Thu 2024.02.22

Synthetic Data (Almost) from Scratch: Generalized Instruction Tuning for Language Models

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, Furu Wei

We introduce Generalized Instruction Tuning (called GLAN), a general and scalable method for instruction tuning of Large Language Models (LLMs). Unlike prior work that relies on seed examples or existing datasets to construct instruction tuning data, GLAN exclusively utilizes a pre-curated taxonomy of human knowledge and capabilities as input and generates large-scale synthetic instruction data across all disciplines. Specifically, inspired by the systematic structure in human education system, we build the taxonomy by decomposing human knowledge and capabilities to various fields, sub-fields and ultimately, distinct disciplines semi-automatically, facilitated by LLMs. Subsequently, we generate a comprehensive list of subjects for every discipline and proceed to design a syllabus tailored to each subject, again utilizing LLMs. With the fine-grained key concepts detailed in every class session of the syllabus, we are able to generate diverse instructions with a broad coverage across the entire spectrum of human knowledge and skills. Extensive experiments on large language models (e.g., Mistral) demonstrate that GLAN excels in multiple dimensions from mathematical reasoning, coding, academic exams, logical reasoning to general instruction following without using task-specific training data of these tasks. In addition, GLAN allows for easy customization and new fields or skills can be added by simply incorporating a new node into our taxonomy.

link: http://arxiv.org/abs/2402.13064v1

Not All Weights Are Created Equal: Enhancing Energy Efficiency in On-Device Streaming Speech Recognition

Yang Li, Yuan Shangguan, Yuhao Wang, Liangzhen Lai, Ernie Chang, Changsheng Zhao, Yangyang Shi, Vikas Chandra

Power consumption plays an important role in on-device streaming speech recognition, as it has a direct impact on the user experience. This study delves into how weight parameters in speech recognition models influence the overall power consumption of these models. We discovered that the impact of weight parameters on power consumption varies, influenced by factors including how often they are invoked and their placement in memory. Armed with this insight, we developed design guidelines aimed at optimizing on-device speech recognition models. These guidelines focus on minimizing power use without substantially affecting accuracy. Our method, which employs targeted compression based on the varying sensitivities of weight parameters, demonstrates superior performance compared to state-of-the-art compression methods. It achieves a reduction in energy usage of up to 47% while maintaining similar model accuracy and improving the real-time factor.

link: http://arxiv.org/abs/2402.13076v1

Mechanistic Neural Networks for Scientific Machine Learning

Adeel Pervez, Francesco Locatello, Efstratios Gavves

This paper presents Mechanistic Neural Networks, a neural network design for machine learning applications in the sciences. It incorporates a new Mechanistic Block in standard architectures to explicitly learn governing differential equations as representations, revealing the underlying dynamics of data and enhancing interpretability and efficiency in data modeling. Central to our approach is a novel Relaxed Linear Programming Solver (NeuRLP) inspired by a technique that reduces solving linear ODEs to solving linear programs. This integrates well with neural networks and surpasses the limitations of traditional ODE solvers enabling scalable GPU parallel processing. Overall, Mechanistic Neural Networks demonstrate their versatility for scientific machine learning applications, adeptly managing tasks from equation discovery to dynamic systems modeling. We prove their comprehensive capabilities in analyzing and interpreting complex scientific data across various applications, showing significant performance against specialized state-of-the-art methods.

link: http://arxiv.org/abs/2402.13077v1

Mode Estimation with Partial Feedback

Charles Arnal, Vivien Cabannes, Vianney Perchet

The combination of lightly supervised pre-training and online fine-tuning has played a key role in recent AI developments. These new learning pipelines call for new theoretical frameworks. In this paper, we formalize core aspects of weakly supervised and active learning with a simple problem: the estimation of the mode of a distribution using partial feedback. We show how entropy coding allows for optimal information acquisition from partial feedback, develop coarse sufficient statistics for mode identification, and adapt bandit algorithms to our new setting. Finally, we combine those contributions into a statistically and computationally efficient solution to our problem.

