Wed 2024.03.06

Iterated Q-Network: Beyond the One-Step Bellman Operator

Théo Vincent, Daniel Palenicek, Boris Belousov, Jan Peters, Carlo D'Eramo

Value-based Reinforcement Learning (RL) methods rely on the application of the Bellman operator, which needs to be approximated from samples. Most approaches consist of an iterative scheme alternating the application of the Bellman operator and a subsequent projection step onto a considered function space. However, we observe that these algorithms can be improved by considering multiple iterations of the Bellman operator at once. Thus, we introduce iterated Q-Networks (iQN), a novel approach that learns a sequence of Q-function approximations where each Q-function serves as the target for the next one in a chain of consecutive Bellman iterations. We demonstrate that iQN is theoretically sound and show how it can be seamlessly used in value-based and actor-critic methods. We empirically demonstrate its advantages on Atari 2600 games and in continuous-control MuJoCo environments.

link: http://arxiv.org/abs/2403.02107v1

A New Perspective on Smiling and Laughter Detection: Intensity Levels Matter

Hugo Bohy, Kevin El Haddad, Thierry Dutoit

Smiles and laughs detection systems have attracted a lot of attention in the past decade contributing to the improvement of human-agent interaction systems. But very few considered these expressions as distinct, although no prior work clearly proves them to belong to the same category or not. In this work, we present a deep learning-based multimodal smile and laugh classification system, considering them as two different entities. We compare the use of audio and vision-based models as well as a fusion approach. We show that, as expected, the fusion leads to a better generalization on unseen data. We also present an in-depth analysis of the behavior of these models on the smiles and laughs intensity levels. The analyses on the intensity levels show that the relationship between smiles and laughs might not be as simple as a binary one or even grouping them in a single category, and so, a more complex approach should be taken when dealing with them. We also tackle the problem of limited resources by showing that transfer learning allows the models to improve the detection of confusing intensity levels.

link: http://dx.doi.org/10.1109/ACII55700.2022.9953896

Inf2Guard: An Information-Theoretic Framework for Learning Privacy-Preserving Representations against Inference Attacks

Sayedeh Leila Noorbakhsh, Binghui Zhang, Yuan Hong, Binghui Wang

Machine learning (ML) is vulnerable to inference (e.g., membership inference, property inference, and data reconstruction) attacks that aim to infer the private information of training data or dataset. Existing defenses are only designed for one specific type of attack and sacrifice significant utility or are soon broken by adaptive attacks. We address these limitations by proposing an information-theoretic defense framework, called Inf2Guard, against the three major types of inference attacks. Our framework, inspired by the success of representation learning, posits that learning shared representations not only saves time/costs but also benefits numerous downstream tasks. Generally, Inf2Guard involves two mutual information objectives, for privacy protection and utility preservation, respectively. Inf2Guard exhibits many merits: it facilitates the design of customized objectives against the specific inference attack; it provides a general defense framework which can treat certain existing defenses as special cases; and importantly, it aids in deriving theoretical results, e.g., inherent utility-privacy tradeoff and guaranteed privacy leakage. Extensive evaluations validate the effectiveness of Inf2Guard for learning privacy-preserving representations against inference attacks and demonstrate the superiority over the baselines.

link: http://arxiv.org/abs/2403.02116v1

Position Paper: Towards Implicit Prompt For Text-To-Image Models

Yue Yang, Yuqi lin, Hong Liu, Wenqi Shao, Runjian Chen, Hailong Shang, Yu Wang, Yu Qiao, Kaipeng Zhang, Ping Luo

