TT-ZO: Tensor-Train Enhanced Zeroth-order Fine-tuning of Large Language Models

Yifan Yang, Kai Zhen, Athanasios Mouchtaris, Siegfried Kunzmann, Zheng Zhang

May 2024

Fine-tuning large language models (LLMs) has achieved remarkable performance in recent years across various natural language processing tasks, while the backpropagation graph consumes a substantial proportion of memory. To address this issue, the recently proposed Memory-efficient Zeroth-order (MeZO) methods attempt to fine-tune LLMs using only forward passes. Nonetheless, a significant performance drop and increased risk of divergence prevent widespread adoption of this method. In this paper, we propose the Tensor-Train enhanced zeroth-order (TT-ZO) framework, specifically designed to improve the performance and convergence of the ZO method. To further accelerate the forward pass of the Tensor-Train adaptation, we introduce a new contraction method for the tensorized layer. To address the divergence problem in large-scale model fine-tuning using ZO methods, we propose an adaptive query number schedule that ensures convergence with detailed theoretical analysis. Additionally, we present substantial experimental results on Roberta-large and Llama-2-7B models, demonstrating the effectiveness of our proposed ZO training framework in terms of accuracy, memory cost, and convergence speed. Further experimental results suggest that the proposed adaptive query schedule could be successfully applied to other ZO fine-tuning methods to improve their convergence and performance.