link: http://arxiv.org/abs/2402.13079v1

IT Intrusion Detection Using Statistical Learning and Testbed Measurements

Xiaoxuan Wang, Rolf Stadler

We study automated intrusion detection in an IT infrastructure, specifically the problem of identifying the start of an attack, the type of attack, and the sequence of actions an attacker takes, based on continuous measurements from the infrastructure. We apply statistical learning methods, including Hidden Markov Model (HMM), Long Short-Term Memory (LSTM), and Random Forest Classifier (RFC) to map sequences of observations to sequences of predicted attack actions. In contrast to most related research, we have abundant data to train the models and evaluate their predictive power. The data comes from traces we generate on an in-house testbed where we run attacks against an emulated IT infrastructure. Central to our work is a machine-learning pipeline that maps measurements from a high-dimensional observation space to a space of low dimensionality or to a small set of observation symbols. Investigating intrusions in offline as well as online scenarios, we find that both HMM and LSTM can be effective in predicting attack start time, attack type, and attack actions. If sufficient training data is available, LSTM achieves higher prediction accuracy than HMM. HMM, on the other hand, requires less computational resources and less training data for effective prediction. Also, we find that the methods we study benefit from data produced by traditional intrusion detection systems like SNORT.

link: http://arxiv.org/abs/2402.13081v1

How Does Selection Leak Privacy: Revisiting Private Selection and Improved Results for Hyper-parameter Tuning

Zihang Xiang, Chenglong Wang, Di Wang

We study the problem of guaranteeing Differential Privacy (DP) in hyper-parameter tuning, a crucial process in machine learning involving the selection of the best run from several. Unlike many private algorithms, including the prevalent DP-SGD, the privacy implications of tuning remain insufficiently understood. Recent works propose a generic private solution for the tuning process, yet a fundamental question still persists: is the current privacy bound for this solution tight? This paper contributes both positive and negative answers to this question. Initially, we provide studies affirming the current privacy analysis is indeed tight in a general sense. However, when we specifically study the hyper-parameter tuning problem, such tightness no longer holds. This is first demonstrated by applying privacy audit on the tuning process. Our findings underscore a substantial gap between the current theoretical privacy bound and the empirical bound derived even under the strongest audit setup. The gap found is not a fluke. Our subsequent study provides an improved privacy result for private hyper-parameter tuning due to its distinct properties. Our privacy results are also more generalizable compared to prior analyses that are only easily applicable in specific setups.

link: http://arxiv.org/abs/2402.13087v1

Slot-VLM: SlowFast Slots for Video-Language Modeling

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, Yan Lu

Video-Language Models (VLMs), powered by the advancements in Large Language Models (LLMs), are charting new frontiers in video understanding. A pivotal challenge is the development of an efficient method to encapsulate video content into a set of representative tokens to align with LLMs. In this work, we introduce Slot-VLM, a novel framework designed to generate semantically decomposed video tokens, in terms of object-wise and event-wise visual representations, to facilitate LLM inference. Particularly, we design a SlowFast Slots module, i.e., SF-Slots, that adaptively aggregates the dense video tokens from the CLIP vision encoder to a set of representative slots. In order to take into account both the spatial object details and the varied temporal dynamics, SF-Slots is built with a dual-branch structure. The Slow-Slots branch focuses on extracting object-centric slots from features at high spatial resolution but low (slow) frame sample rate, emphasizing detailed object information. Conversely, Fast-Slots branch is engineered to learn event-centric slots from high temporal sample rate but low spatial resolution features. These complementary slots are combined to form the vision context, serving as the input to the LLM for efficient question answering. Our experimental results demonstrate the effectiveness of our Slot-VLM, which achieves the state-of-the-art performance on video question-answering.

link: http://arxiv.org/abs/2402.13088v1

Towards an empirical understanding of MoE design choices

Dongyang Fan, Bettina Messmer, Martin Jaggi

In this study, we systematically evaluate the impact of common design choices in Mixture of Experts (MoEs) on validation performance, uncovering distinct influences at token and sequence levels. We also present empirical evidence showing comparable performance between a learned router and a frozen, randomly initialized router, suggesting that learned routing may not be essential. Our study further reveals that Sequence-level routing can result in topic-specific weak expert specialization, in contrast to syntax specialization observed with Token-level routing.

