

Title: Low-Latency Hardware Accelerator for Real-Time Neural Network-Based Qubit-State Discrimination

Abstract body (<1300 characters, including funding acknowledgement)

Fast, high-fidelity processing of qubit readout signals is essential for advancing quantum processors, particularly in supporting effective quantum error correction. While neural networks (NNs) have demonstrated significant potential in achieving high-fidelity qubit-state discrimination, their computational demands can lead to substantial processing latency. This work presents a hardware accelerator designed for low-latency implementation of an optimized fully connected neural network (FCNN) for real-time multi-qubit-state discrimination. Inspired by [1], our FCNN incorporates quantization-aware training and features a streamlined 4-layer architecture. This results in a 51.3x reduction in parameters, down to 31,879, with only a 0.3% decrease in discrimination accuracy, achieving a final accuracy of 90.86%. The hardware accelerator utilizes a scalable array of multipliers and adders for parallel processing, effectively minimizing latency while maintaining a compact silicon footprint. It is suitable for implementation on both FPGAs and ASICs.

[1] B. Lienhard et al., Phys. Rev. Applied 17, 014024 (2022)

We thank A*STAR, Singapore (C210917009) & NRF, Singapore (NR2021-QEP2-03-P07). B.L. is supported by Postdoc.Mobility Fellowship grant #P500PT_211060.