LoRA and QLoRA: Efficient Fine-Tuning of Large Language Models

Omar Arnous Cellula Technologies
NLP Internship
omar99.arnous@gmail.com

September 20, 2025

Abstract

Large Language Models (LLMs) achieve state-of-the-art results in natural language processing but are expensive to fine-tune due to their billions of parameters. LoRA (Low-Rank Adaptation) and QLoRA (Quantized LoRA) offer parameter-efficient alternatives, reducing computational cost while maintaining strong performance. This paper introduces LoRA and QLoRA, compares their methodologies, and highlights their role in democratizing access to LLM fine-tuning.

1 Introduction

LLMs such as GPT and LLaMA have transformed NLP but require significant resources for training and fine-tuning. Traditional fine-tuning updates all parameters, demanding large memory and compute budgets. Parameter-efficient fine-tuning methods (PEFT) address this issue by updating only a small subset of parameters. Among these, LoRA [1] and QLoRA [2] are widely adopted for adapting LLMs to downstream tasks with modest hardware.

2 LoRA: Low-Rank Adaptation

LoRA introduces trainable low-rank matrices into the transformer architecture:

2.1 Methodology

Instead of updating all parameters, LoRA freezes the pretrained model weights and injects low-rank adapters into attention and feed-forward layers. Only these adapters are trained, drastically reducing the number of trainable parameters.

2.2 Efficiency

This approach reduces memory and compute requirements while retaining accuracy close to full fine-tuning. LoRA has been successfully applied in machine translation, dialogue systems, and instruction tuning.

3 QLoRA: Quantized LoRA

QLoRA extends LoRA by combining quantization and adapters:

3.1 Quantization

The base model is loaded in 4-bit precision, reducing GPU memory usage without significant performance loss.

3.2 Methodology

Like LoRA, QLoRA adds low-rank adapters, but it trains them on top of a quantized model. Gradient checkpointing and double quantization are used to further optimize efficiency.

3.3 Impact

QLoRA enables fine-tuning of models with up to 65B parameters on a single GPU, making large-scale adaptation accessible to researchers with limited resources.

4 Comparison

4.1 Approach

LoRA reduces trainable parameters via low-rank adaptation. QLoRA applies LoRA on top of quantized models, minimizing both memory and compute cost.

4.2 Efficiency

LoRA significantly reduces training costs compared to full fine-tuning. QLoRA improves on this by lowering memory footprint, allowing fine-tuning of very large models on consumer-grade hardware.

4.3 Use Cases

LoRA is ideal when moderate efficiency gains are sufficient. QLoRA is designed for extreme-scale models where memory constraints are critical.

5 Conclusion

LoRA and QLoRA represent a paradigm shift in adapting LLMs. LoRA reduces parameter updates while preserving performance, and QLoRA extends this to quantized settings, enabling fine-tuning of trillion-scale models on accessible hardware. Together, they make LLM research more inclusive and practical.

References

References

[1] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient fine-tuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023.