Fri 2024.05.10

Choose What You Need: Disentangled Representation Learning for Scene Text Recognition, Removal and Editing

Boqiang Zhang, Hongtao Xie, Zuan Gao, Yuxin Wang

Scene text images contain not only style information (font, background) but also content information (character, texture). Different scene text tasks need different information, but previous representation learning methods use tightly coupled features for all tasks, resulting in sub-optimal performance. We propose a Disentangled Representation Learning framework (DARLING) aimed at disentangling these two types of features for improved adaptability in better addressing various downstream tasks (choose what you really need). Specifically, we synthesize a dataset of image pairs with identical style but different content. Based on the dataset, we decouple the two types of features by the supervision design. Clearly, we directly split the visual representation into style and content features, the content features are supervised by a text recognition loss, while an alignment loss aligns the style features in the image pairs. Then, style features are employed in reconstructing the counterpart image via an image decoder with a prompt that indicates the counterpart's content. Such an operation effectively decouples the features based on their distinctive properties. To the best of our knowledge, this is the first time in the field of scene text that disentangles the inherent properties of the text images. Our method achieves state-of-the-art performance in Scene Text Recognition, Removal, and Editing.

link: http://arxiv.org/abs/2405.04377v1

$\textbf{Splat-MOVER}$: Multi-Stage, Open-Vocabulary Robotic Manipulation via Editable Gaussian Splatting

Ola Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe Kennedy III, Mac Schwager

We present Splat-MOVER, a modular robotics stack for open-vocabulary robotic manipulation, which leverages the editability of Gaussian Splatting (GSplat) scene representations to enable multi-stage manipulation tasks. Splat-MOVER consists of: (i) $\textit{ASK-Splat}$, a GSplat representation that distills latent codes for language semantics and grasp affordance into the 3D scene. ASK-Splat enables geometric, semantic, and affordance understanding of 3D scenes, which is critical for many robotics tasks; (ii) $\textit{SEE-Splat}$, a real-time scene-editing module using 3D semantic masking and infilling to visualize the motions of objects that result from robot interactions in the real-world. SEE-Splat creates a "digital twin" of the evolving environment throughout the manipulation task; and (iii) $\textit{Grasp-Splat}$, a grasp generation module that uses ASK-Splat and SEE-Splat to propose candidate grasps for open-world objects. ASK-Splat is trained in real-time from RGB images in a brief scanning phase prior to operation, while SEE-Splat and Grasp-Splat run in real-time during operation. We demonstrate the superior performance of Splat-MOVER in hardware experiments on a Kinova robot compared to two recent baselines in four single-stage, open-vocabulary manipulation tasks, as well as in four multi-stage manipulation tasks using the edited scene to reflect scene changes due to prior manipulation stages, which is not possible with the existing baselines. Code for this project and a link to the project page will be made available soon.

link: http://arxiv.org/abs/2405.04378v1

Pragmatist Intelligence: Where the Principle of Usefulness Can Take ANNs

Antonio Biki■, Sayan Mukherjee

Artificial neural networks (ANNs) perform extraordinarily on numerous tasks including classification or prediction, e.g., speech processing and image classification. These new functions are based on a computational model that is enabled to select freely all necessary internal model parameters as long as it eventually delivers the functionality it is supposed to exhibit. Here, we review the connection between the model parameter selection in machine learning (ML) algorithms running on

ANNs and the epistemological theory of neopragmatism focusing on the theory's utility and anti-representationalist aspects. To understand the consequences of the model parameter selection of an ANN, we suggest using neopragmatist theories whose implications are well studied. Incidentally, neopragmatism's notion of optimization is also based on utility considerations. This means that applying this approach elegantly reveals the inherent connections between optimization in ML, using a numerical method during the learning phase, and optimization in the ethical theory of consequentialism, where it occurs as a maxim of action. We suggest that these connections originate from the way relevance is calculated in ML systems. This could ultimately reveal a tendency for specific actions in ML systems.

