Efficient Fine-Tuning of Large Language Models Using LoRA and Quantization

Abstract

This paper explores parameter-efficient fine-tuning techniques for large language models (LLMs). Specifically, we investigate Low-Rank Adaptation (LoRA) combined with 4-bit quantization to reduce memory usage while maintaining performance. Experiments were conducted using the Mistral-7B model on a limited GPU environment.

1. Introduction

Large language models require substantial computational resources for training and inference. Full fine-tuning is often infeasible for researchers with limited hardware. LoRA enables fine-tuning by injecting trainable low-rank matrices into transformer layers while freezing the base model weights.

2. Methodology

We applied LoRA to the attention layers of Mistral-7B. The model was loaded using 4-bit NF4 quantization to reduce GPU memory consumption. Training was performed on a Colab T4 GPU with gradient accumulation. Evaluation metrics included perplexity and qualitative response analysis.

3. Results

The quantized LoRA model achieved comparable performance to full fine-tuning while using significantly less memory. GPU usage remained under 15GB, enabling training on consumer-grade hardware.

4. Conclusion

Parameter-efficient methods such as LoRA combined with quantization provide a practical solution for adapting large language models in resource-constrained environments. Future work includes testing on larger datasets and evaluating different quantization strategies.