based optimization capabilities. First, a multi-task AIGC computational offloading model is presented to ensure the efficient execution of heterogeneous AIGC tasks on edge servers. Then, we propose an attention-enhanced multi-agent reinforcement learning (AMARL) algorithm aimed at refining offloading policies within the IoT system, thereby supporting generative model-driven edge learning. Finally, our experimental results demonstrate the effectiveness of the proposed algorithm in optimizing the total system latency of the edge-based AIGC task completion.

link: http://arxiv.org/abs/2405.02972v1

SkelCap: Automated Generation of Descriptive Text from Skeleton Keypoint Sequences

Ali Emre Keskin, Hacer Yalim Keles

Numerous sign language datasets exist, yet they typically cover only a limited selection of the thousands of signs used globally. Moreover, creating diverse sign language datasets is an expensive and challenging task due to the costs associated with gathering a varied group of signers. Motivated by these challenges, we aimed to develop a solution that addresses these limitations. In this context, we focused on textually describing body movements from skeleton keypoint sequences, leading to the creation of a new dataset. We structured this dataset around AUTSL, a comprehensive isolated Turkish sign language dataset. We also developed a baseline model, SkelCap, which can generate textual descriptions of body movements. This model processes the skeleton keypoints data as a vector, applies a fully connected layer for embedding, and utilizes a transformer neural network for sequence-to-sequence modeling. We conducted extensive evaluations of our model, including signer-agnostic and sign-agnostic assessments. The model achieved promising results, with a ROUGE-L score of 0.98 and a BLEU-4 score of 0.94 in the signer-agnostic evaluation. The dataset we have prepared, namely the AUTSL-SkelCap, will be made publicly available soon.

link: http://arxiv.org/abs/2405.02977v1

Self-Organized Construction by Minimal Surprise

Tanja Katharina Kaiser, Heiko Hamann

For the robots to achieve a desired behavior, we can program them directly, train them, or give them an innate driver that makes the robots themselves desire the targeted behavior. With the minimal surprise approach, we implant in our robots the desire to make their world predictable. Here, we apply minimal surprise to collective construction. Simulated robots push blocks in a 2D torus grid world. In two variants of our experiment we either allow for emergent behaviors or predefine the expected environment of the robots. In either way, we evolve robot behaviors that move blocks to structure their environment and make it more predictable. The resulting controllers can be applied in collective construction by robots.

link: http://dx.doi.org/10.1109/FAS-W.2019.00057

Paintings and Drawings Aesthetics Assessment with Rich Attributes for Various Artistic Categories

Xin Jin, Qianqian Qiao, Yi Lu, Shan Gao, Heng Huang, Guangdong Li

Image aesthetic evaluation is a highly prominent research domain in the field of computer vision. In recent years, there has been a proliferation of datasets and corresponding evaluation methodologies for assessing the aesthetic quality of photographic works, leading to the establishment of a relatively mature research environment. However, in contrast to the extensive research in photographic aesthetics, the field of aesthetic evaluation for paintings and Drawings has seen limited attention until the introduction of the BAID dataset in March 2023. This dataset solely comprises overall scores for high-quality artistic images. Our research marks the pioneering introduction of a multi-attribute, multi-category dataset specifically tailored to the field of painting: Aesthetics of Paintings and Drawings Dataset (APDD). The construction of APDD received active participation from 28 professional artists worldwide, along with dozens of students specializing in the field of art. This dataset encompasses 24 distinct artistic categories and 10 different aesthetic attributes. Each image in APDD has been evaluated by six professionally trained experts in the field of art, including assessments for both total aesthetic scores and aesthetic attribute scores. The final APDD dataset comprises a total of 4985 images, with an annotation count exceeding 31100 entries. Concurrently, we propose an innovative approach: Art Assessment Network for Specific Painting Styles (AANSPS), designed for the assessment of aesthetic attributes in mixed-attribute art datasets. Through this research, our goal is to catalyze advancements in the field of aesthetic evaluation for paintings and drawings, while enriching the available resources and methodologies

for its further development and application.

link: http://arxiv.org/abs/2405.02982v1

E-TSL: A Continuous Educational Turkish Sign Language Dataset with Baseline Methods

■ükrü Öztürk, Hacer Yalim Keles

This study introduces the continuous Educational Turkish Sign Language (E-TSL) dataset, collected from online Turkish language lessons for 5th, 6th, and 8th grades. The dataset comprises 1,410 videos totaling nearly 24 hours and includes performances from 11 signers. Turkish, an agglutinative language, poses unique challenges for sign language translation, particularly with a vocabulary where 64% are singleton words and 85% are rare words, appearing less than five times. We developed two baseline models to address these challenges: the Pose to Text Transformer (P2T-T) and the Graph Neural Network based Transformer (GNN-T) models. The GNN-T model achieved 19.13% BLEU-1 score and 3.28% BLEU-4 score, presenting a significant challenge compared to existing benchmarks. The P2T-T model, while demonstrating slightly lower performance in BLEU scores, achieved a higher ROUGE-L score of 22.09%. Additionally, we benchmarked our model using the well-known PHOENIX-Weather 2014T dataset to validate our approach.

link: http://arxiv.org/abs/2405.02984v1

Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability to Mark Short Answer Questions in K-12 Education