Recent text-to-image (T2I) models have had great success, and many benchmarks have been proposed to evaluate their performance and safety. However, they only consider explicit prompts while neglecting implicit prompts (hint at a target without explicitly mentioning it). These prompts may get rid of safety constraints and pose potential threats to the applications of these models. This position paper highlights the current state of T2I models toward implicit prompts. We present a benchmark named ImplicitBench and conduct an investigation on the performance and impacts of implicit prompts with popular T2I models. Specifically, we design and collect more than 2,000 implicit prompts of three aspects: General Symbols, Celebrity Privacy, and Not-Safe-For-Work (NSFW) Issues, and evaluate six well-known T2I models' capabilities under these implicit prompts. Experiment results show that (1) T2I models are able to accurately create various target symbols indicated by implicit prompts; (2) Implicit prompts bring potential risks of privacy leakage for T2I models. (3) Constraints of NSFW in most of the evaluated T2I models can be bypassed with implicit prompts. We call for increased attention to the potential and risks of implicit prompts in the T2I community and further investigation into the capabilities and impacts of implicit prompts, advocating for a balanced approach that harnesses their benefits while mitigating their risks.

link: http://arxiv.org/abs/2403.02118v1

Leveraging Weakly Annotated Data for Hate Speech Detection in Code-Mixed Hinglish: A Feasibility-Driven Transfer Learning Approach with Large Language Models

Sargam Yadav, Abhishek Kaushik, Kevin McDaid

The advent of Large Language Models (LLMs) has advanced the benchmark in various Natural Language Processing (NLP) tasks. However, large amounts of labelled training data are required to train LLMs. Furthermore, data annotation and training are computationally expensive and time-consuming. Zero and few-shot learning have recently emerged as viable options for labelling data using large pre-trained models. Hate speech detection in mix-code low-resource languages is an active problem area where the use of LLMs has proven beneficial. In this study, we have compiled a dataset of 100 YouTube comments, and weakly labelled them for coarse and fine-grained misogyny classification in mix-code Hinglish. Weak annotation was applied due to the labor-intensive annotation process. Zero-shot learning, one-shot learning, and few-shot learning and prompting approaches have then been applied to assign labels to the comments and compare them to human-assigned labels. Out of all the approaches, zero-shot classification using the Bidirectional Auto-Regressive Transformers (BART) large model and few-shot prompting using Generative Pre-trained Transformer- 3 (ChatGPT-3) achieve the best results

link: http://arxiv.org/abs/2403.02121v1

LOCR: Location-Guided Transformer for Optical Character Recognition

Yu Sun, Dongzhan Zhou, Chen Lin, Conghui He, Wanli Ouyang, Han-Sen Zhong

Academic documents are packed with texts, equations, tables, and figures, requiring comprehensive understanding for accurate Optical Character Recognition (OCR). While end-to-end OCR methods offer improved accuracy over layout-based approaches, they often grapple with significant repetition issues, especially with complex layouts in Out-Of-Domain (OOD) documents.To tackle this issue, we propose LOCR, a model that integrates location guiding into the transformer architecture during autoregression. We train the model on a dataset comprising over 77M text-location pairs from 125K academic document pages, including bounding boxes for words, tables and mathematical symbols. LOCR adeptly handles various formatting elements and generates content in Markdown language. It outperforms all existing methods in our test set constructed from arXiv, as measured by edit distance, BLEU, METEOR and F-measure.LOCR also reduces repetition frequency from 4.4% of pages to 0.5% in the arXiv dataset, from 13.2% to 1.3% in OOD quantum physics documents and from 8.1% to 1.8% in OOD marketing documents. Additionally, LOCR features an interactive OCR mode, facilitating the generation of complex

documents through a few location prompts from human.

link: http://arxiv.org/abs/2403.02127v1

Demeter: Resource-Efficient Distributed Stream Processing under Dynamic Loads with Multi-Configuration Optimization