link: http://arxiv.org/abs/2402.13089v1

Event-level Knowledge Editing

Hao Peng, Xiaozhi Wang, Chunyang Li, Kaisheng Zeng, Jiangshan Duo, Yixin Cao, Lei Hou, Juanzi Li

Knowledge editing aims at updating knowledge of large language models (LLMs) to prevent them from becoming outdated. Existing work edits LLMs at the level of factual knowledge triplets. However, natural knowledge updates in the real world come from the occurrences of new events rather than direct changes in factual triplets. In this paper, we propose a new task setting: event-level knowledge editing, which directly edits new events into LLMs and improves over conventional triplet-level editing on (1) Efficiency. A single event edit leads to updates in multiple entailed knowledge triplets. (2) Completeness. Beyond updating factual knowledge, event-level editing also requires considering the event influences and updating LLMs' knowledge about future trends. We construct a high-quality event-level editing benchmark ELKEN, consisting of 1,515 event edits, 6,449 questions about factual knowledge, and 10,150 questions about future tendencies. We systematically evaluate the performance of various knowledge editing methods and LLMs on this benchmark. We find that ELKEN poses significant challenges to existing knowledge editing approaches. Our codes and dataset are publicly released to facilitate further research.

link: http://arxiv.org/abs/2402.13093v1

Digital Comprehensibility Assessment of Simplified Texts among Persons with Intellectual Disabilities

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, Sarah Ebling

Text simplification refers to the process of increasing the comprehensibility of texts. Automatic text simplification models are most commonly evaluated by experts or crowdworkers instead of the primary target groups of simplified texts, such as persons with intellectual disabilities. We conducted an evaluation study of text comprehensibility including participants with and without intellectual disabilities reading unsimplified, automatically and manually simplified German texts on a tablet computer. We explored four different approaches to measuring comprehensibility: multiple-choice comprehension questions, perceived difficulty ratings, response time, and reading speed. The results revealed significant variations in these measurements, depending on the reader group and whether the text had undergone automatic or manual simplification. For the target group of persons with intellectual disabilities, comprehension questions emerged as the most reliable measure, while analyzing reading speed provided valuable insights into participants' reading behavior.

link: http://arxiv.org/abs/2402.13094v1

A Lightweight Machine Learning Approach for Delay-Aware Cell-Switching in 6G HAPS Networks

Görkem Berkay Koç, Berk Çilo■lu, Metin Ozturk, Halim Yanikomeroglu

This study investigates the integration of a high altitude platform station (HAPS), a non-terrestrial network (NTN) node, into the cell-switching paradigm for energy saving. By doing so, the sustainability and ubiquitous connectivity targets can be achieved. Besides, a delay-aware approach is also adopted, where the delay profiles of users are respected in such a way that we attempt to meet the latency requirements of users with a best-effort strategy. To this end, a novel, simple, and lightweight Q-learning algorithm is designed to address the cell-switching optimization problem. During the simulation campaigns, different interference scenarios and delay situations between base stations are examined in terms of energy consumption and quality-of-service (QoS), and the results confirm the efficacy of the proposed Q-learning algorithm.

link: http://arxiv.org/abs/2402.13096v1

ELAD: Explanation-Guided Large Language Models Active Distillation

Yifei Zhang, Bo Pan, Chen Ling, Yuntong Hu, Liang Zhao

The deployment and application of Large Language Models (LLMs) is hindered by their memory inefficiency, computational demands, and the high costs of API inferences. Traditional distillation methods, which transfer the capabilities of LLMs to smaller models, often fail to determine whether the knowledge has been sufficiently transferred, potentially resulting in high costs or incomplete distillation. In this paper, we propose an Explanation-Guided LLMs Active Distillation (ELAD) framework that employs an active learning strategy to optimize the balance between annotation costs and model performance. To improve efficient sample selection, we introduce an explanation-guided sample selection method that identifies samples challenging its reasoning by exploiting uncertainties in explanation steps. Additionally, we present a customized LLM-annotated explanation revision technique where the teacher model detects and corrects flaws in the student model's reasoning. Our experiments across various reasoning datasets demonstrate that our framework significantly enhances the efficiency of LLM knowledge distillation.

