Technical Review: Quantum Machine Learning

This document summarizes recent research on Quantum Machine Learning based on arXiv publications.

Paper 1: Allocating Variance to Maximize Expectation

Authors: Renato Purita Paes Leme, Cliff Stein, Yifeng Teng, Pratik Worah

Abstract: We design efficient approximation algorithms for maximizing the expectation of the supremum of families of Gaussian random variables. In particular, let $\mathrm{OPT}:=\max_{\sigma_1,\cdots,\sigma_n}\mathbb{E}\left[\sum_{j=1}^{m}\max_{i\in S_j} X_i\right]$, where X_i are Gaussian, $S_j\subset[n]$ and $\sum_i\sigma_i^2=1$, then our theoretical results include: - We characterize the optimal variance allocation -- it concentrates on a small subset of variables as $|S_j|$ increases, - A polynomial time approximation scheme (PTAS) for computing OPT when $m=1$, and - An $O(\log n)$ approximation algorithm for computing OPT for general $m>1$. Such expectation maximization problems occur in diverse applications, ranging from utility maximization in auctions markets to learning mixture models in quantitative genetics.

Link: http://arxiv.org/pdf/2502.18463v1

Paper 2: Scalable Equilibrium Sampling with Sequential Boltzmann Generators

Authors: Charlie B. Tan, Avishek Joey Bose, Chen Lin, Leon Klein, Michael M. Bronstein, Alexander Tong

Abstract: Scalable sampling of molecular states in thermodynamic equilibrium is a long-standing challenge in statistical physics. Boltzmann generators tackle this problem by pairing powerful normalizing flows with importance sampling to obtain statistically independent samples under the target distribution. In this paper, we extend the Boltzmann generator framework and introduce

Sequential Boltzmann generators (SBG) with two key improvements. The first is a highly efficient non-equivariant Transformer-based normalizing flow operating directly on all-atom Cartesian coordinates. In contrast to equivariant continuous flows of prior methods, we leverage exactly invertible non-equivariant architectures which are highly efficient both during sample generation and likelihood computation. As a result, this unlocks more sophisticated inference strategies beyond standard importance sampling. More precisely, as a second key improvement we perform inference-time scaling of flow samples using annealed Langevin dynamics which transports samples toward the target distribution leading to lower variance (annealed) importance weights which enable higher fidelity resampling with sequential Monte Carlo. SBG achieves state-of-the-art performance w.r.t. all metrics on molecular systems, demonstrating the first equilibrium sampling in Cartesian coordinates of tri, tetra, and hexapeptides that were so far intractable for prior Boltzmann generators.

Link: http://arxiv.org/pdf/2502.18462v1

Paper 3: K-LoRA: Unlocking Training-Free Fusion of Any Subject and Style LoRAs

Authors: Ziheng Ouyang, Zhen Li, Qibin Hou

Abstract: Recent studies have explored combining different LoRAs to jointly generate learned style and content. However, existing methods either fail to effectively preserve both the original subject and style simultaneously or require additional training. In this paper, we argue that the intrinsic properties of LoRA can effectively guide diffusion models in merging learned subject and style. Building on this insight, we propose K-LoRA, a simple yet effective training-free LoRA fusion approach. In each attention layer, K-LoRA compares the Top-K elements in each LoRA to be fused, determining which LoRA to select for optimal fusion. This selection mechanism ensures that the most representative features of both subject and style are retained during the fusion process, effectively balancing their contributions. Experimental results demonstrate that the proposed method effectively integrates the subject and style information learned by the original LoRAs, outperforming state-of-the-art training-based approaches in both qualitative and quantitative results.

Link: http://arxiv.org/pdf/2502.18461v1

Paper 4: DRAMA: Diverse Augmentation from Large Language Models to Smaller Dense Retrievers

Authors: Xueguang Ma, Xi Victoria Lin, Barlas Oguz, Jimmy Lin, Wen-tau Yih, Xilun Chen

Abstract: Large language models (LLMs) have demonstrated strong effectiveness and robustness while fine-tuned as dense retrievers. However, their large parameter size brings significant inference time computational challenges, including high encoding costs for large-scale corpora and increased query latency, limiting their practical deployment. While smaller retrievers offer better efficiency, they often fail to generalize effectively with limited supervised fine-tuning data. In this work, we introduce DRAMA, a training framework that leverages LLMs to train smaller generalizable dense retrievers. In particular, we adopt pruned LLMs as the backbone and train on diverse LLM-augmented data in a single-stage contrastive learning setup. Experiments show that DRAMA offers better multilingual and long-context capabilities than traditional encoder-based retrievers, and achieves strong performance across multiple tasks and languages. These highlight the potential of connecting the training of smaller retrievers with the growing advancements in LLMs, bridging the gap between efficiency and generalization.

Link: http://arxiv.org/pdf/2502.18460v1

Paper 5: LLM-Based Design Pattern Detection

Authors: Christian Schindler, Andreas Rausch

Abstract: Detecting design pattern instances in unfamiliar codebases remains a challenging yet essential task for improving software quality and maintainability. Traditional static analysis tools often struggle with the complexity, variability, and lack of explicit annotations that characterize real-world pattern implementations. In this paper, we present a novel approach leveraging Large

Language Models to automatically identify design pattern instances across diverse codebases. Our method focuses on recognizing the roles classes play within the pattern instances. By providing clearer insights into software structure and intent, this research aims to support developers, improve comprehension, and streamline tasks such as refactoring, maintenance, and adherence to best practices.

Link: http://arxiv.org/pdf/2502.18458v1