link: http://arxiv.org/abs/2405.04386v1

Parallelized Multi-Agent Bayesian Optimization in Lava

Shay Snyder, Derek Gobin, Victoria Clerico, Sumedh R. Risbud, Maryam Parsa

In parallel with the continuously increasing parameter space dimensionality, search and optimization algorithms should support distributed parameter evaluations to reduce cumulative runtime. Intel's neuromorphic optimization library, Lava-Optimization, was introduced as an abstract optimization system compatible with neuromorphic systems developed in the broader Lava software framework. In this work, we introduce Lava Multi-Agent Optimization (LMAO) with native support for distributed parameter evaluations communicating with a central Bayesian optimization system. LMAO provides an abstract framework for deploying distributed optimization and search algorithms within the Lava software framework. Moreover, LMAO introduces support for random and grid search along with process connections across multiple levels of mathematical precision. We evaluate the algorithmic performance of LMAO with a traditional non-convex optimization problem, a fixed-precision transductive spiking graph neural network for citation graph classification, and a neuromorphic satellite scheduling problem. Our results highlight LMAO's efficient scaling to multiple processes, reducing cumulative runtime and minimizing the likelihood of converging to local optima.

link: http://arxiv.org/abs/2405.04387v1

DriveWorld: 4D Pre-trained Scene Understanding via World Models for Autonomous Driving

Chen Min, Dawei Zhao, Liang Xiao, Jian Zhao, Xinli Xu, Zheng Zhu, Lei Jin, Jianshu Li, Yulan Guo, Junliang Xing, Liping Jing, Yiming Nie, Bin Dai

Vision-centric autonomous driving has recently raised wide attention due to its lower cost. Pre-training is essential for extracting a universal representation. However, current vision-centric pre-training typically relies on either 2D or 3D pre-text tasks, overlooking the temporal characteristics of autonomous driving as a 4D scene understanding task. In this paper, we address this challenge by introducing a world model-based autonomous driving 4D representation learning framework, dubbed \emph{DriveWorld}, which is capable of pre-training from multi-camera driving videos in a spatio-temporal fashion. Specifically, we propose a Memory State-Space Model for spatio-temporal modelling, which consists of a Dynamic Memory Bank module for learning temporal-aware latent dynamics to predict future changes and a Static Scene Propagation module for learning spatial-aware latent statics to offer comprehensive scene contexts. We additionally introduce a Task Prompt to decouple task-aware features for various downstream tasks. The experiments demonstrate that DriveWorld delivers promising results on various autonomous driving tasks. When pre-trained with the OpenScene dataset, DriveWorld achieves a 7.5% increase in mAP for 3D object detection, a 3.0% increase in IoU for online mapping, a 5.0% increase in AMOTA for multi-object tracking, a 0.1m decrease in minADE for motion forecasting, a 3.0% increase in IoU for occupancy prediction, and a 0.34m reduction in average L2 error for planning.

link: http://arxiv.org/abs/2405.04390v1

BILTS: A novel bi-invariant local trajectory-shape descriptor for rigid-body motion

Arno Verduyn, Erwin Aertbeliën, Glenn Maes, Joris De Schutter, Maxim Vochten

Measuring the similarity between motions and established motion models is crucial for motion analysis, recognition, generation, and adaptation. To enhance similarity measurement across diverse contexts, invariant motion descriptors have been proposed. However, for rigid-body motion, few invariant descriptors exist that are bi-invariant, meaning invariant to both the body and world reference frames used to describe the motion. Moreover, their robustness to singularities is limited. This paper introduces a novel Bi-Invariant Local Trajectory-Shape descriptor (BILTS) and a corresponding dissimilarity measure. Mathematical relationships between BILTS and existing descriptors are derived, providing new insights into their properties. The paper also includes an algorithm to reproduce the motion from the BILTS descriptor, demonstrating its bidirectionality and usefulness for trajectory generation. Experimental validation using datasets of daily-life activities shows the higher robustness of the BILTS descriptor compared to the bi-invariant ISA descriptor. This higher robustness supports the further application of bi-invariant descriptors for motion recognition and generalization.