Morgan Geldenhuys, Dominik Scheinert, Odej Kao, Lauritz Thamsen

Distributed Stream Processing (DSP) focuses on the near real-time processing of large streams of unbounded data. To increase processing capacities, DSP systems are able to dynamically scale across a cluster of commodity nodes, ensuring a good Quality of Service despite variable workloads. However, selecting scaleout configurations which maximize resource utilization remains a challenge. This is especially true in environments where workloads change over time and node failures are all but inevitable. Furthermore, configuration parameters such as memory allocation and checkpointing intervals impact performance and resource usage as well. Sub-optimal configurations easily lead to high operational costs, poor performance, or unacceptable loss of service. In this paper, we present Demeter, a method for dynamically optimizing key DSP system configuration parameters for resource efficiency. Demeter uses Time Series Forecasting to predict future workloads and Multi-Objective Bayesian Optimization to model runtime behaviors in relation to parameter settings and workload rates. Together, these techniques allow us to determine whether or not enough is known about the predicted workload rate to proactively initiate short-lived parallel profiling runs for data gathering. Once trained, the models guide the adjustment of multiple, potentially dependent system configuration parameters ensuring optimized performance and resource usage in response to changing workload rates. Our experiments on a commodity cluster using Apache Flink demonstrate that Demeter significantly improves the operational efficiency of long-running benchmark jobs.

link: http://arxiv.org/abs/2403.02129v1

Using LLMs for the Extraction and Normalization of Product Attribute Values

Nick Baumann, Alexander Brinkmann, Christian Bizer

Product offers on e-commerce websites often consist of a textual product title and a textual product description. In order to provide features such as faceted product filtering or content-based product recommendation, the websites need to extract attribute-value pairs from the unstructured product descriptions. This paper explores the potential of using large language models (LLMs), such as OpenAI's GPT-3.5 and GPT-4, to extract and normalize attribute values from product titles and product descriptions. For our experiments, we introduce the WDC Product Attribute-Value Extraction (WDC PAVE) dataset. WDC PAVE consists of product offers from 87 websites that provide schema$.$org annotations. The offers belong to five different categories, each featuring a specific set of attributes. The dataset provides manually verified attribute-value pairs in two forms: (i) directly extracted values and (ii) normalized attribute values. The normalization of the attribute values requires systems to perform the following types of operations: name expansion, generalization, unit of measurement normalization, and string wrangling. Our experiments demonstrate that GPT-4 outperforms PLM-based extraction methods by 10%, achieving an F1-Score of 91%. For the extraction and normalization of product attribute values, GPT-4 achieves a similar performance to the extraction scenario, while being particularly strong at string wrangling and name expansion.

link: http://arxiv.org/abs/2403.02130v2

Deep Reinforcement Learning for Dynamic Algorithm Selection: A Proof-of-Principle Study on Differential Evolution

Hongshu Guo, Yining Ma, Zeyuan Ma, Jiacheng Chen, Xinglin Zhang, Zhiguang Cao, Jun Zhang, Yue-Jiao Gong

Evolutionary algorithms, such as Differential Evolution, excel in solving real-parameter optimization challenges. However, the effectiveness of a single algorithm varies across different problem instances, necessitating considerable efforts in algorithm selection or configuration. This paper

aims to address the limitation by leveraging the complementary strengths of a group of algorithms and dynamically scheduling them throughout the optimization progress for specific problems. We propose a deep reinforcement learning-based dynamic algorithm selection framework to accomplish this task. Our approach models the dynamic algorithm selection a Markov Decision Process, training an agent in a policy gradient manner to select the most suitable algorithm according to the features observed during the optimization process. To empower the agent with the necessary information, our framework incorporates a thoughtful design of landscape and algorithmic features. Meanwhile, we employ a sophisticated deep neural network model to infer the optimal action, ensuring informed algorithm selections. Additionally, an algorithm context restoration mechanism is embedded to facilitate smooth switching among different algorithms. These mechanisms together enable our framework to seamlessly select and switch algorithms in a dynamic online fashion. Notably, the proposed framework is simple and generic, offering potential improvements across a broad spectrum of evolutionary algorithms. As a proof-of-principle study, we apply this framework to a group of Differential Evolution algorithms. The experimental results showcase the remarkable effectiveness of the proposed framework, not only enhancing the overall optimization performance but also demonstrating favorable generalization ability across different problem classes.

link: http://arxiv.org/abs/2403.02131v1

UB-FineNet: Urban Building Fine-grained Classification Network for Open-access Satellite Images

