Generative AI
Assignment 3

Name: Alaiba Nawaz
Roll No: 21L-5650
Section: BDS-8A

1. Full Fine-Tuning

- **Accuracy:** 0.8930
- **Trainable Parameters:** 124,647,170
- **Training Time:** 87.32 seconds
- **GPU Memory Usage:** 3752 MB

Explanation:
A complete update of full parameter model values through Full Fine-Tuning allows for maximum adaptability to new tasks. The method produces greater results shown in this example but requires enormous computational capabilities and substantial training time and memory resources. Having plenty of resources coupled with a desire for utmost accuracy makes Full Fine-Tuning the best choice.

2. LoRA Fine-Tuning

- **Accuracy:** 0.8165
- **Trainable Parameters:** 1,034,498
- **Training Time:** 63.95 seconds
- **GPU Memory Usage:** 2206 MB

Explanation:
The model receives LoRA (Low-Rank Adaptation) adaptable low-rank matrices which maintain original weight values unchanged. With this approach the quantity of parameters becomes significantly smaller when training while both speed and consumed memory increase but accuracy shows a moderate decrease. The speed and efficiency characteristics of LoRA accommodate applications that require some reduction in accuracy compared to normal operation.

3. QLoRA Fine-Tuning

- **Accuracy:** 0.7890
- **Trainable Parameters:** 1,034,498
- **Training Time:** 89.54 seconds
- **GPU Memory Usage:** 820 MB

Explanation:
QLoRA implements weight compression by quantization on top of LoRA efficiency methods with 4-bit precision levels. The reduction in GPU memory needs enables training of large models through consumer-grade hardware systems. The process of quantization does slightly damage numerical precision thus likely causing the accuracy drop and extended training durations due to dequantization-related computation delays.

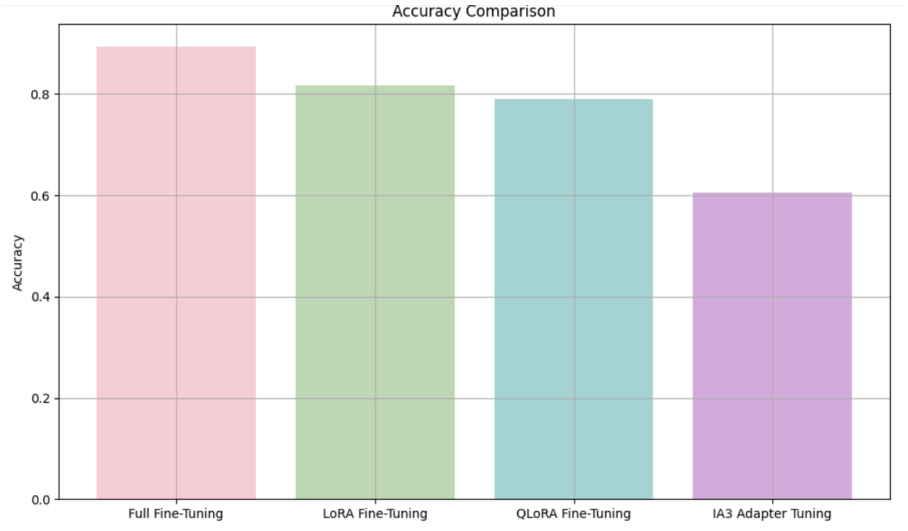
4. IA3 Adapter Tuning

- **Accuracy:** 0.6045
- **Trainable Parameters:** 665,858
- **Training Time:** 61.21 seconds
- **GPU Memory Usage:** 2608 MB

