A Quantum Circuit-Based Compression Perspective for Parameter-Efficient Learning

Chen-Yu Liu, Chao-Han Huck Yang, Hsi-Sheng Goan, Min-Hsiu Hsieh

Abstract

Quantum-centric supercomputing presents a promising framework for large-scale hybrid quantum-classical computing. While quantum machine learning (QML) offers theoretical advantages across various applications, practical challenges—such as encoding large-scale data at the input stage and reliance on quantum resources during inference—limit its feasibility for tasks like fine-tuning large language models (LLMs). Quantum parameter generation, a novel QML approach, addresses these limitations by leveraging quantum neural networks (QNNs) to generate classical model parameters exclusively during training, eliminating the need for quantum hardware in inference. In this work, we introduce Quantum Parameter Adaptation (QPA) within the quantum parameter generation framework, integrating QNNs with a classical multi-layer perceptron (MLP) mapping model to optimize fine-tuning methods for LLMs. Using Gemma-2 and GPT-2 as case studies, QPA enables substantial parameter reduction in parameter-efficient fine-tuning techniques, such as Low-Rank Adaptation (LoRA), while maintaining or even enhancing performance in text generation tasks. Compared to standard LoRA, QPA significantly reduces the number of trainable parameters while preserving or slightly improving model performance. These results demonstrate QPA's ability to achieve efficient parameter reduction without compromising performance within the quantum parameter generation paradigm, highlighting the potential of quantum-enhanced parameter reduction as a scalable quantum-classical approach for fine-tuning LLMs while ensuring inference remains feasible on classical hardware.