Zhiyi He, Wei Yao, Jie Shao, Puzuo Wang

Fine classification of city-scale buildings from satellite remote sensing imagery is a crucial research area with significant implications for urban planning, infrastructure development, and population distribution analysis. However, the task faces big challenges due to low-resolution overhead images acquired from high altitude space-borne platforms and the long-tail sample distribution of fine-grained urban building categories, leading to severe class imbalance problem. To address these issues, we propose a deep network approach to fine-grained classification of urban buildings using open-access satellite images. A Denoising Diffusion Probabilistic Model (DDPM) based super-resolution method is first introduced to enhance the spatial resolution of satellite images, which benefits from domain-adaptive knowledge distillation. Then, a new fine-grained classification network with Category Information Balancing Module (CIBM) and Contrastive Supervision (CS) technique is proposed to mitigate the problem of class imbalance and improve the classification robustness and accuracy. Experiments on Hong Kong data set with 11 fine building types revealed promising classification results with a mean Top-1 accuracy of 60.45\%, which is on par with street-view image based approaches. Extensive ablation study shows that CIBM and CS improve Top-1 accuracy by 2.6\% and 3.5\% compared to the baseline method, respectively. And both modules can be easily inserted into other classification networks and similar enhancements have been achieved. Our research contributes to the field of urban analysis by providing a practical solution for fine classification of buildings in challenging mega city scenarios solely using open-access satellite images. The proposed method can serve as a valuable tool for urban planners, aiding in the understanding of economic, industrial, and population distribution.

link: http://arxiv.org/abs/2403.02132v1

Point2Building: Reconstructing Buildings from Airborne LiDAR Point Clouds

Yujia Liu, Anton Obukhov, Jan Dirk Wegner, Konrad Schindler

We present a learning-based approach to reconstruct buildings as 3D polygonal meshes from airborne LiDAR point clouds. What makes 3D building reconstruction from airborne LiDAR hard is the large diversity of building designs and especially roof shapes, the low and varying point density across the scene, and the often incomplete coverage of building facades due to occlusions by vegetation or to the viewing angle of the sensor. To cope with the diversity of shapes and inhomogeneous and incomplete object coverage, we introduce a generative model that directly predicts 3D polygonal meshes from input point clouds. Our autoregressive model, called Point2Building, iteratively builds up the mesh by generating sequences of vertices and faces. This

approach enables our model to adapt flexibly to diverse geometries and building structures. Unlike many existing methods that rely heavily on pre-processing steps like exhaustive plane detection, our model learns directly from the point cloud data, thereby reducing error propagation and increasing the fidelity of the reconstruction. We experimentally validate our method on a collection of airborne LiDAR data of Zurich, Berlin and Tallinn. Our method shows good generalization to diverse urban styles.

link: http://arxiv.org/abs/2403.02136v1

Self-Supervised Facial Representation Learning with Facial Region Awareness

Zheng Gao, Ioannis Patras

Self-supervised pre-training has been proved to be effective in learning transferable representations that benefit various visual tasks. This paper asks this question: can self-supervised pre-training learn general facial representations for various facial analysis tasks? Recent efforts toward this goal are limited to treating each face image as a whole, i.e., learning consistent facial representations at the image-level, which overlooks the consistency of local facial representations (i.e., facial regions like eyes, nose, etc). In this work, we make a first attempt to propose a novel self-supervised facial representation learning framework to learn consistent global and local facial representations, Facial Region Awareness (FRA). Specifically, we explicitly enforce the consistency of facial regions by matching the local facial representations across views, which are extracted with learned heatmaps highlighting the facial regions. Inspired by the mask prediction in supervised semantic segmentation, we obtain the heatmaps via cosine similarity between the per-pixel projection of feature maps and facial mask embeddings computed from learnable positional embeddings, which leverage the attention mechanism to globally look up the facial image for facial regions. To learn such heatmaps, we formulate the learning of facial mask embeddings as a deep clustering problem by assigning the pixel features from the feature maps to them. The transfer learning results on facial classification and regression tasks show that our FRA outperforms previous pre-trained models and more importantly, using ResNet as the unified backbone for various tasks, our FRA achieves comparable or even better performance compared with SOTA methods in facial analysis tasks.