link: http://arxiv.org/abs/2405.04392v1

Efficient Online Set-valued Classification with Bandit Feedback

Zhou Wang, Xingye Qiao

Conformal prediction is a distribution-free method that wraps a given machine learning model and returns a set of plausible labels that contain the true label with a prescribed coverage rate. In practice, the empirical coverage achieved highly relies on fully observed label information from data both in the training phase for model fitting and the calibration phase for quantile estimation. This dependency poses a challenge in the context of online learning with bandit feedback, where a learner only has access to the correctness of actions (i.e., pulled an arm) but not the full information of the true label. In particular, when the pulled arm is incorrect, the learner only knows that the pulled one is not the true class label, but does not know which label is true. Additionally, bandit feedback further results in a smaller labeled dataset for calibration, limited to instances with correct actions, thereby affecting the accuracy of quantile estimation. To address these limitations, we propose Bandit Class-specific Conformal Prediction (BCCP), offering coverage guarantees on a class-specific granularity. Using an unbiased estimation of an estimand involving the true label, BCCP trains the model and makes set-valued inferences through stochastic gradient descent. Our approach overcomes the challenges of sparsely labeled data in each iteration and generalizes the reliability and applicability of conformal prediction to online decision-making environments.

link: http://arxiv.org/abs/2405.04393v1

PACIFISTA: Conflict Evaluation and Management in Open RAN

Pietro Brach del Prever, Salvatore D'Oro, Leonardo Bonati, Michele Polese, Maria Tsampazi, Heiko Lehmann, Tommaso Melodia

The O-RAN ALLIANCE is defining architectures, interfaces, operations, and security requirements for cellular networks based on Open Radio Access Network (RAN) principles. In this context, O-RAN introduced the RAN Intelligent Controllers (RICs) to enable dynamic control of cellular networks via data-driven applications referred to as rApps and xApps. RICs enable for the first time truly intelligent and self-organizing cellular networks. However, enabling the execution of many Artificial Intelligence (AI) algorithms taking autonomous control decisions to fulfill diverse (and possibly conflicting) goals poses unprecedented challenges. For instance, the execution of one xApp aiming at maximizing throughput and one aiming at minimizing energy consumption would inevitably result in diametrically opposed resource allocation strategies. Therefore, conflict management becomes a crucial component of any functional intelligent O-RAN system. This article studies the problem of conflict mitigation in O-RAN and proposes PACIFISTA, a framework to detect, characterize, and mitigate conflicts. PACIFISTA leverages a profiling pipeline to tests O-RAN applications in a sandbox environment, and combines hierarchical graphs with statistical models to detect the existence of conflicts and evaluate their severity. Experiments on Colosseum and OpenRAN Gym demonstrate PACIFISTA's ability to predict conflicts and provide valuable information before potentially conflicting xApps are deployed in production systems. We demonstrate that even O-RAN applications with similar goals can result in 16% throughput loss, and show how applications with conflicting goals might cause severe instability and result in up to 30% performance degradation. We also show that PACIFISTA can help operators to identify coexisting applications and maintain performance degradation below a tolerable threshold.

link: http://arxiv.org/abs/2405.04395v1

Predicting Transonic Flowfields in Non-Homogeneous Unstructured Grids Using Autoencoder Graph Convolutional Networks