link: http://arxiv.org/abs/2402.13098v1

A Microstructure-based Graph Neural Network for Accelerating Multiscale Simulations

J. Storm, I. B. C. M. Rocha, F. P. van der Meer

Simulating the mechanical response of advanced materials can be done more accurately using concurrent multiscale models than with single-scale simulations. However, the computational costs stand in the way of the practical application of this approach. The costs originate from microscale

Finite Element (FE) models that must be solved at every macroscopic integration point. A plethora of surrogate modeling strategies attempt to alleviate this cost by learning to predict macroscopic stresses from macroscopic strains, completely replacing the microscale models. In this work, we introduce an alternative surrogate modeling strategy that allows for keeping the multiscale nature of the problem, allowing it to be used interchangeably with an FE solver for any time step. Our surrogate provides all microscopic quantities, which are then homogenized to obtain macroscopic quantities of interest. We achieve this for an elasto-plastic material by predicting full-field microscopic strains using a graph neural network (GNN) while retaining the microscopic constitutive material model to obtain the stresses. This hybrid data-physics graph-based approach avoids the high dimensionality originating from predicting full-field responses while allowing non-locality to arise. By training the GNN on a variety of meshes, it learns to generalize to unseen meshes, allowing a single model to be used for a range of microstructures. The embedded microscopic constitutive model in the GNN implicitly tracks history-dependent variables and leads to improved accuracy. We demonstrate for several challenging scenarios that the surrogate can predict complex macroscopic stress-strain paths. As the computation time of our method scales favorably with the number of elements in the microstructure compared to the FE method, our method can significantly accelerate FE2 simulations.

link: http://arxiv.org/abs/2402.13101v1

On Generalization Bounds for Deep Compound Gaussian Neural Networks

Carter Lyons, Raghu G. Raj, Margaret Cheney

Algorithm unfolding or unrolling is the technique of constructing a deep neural network (DNN) from an iterative algorithm. Unrolled DNNs often provide better interpretability and superior empirical performance over standard DNNs in signal estimation tasks. An important theoretical question, which has only recently received attention, is the development of generalization error bounds for unrolled DNNs. These bounds deliver theoretical and practical insights into the performance of a DNN on empirical datasets that are distinct from, but sampled from, the probability density generating the DNN training data. In this paper, we develop novel generalization error bounds for a class of unrolled DNNs that are informed by a compound Gaussian prior. These compound Gaussian networks have been shown to outperform comparative standard and unfolded deep neural networks in compressive sensing and tomographic imaging problems. The generalization error bound is formulated by bounding the Rademacher complexity of the class of compound Gaussian network estimates with Dudley's integral. Under realistic conditions, we show that, at worst, the generalization error scales $\mathcal{O}(n\sqrt{\ln(n)})$ in the signal dimension and $\mathcal{O}(($Network Size$)^{3/2})$ in network size.

link: http://arxiv.org/abs/2402.13106v1

On the Stability of Gradient Descent for Large Learning Rate

Alexandru Cr■ciun, Debarghya Ghoshdastidar

There currently is a significant interest in understanding the Edge of Stability (EoS) phenomenon, which has been observed in neural networks training, characterized by a non-monotonic decrease of the loss function over epochs, while the sharpness of the loss (spectral norm of the Hessian) progressively approaches and stabilizes around 2/(learning rate). Reasons for the existence of EoS when training using gradient descent have recently been proposed -- a lack of flat minima near the gradient descent trajectory together with the presence of compact forward-invariant sets. In this paper, we show that linear neural networks optimized under a quadratic loss function satisfy the first assumption and also a necessary condition for the second assumption. More precisely, we prove that the gradient descent map is non-singular, the set of global minimizers of the loss function forms a smooth manifold, and the stable minima form a bounded subset in parameter space. Additionally, we prove that if the step-size is too big, then the set of initializations from which gradient descent converges to a critical point has measure zero.

link: http://arxiv.org/abs/2402.13108v1