**Title:** QuZO: Quantized Zeroth-Order Fine-Tuning for Large Language Models on Precision-Limited Platforms

**Author:**  Jiajun Zhou, Ziyue Liu, Yifan Yang, Yequan Zhao, Ngai Wong, Athanasios Mouchtaris, Siegfried Kunzmann, Zheng Zhang

**Abstract:**

Large language models (LLM) are often quantized to a low precision in order to improve inference speed and to reduce hardware cost (e.g., memory and energy) when deployed on resource-constraint computing platforms. Can we leverage a limited-precision inference engine to perform fine-tuning with minimal hardware adjustment? To achieve this goal, this paper presents a quantized zeroth-order method, QuZO, for fine tuning LLMs by using low-precision forward propagation only. QuZO is tailored for different precisions with integer (INT) and floating-point (FP) arithmetic to achieve high training accuracy in low-precision settings. Furthermore, the memory-efficient QuZO even achieves higher fine-tuning accuracy than first-order training when the precision is very low (e.g., below INT8), because QuZO can avoid the error-prone straight-through estimator that is widely used in quantized first-order training. Our method demonstrates competitive accuracy across various downstream tasks of LLMs. Our result shows that QuZO can reduce the memory consumption by 5x in LLaMa-7B fine-tuning compared to first-order optimization with the FSDP engine when using INT8 quantization.