Research Report: LoRA and QLoRA

Yousef Mahmoud Ali

September 16, 2025

Abstract

This report introduces two modern methods for fine-tuning large language models: LoRA and QLoRA. Both are designed to make fine-tuning more efficient, but they use different ideas. LoRA reduces the number of trainable parameters by using low-rank decomposition, while QLoRA adds quantization techniques to make training possible even on normal GPUs. This report explains their concepts, advantages, limitations, and applications in simple terms.

1 Introduction

Large language models like GPT-3 or PaLM are very powerful but also very expensive to fine-tune. They contain billions of parameters, which makes training require huge memory and expensive hardware. To solve this problem, researchers developed methods that allow fine-tuning without needing to update all parameters. Two of these methods are LoRA and QLoRA.

2 LoRA (Low-Rank Adaptation)

2.1 Main Idea

LoRA is based on the observation that weight updates during fine-tuning often lie in a low-dimensional space. Instead of training a full weight matrix, LoRA trains two much smaller matrices. This reduces the number of trainable parameters by a large factor.

2.2 Advantages

- Saves memory because only small matrices are trained.
- Can be applied to many layers of a transformer.
- Multiple LoRA adapters can be combined, which makes it flexible.

2.3 Limitations

- Accuracy might be slightly lower than full fine-tuning.
- Needs good choice of rank r to balance between efficiency and performance.

3 QLoRA (Quantized LoRA)

3.1 Main Idea

QLoRA builds on LoRA by adding **quantization**. Quantization means representing weights in lower precision (like 4-bit instead of 16-bit). This makes it possible to fine-tune very large models (up to 65B parameters) on a single modern GPU.

3.2 Key Techniques

- **4-bit NormalFloat (NF4):** A quantization method designed for normally distributed weights.
- **Double Quantization:** Even quantizes the constants used in quantization for more memory savings.
- **Paged Optimizers:** Uses smart memory management to avoid crashes from memory spikes.

3.3 Advantages

- Makes it possible to fine-tune huge models on consumer hardware.
- Uses very little GPU memory compared to normal fine-tuning.
- Maintains high accuracy despite heavy compression.

3.4 Limitations

- Quantization can sometimes slightly reduce precision.
- More complex setup compared to LoRA.

4 Comparison of LoRA and QLoRA

Feature	LoRA	QLoRA
Main Goal	Reduce trainable parameters	Fine-tune huge models on small GPUs
Key Idea	Low-rank decomposition	LoRA + quantization
Memory Usage	Medium	Very low
Hardware Needs	Moderate GPU	Even consumer GPUs

Table 1: Comparison between LoRA and QLoRA

5 Applications

5.1 LoRA Applications

- Domain adaptation (e.g., medical or legal text).
- Personalized chatbots and assistants.
- Multi-task learning using different adapters.

5.2 QLoRA Applications

- Fine-tuning 65B+ parameter models on a single GPU.
- Research for people with limited hardware.
- Making large models accessible to smaller companies and universities.

6 Conclusion

LoRA and QLoRA are two important steps in making large language models easier and cheaper to fine-tune. LoRA focuses on reducing trainable parameters through low-rank adaptation, while QLoRA goes further by adding quantization to make massive models possible to train on regular GPUs. Both methods are now widely used in the NLP community because they balance efficiency with performance.