LoRA and QLoRA

Abdelrahman Ahmed

September 19, 2025

1 Introduction

Large Language Models (LLMs) are powerful but fine-tuning them is resource-intensive. LoRA (Low-Rank Adaptation) and QLoRA (Quantized LoRA) are efficient fine-tuning techniques designed to reduce resource usage while maintaining performance.

2 LoRA

Definition: LoRA introduces low-rank trainable matrices into existing layers, reducing the number of trainable parameters.

Advantages: Efficient fine-tuning without retraining all parameters, lower memory footprint, and faster adaptation.

Applications: Customizing large models for specific domains (legal, medical, chatbots).

3 QLoRA

Definition: QLoRA combines LoRA with quantization, reducing weights to 4-bit precision while preserving performance.

Advantages: Significant memory savings, enabling fine-tuning of very large models on a single GPU.

Applications: Fine-tuning models like LLaMA or Falcon on consumer-grade hardware.

4 Comparison

Feature	**LoRA**	**QLoRA**
Technique	Low-Rank Adaptation	Low-Rank + Quantization (4-bit precision)
Efficiency	Reduces trainable parameters	Reduces parameters + memory footprint
Hardware Needs	May require multiple GPUs for very large models	Can fine-tune very large models on a single GPU
Use Cases	Domain-specific adaptation	Large-scale fine-tuning on limited hardware

5 Conclusion

LoRA enables efficient fine-tuning with reduced parameters, while QLoRA extends this by adding quantization for even greater memory savings. Together, they make LLM fine-tuning more accessible and cost-effective.