AQT Whitepaper

Łukasz Lew Yichi Zhang Vlad Feinberg

Shivani Agrawal Jihwan Lee Jonathan Malmaud

Lisa Wang Pouya Dormiani Reiner Pope

Abstract

To sustain AI innovation, minimizing training and serving costs for large models is crucial. Tensor operations, particularly matrix multiplications, are computationally intensive. Accelerators like TPU v5e and certain GPUs can perform INT8 or Float8 tensor operations 2× faster than BFloat16, but comprehensive software support is essential. We present Accurate Quantized Training (AQT), an open-source JAX library for quantization, boosting model training and serving speed in production without extensive tuning. AQT offers a flexible API for both production and research. INT8 experiments using TPU v5e demonstrate a 124% improvement in the training speed of a 16B language model in MaxText codebase and a 139% improvement for the 175B GPT-3 MLPerf 3.1 model.