Gabriele Immordino, Andrea Vaiuso, Andrea Da Ronch, Marcello Righi

This paper focuses on addressing challenges posed by non-homogeneous unstructured grids, commonly used in Computational Fluid Dynamics (CFD). Their prevalence in CFD scenarios has motivated the exploration of innovative approaches for generating reduced-order models. The core of our approach centers on geometric deep learning, specifically the utilization of graph convolutional network (GCN). The novel Autoencoder GCN architecture enhances prediction accuracy by propagating information to distant nodes and emphasizing influential points. This architecture, with GCN layers and encoding/decoding modules, reduces dimensionality based on pressure-gradient values. The autoencoder structure improves the network capability to identify key features, contributing to a more robust and accurate predictive model. To validate the proposed methodology, we analyzed two different test cases: wing-only model and wing--body configuration. Precise reconstruction of steady-state distributed quantities within a two-dimensional parametric space underscores the reliability and versatility of the implemented approach.

link: http://arxiv.org/abs/2405.04396v1

Utility-driven Optimization of TTL Cache Hierarchies under Network Delays

Karim S. Elsayed, Fabien Geyer, Amr Rizk

We optimize hierarchies of Time-to-Live (TTL) caches under random network delays. A TTL cache assigns individual eviction timers to cached objects that are usually refreshed upon a hit where upon a miss the object requires a random time to be fetched from a parent cache. Due to their object decoupling property, TTL caches are of particular interest since the optimization of a per-object utility enables service differentiation. However, state-of-the-art exact TTL cache optimization does not extend beyond single TTL caches, especially under network delays. In this paper, we leverage the object decoupling effect to formulate the non-linear utility maximization problem for TTL cache hierarchies in terms of the exact object hit probability under random network delays. We iteratively solve the utility maximization problem to find the optimal per-object TTLs. Further, we show that the exact model suffers from tractability issues for large hierarchies and propose a machine learning approach to estimate the optimal TTL values for large systems. Finally, we provide numerical and data center trace-based evaluations for both methods showing the significant offloading improvement due to TTL optimization considering the network delays.

link: http://arxiv.org/abs/2405.04402v1

Learning To See But Forgetting To Follow: Visual Instruction Tuning Makes LLMs More Prone To Jailbreak Attacks

Georgios Pantazopoulos, Amit Parekh, Malvina Nikandrou, Alessandro Suglia

Augmenting Large Language Models (LLMs) with image-understanding capabilities has resulted in a boom of high-performing Vision-Language models (VLMs). While studying the alignment of LLMs to human values has received widespread attention, the safety of VLMs has not received the same attention. In this paper, we explore the impact of jailbreaking on three state-of-the-art VLMs, each using a distinct modeling approach. By comparing each VLM to their respective LLM backbone, we find that each VLM is more susceptible to jailbreaking. We consider this as an undesirable outcome from visual instruction-tuning, which imposes a forgetting effect on an LLM's safety guardrails. Therefore, we provide recommendations for future work based on evaluation strategies that aim to highlight the weaknesses of a VLM, as well as take safety measures into account during visual instruction tuning.

link: http://arxiv.org/abs/2405.04403v1

Vision Mamba: A Comprehensive Survey and Taxonomy

Xiao Liu, Chenxu Zhang, Lei Zhang

State Space Model (SSM) is a mathematical model used to describe and analyze the behavior of dynamic systems. This model has witnessed numerous applications in several fields, including control theory, signal processing, economics and machine learning. In the field of deep learning, state space models are used to process sequence data, such as time series analysis, natural language processing (NLP) and video understanding. By mapping sequence data to state space, long-term dependencies in the data can be better captured. In particular, modern SSMs have shown strong representational capabilities in NLP, especially in long sequence modeling, while maintaining linear time complexity. Notably, based on the latest state-space models, Mamba merges time-varying parameters into SSMs and formulates a hardware-aware algorithm for efficient training and inference. Given its impressive efficiency and strong long-range dependency modeling capability, Mamba is expected to become a new AI architecture that may outperform Transformer. Recently, a number of works have attempted to study the potential of Mamba in various fields, such as general vision, multi-modal, medical image analysis and remote sensing image analysis, by extending Mamba from natural language domain to visual domain. To fully understand Mamba in the visual domain, we conduct a comprehensive survey and present a taxonomy study. This survey focuses on Mamba's application to a variety of visual tasks and data types, and discusses its predecessors, recent advances and far-reaching impact on a wide range of domains. Since Mamba is now on an upward trend, please actively notice us if you have new findings, and new progress on Mamba will be included in this survey in a timely manner and updated on the Mamba project at https://github.com/lx6c78/Vision-Mamba-A-Comprehensive-Survey-and-Taxonomy.