link: http://arxiv.org/abs/2403.02138v1

MiM-ISTD: Mamba-in-Mamba for Efficient Infrared Small Target Detection

Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Jieping Ye, Nenghai Yu

Thanks to the development of basic models, infrared small target detection (ISTD) algorithms have made significant progress. Specifically, the structures combining convolutional networks with transformers can well extract both local and global features. At the same time, they also inherit defects from the basic model, e.g., the quadratic computational complexity of transformers, which impacts efficiency. Inspired by a recent basic model with linear complexity for long-distance modeling, called Mamba, we explore the potential of this state space model in ISTD in this paper. However, direct application is unsuitable since local features, which are critical to detecting small targets, cannot be fully exploited. Instead, we tailor a Mamba-in-Mamba (MiM-ISTD) structure for efficient ISTD. For example, we treat the local patches as "visual sentences" and further decompose them into sub-patches as "visual words" to further explore the locality. The interactions among each word in a given visual sentence will be calculated with negligible computational costs. By aggregating the word and sentence features, the representation ability of MiM-ISTD can be significantly bolstered. Experiments on NUAA-SIRST and IRSTD-1k prove the superior accuracy and efficiency of our method. Specifically, MiM-ISTD is $10 \times$ faster than the SOTA and reduces GPU memory usage by 73.4$\%$ per 2048×2048 image during inference, overcoming the computation$\&$memory constraints on performing Mamba-based understanding on high-resolution infrared images.Source code is available at https://github.com/txchen-USTC/MiM-ISTD.

link: http://arxiv.org/abs/2403.02148v1

Recency-Weighted Temporally-Segmented Ensemble for Time-Series Modeling

Pål V. Johnsen, Eivind Bøhn, Sølve Eidnes, Filippo Remonato, Signe Riemer-Sørensen

Time-series modeling in process industries faces the challenge of dealing with complex, multi-faceted, and evolving data characteristics. Conventional single model approaches often struggle to capture the interplay of diverse dynamics, resulting in suboptimal forecasts. Addressing this, we introduce the Recency-Weighted Temporally-Segmented (ReWTS, pronounced `roots') ensemble model, a novel chunk-based approach for multi-step forecasting. The key characteristics of the ReWTS model are twofold: 1) It facilitates specialization of models into different dynamics by segmenting the training data into `chunks' of data and training one model per chunk. 2) During inference, an optimization procedure assesses each model on the recent past and selects the active models, such that the appropriate mixture of previously learned dynamics can be recalled to forecast the future. This method not only captures the nuances of each period, but also adapts more effectively to changes over time compared to conventional `global' models trained on all data in one go. We present a comparative analysis, utilizing two years of data from a wastewater treatment plant and a drinking water treatment plant in Norway, demonstrating the ReWTS ensemble's superiority. It consistently outperforms the global model in terms of mean squared forecasting error across various model architectures by 10-70\% on both datasets, notably exhibiting greater resilience to outliers. This approach shows promise in developing automatic, adaptable forecasting models for decision-making and control systems in process industries and other complex systems.

link: http://arxiv.org/abs/2403.02150v1

TripoSR: Fast 3D Object Reconstruction from a Single Image

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, Yan-Pei Cao

This technical report introduces TripoSR, a 3D reconstruction model leveraging transformer architecture for fast feed-forward 3D generation, producing 3D mesh from a single image in under 0.5 seconds. Building upon the LRM network architecture, TripoSR integrates substantial improvements in data processing, model design, and training techniques. Evaluations on public datasets show that TripoSR exhibits superior performance, both quantitatively and qualitatively, compared to other open-source alternatives. Released under the MIT license, TripoSR is intended to empower researchers, developers, and creatives with the latest advancements in 3D generative AI.

link: http://arxiv.org/abs/2403.02151v1