Owen Henkel, Adam Boxer, Libby Hills, Bill Roberts

This paper presents reports on a series of experiments with a novel dataset evaluating how well Large Language Models (LLMs) can mark (i.e. grade) open text responses to short answer questions, Specifically, we explore how well different combinations of GPT version and prompt engineering strategies performed at marking real student answers to short answer across different domain areas (Science and History) and grade-levels (spanning ages 5-16) using a new, never-used-before dataset from Carousel, a quizzing platform. We found that GPT-4, with basic few-shot prompting performed well (Kappa, 0.70) and, importantly, very close to human-level performance (0.75). This research builds on prior findings that GPT-4 could reliably score short answer reading comprehension questions at a performance-level very close to that of expert human raters. The proximity to human-level performance, across a variety of subjects and grade levels suggests that LLMs could be a valuable tool for supporting low-stakes formative assessment tasks in K-12 education and has important implications for real-world education delivery.

link: http://arxiv.org/abs/2405.02985v1

RepAugment: Input-Agnostic Representation-Level Augmentation for Respiratory Sound Classification

June-Woo Kim, Miika Toikkanen, Sangmin Bae, Minseok Kim, Ho-Young Jung

Recent advancements in AI have democratized its deployment as a healthcare assistant. While pretrained models from large-scale visual and audio datasets have demonstrably generalized to this task, surprisingly, no studies have explored pretrained speech models, which, as human-originated sounds, intuitively would share closer resemblance to lung sounds. This paper explores the efficacy of pretrained speech models for respiratory sound classification. We find that there is a characterization gap between speech and lung sound samples, and to bridge this gap, data augmentation is essential. However, the most widely used augmentation technique for audio and speech, SpecAugment, requires 2-dimensional spectrogram format and cannot be applied to models pretrained on speech waveforms. To address this, we propose RepAugment, an input-agnostic representation-level augmentation technique that outperforms SpecAugment, but is also suitable for respiratory sound classification with waveform pretrained models. Experimental results show that our approach outperforms the SpecAugment, demonstrating a substantial improvement in the accuracy of minority disease classes, reaching up to 7.14%.

link: http://arxiv.org/abs/2405.02996v1

MedAdapter: Efficient Test-Time Adaptation of Large Language Models towards Medical Reasoning

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Hang Wu, Carl Yang, May D. Wang

Despite their improved capabilities in generation and reasoning, adapting large language models (LLMs) to the biomedical domain remains challenging due to their immense size and corporate privacy. In this work, we propose MedAdapter, a unified post-hoc adapter for test-time adaptation of LLMs towards biomedical applications. Instead of fine-tuning the entire LLM, MedAdapter effectively adapts the original model by fine-tuning only a small BERT-sized adapter to rank candidate solutions generated by LLMs. Experiments demonstrate that MedAdapter effectively adapts both white-box and black-box LLMs in biomedical reasoning, achieving average performance improvements of 25.48% and 11.31%, respectively, without requiring extensive computational resources or sharing data with third parties. MedAdapter also yields superior performance when combined with train-time adaptation, highlighting a flexible and complementary solution to existing adaptation methods. Faced with the challenges of balancing model performance, computational resources, and data privacy, MedAdapter provides an efficient, privacy-preserving, cost-effective, and transparent solution for adapting LLMs to the biomedical domain.

link: http://arxiv.org/abs/2405.03000v1

Parameter-Efficient Fine-Tuning with Discrete Fourier Transform

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, Jia Li

Low-rank adaptation~(LoRA) has recently gained much interest in fine-tuning foundation models. It effectively reduces the number of trainable parameters by incorporating low-rank matrices A and B to represent the weight change, i.e., $\Delta W=BA$. Despite LoRA's progress, it faces storage challenges when handling extensive customization adaptations or larger base models. In this work, we aim to further compress trainable parameters by enjoying the powerful expressiveness of the Fourier transform. Specifically, we introduce FourierFT, which treats ΔW as a matrix in the spatial domain and learns only a small fraction of its spectral coefficients. With the trained spectral coefficients, we implement the inverse discrete Fourier transform to recover ΔW. Empirically, our FourierFT method shows comparable or better performance with fewer parameters than LoRA on various tasks, including natural language understanding, natural language generation, instruction tuning, and image classification. For example, when performing instruction tuning on the LLaMA2-7B model, FourierFT surpasses LoRA with only 0.064M trainable parameters, compared to LoRA's 33.5M. Our code is released at \url{https://github.com/Chaos96/fourierft}.

link: http://arxiv.org/abs/2405.03003v1

Exploring prompts to elicit memorization in masked language model-based named entity recognition

Yuxi Xia, Anastasiia Sedova, Pedro Henrique Luz de Araujo, Vasiliki Kougia, Lisa Nußbaumer, Benjamin Roth

Training data memorization in language models impacts model capability (generalization) and safety (privacy risk). This paper focuses on analyzing prompts' impact on detecting the memorization of 6 masked language model-based named entity recognition models. Specifically, we employ a diverse set of 400 automatically generated prompts, and a pairwise dataset where each pair consists of one person's name from the training set and another name out of the set. A prompt completed with a person's name serves as input for getting the model's confidence in predicting this name. Finally, the prompt performance of detecting model memorization is quantified by the percentage of name pairs for which the model has higher confidence for the name from the training set. We show that the performance of different prompts varies by as much as 16 percentage points on the same model, and prompt engineering further increases the gap.

Moreover, our experiments demonstrate that prompt performance is model-dependent but does generalize across different name sets. A comprehensive analysis indicates how prompt performance is influenced by prompt properties, contained tokens, and the model's self-attention weights on the prompt.

link: http://arxiv.org/abs/2405.03004v1