link: http://arxiv.org/abs/2405.04404v1

Weakly-Supervised Residual Evidential Learning for Multi-Instance Uncertainty Estimation

Pei Liu, Luping Ji

Uncertainty estimation (UE), as an effective means of quantifying predictive uncertainty, is crucial for safe and reliable decision-making, especially in high-risk scenarios. Existing UE schemes usually assume that there are completely-labeled samples to support fully-supervised learning. In practice, however, many UE tasks often have no sufficiently-labeled data to use, such as the Multiple Instance Learning (MIL) with only weak instance annotations. To bridge this gap, this paper, for the first time, addresses the weakly-supervised issue of Multi-Instance UE (MIUE) and proposes a new baseline scheme, Multi-Instance Residual Evidential Learning (MIREL). Particularly, at the fine-grained instance UE with only weak supervision, we derive a multi-instance residual operator through the Fundamental Theorem of Symmetric Functions. On this operator derivation, we further propose MIREL to jointly model the high-order predictive distribution at bag and instance levels for MIUE. Extensive experiments empirically demonstrate that our MIREL not only could often make existing MIL networks perform better in MIUE, but also could surpass representative UE methods by large margins, especially in instance-level UE tasks. Our source code is available at https://github.com/liupei101/MIREL.

link: http://arxiv.org/abs/2405.04405v2

Super-Exponential Regret for UCT, AlphaGo and Variants

Laurent Orseau, Remi Munos

We improve the proofs of the lower bounds of Coquelin and Munos (2007) that demonstrate that UCT can have $\exp(\dots\exp(1)\dots)$ regret (with $\Omega(D)$ exp terms) on the D-chain environment, and that a `polynomial' UCT variant has $\exp_2(\exp_2(D - O(\log D)))$ regret on the same environment -- the original proofs contain an oversight for rewards bounded in $[0, 1]$, which we fix in the present draft. We also adapt the proofs to AlphaGo's MCTS and its descendants (e.g., AlphaZero, Leela Zero) to also show $\exp_2(\exp_2(D - O(\log D)))$ regret.

link: http://arxiv.org/abs/2405.04407v1

DocRes: A Generalist Model Toward Unifying Document Image Restoration Tasks

Jiaxin Zhang, Dezhi Peng, Chongyu Liu, Peirong Zhang, Lianwen Jin

Document image restoration is a crucial aspect of Document AI systems, as the quality of document images significantly influences the overall performance. Prevailing methods address distinct restoration tasks independently, leading to intricate systems and the incapability to harness the potential synergies of multi-task learning. To overcome this challenge, we propose DocRes, a generalist model that unifies five document image restoration tasks including dewarping, deshadowing, appearance enhancement, deblurring, and binarization. To instruct DocRes to perform various restoration tasks, we propose a novel visual prompt approach called Dynamic Task-Specific Prompt (DTSPrompt). The DTSPrompt for different tasks comprises distinct prior features, which are additional characteristics extracted from the input image. Beyond its role as a cue for task-specific execution, DTSPrompt can also serve as supplementary information to enhance the model's performance. Moreover, DTSPrompt is more flexible than prior visual prompt approaches as it can be seamlessly applied and adapted to inputs with high and variable resolutions. Experimental results demonstrate that DocRes achieves competitive or superior performance compared to existing state-of-the-art task-specific models. This underscores the potential of DocRes across a broader spectrum of document image restoration tasks. The source code is publicly available at https://github.com/ZZZHANG-jx/DocRes

link: http://arxiv.org/abs/2405.04408v1