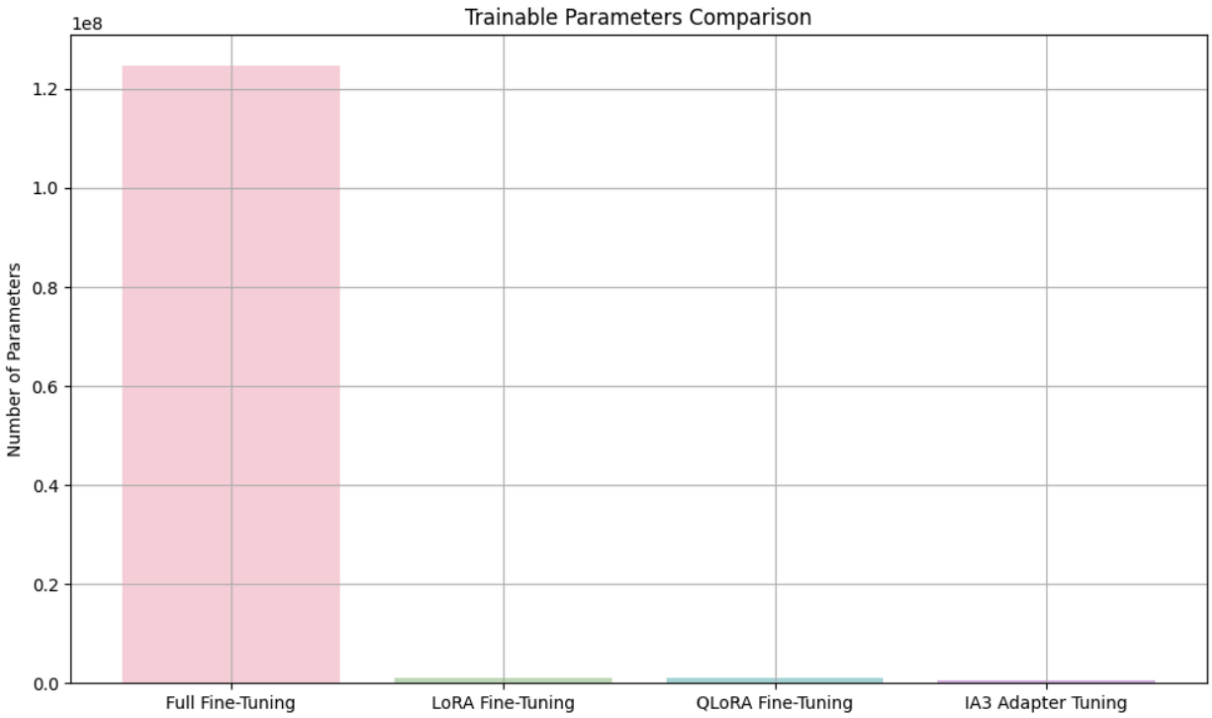
Explanation:
The lightweight adapters of IA3 get inserted into particular layers of the model (attention and feedforward) while maintaining the other parts as frozen. Because its adjustments touch only a few parameters the system has short training times yet achieves minimum accuracy as learning stays restricted to specific tasks. The limited modifications work well for constrained tasks yet show restricted application range between different problem domains.

BAR CHARTS

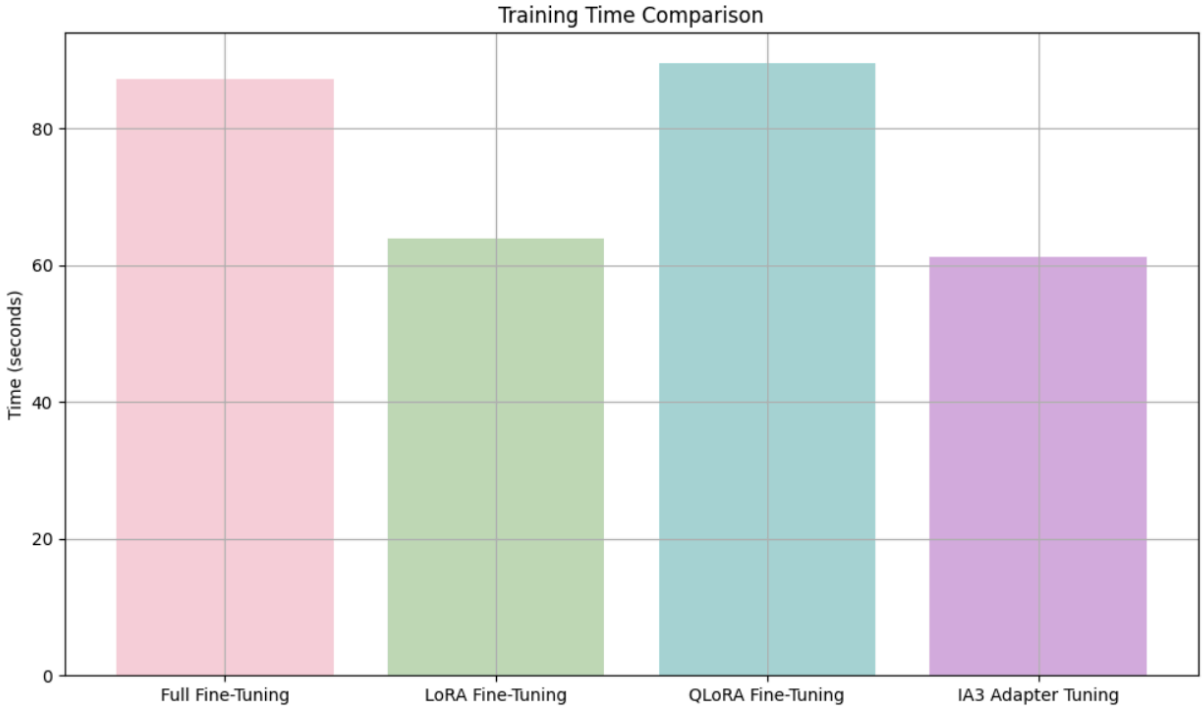
Accuracy Comparisons:



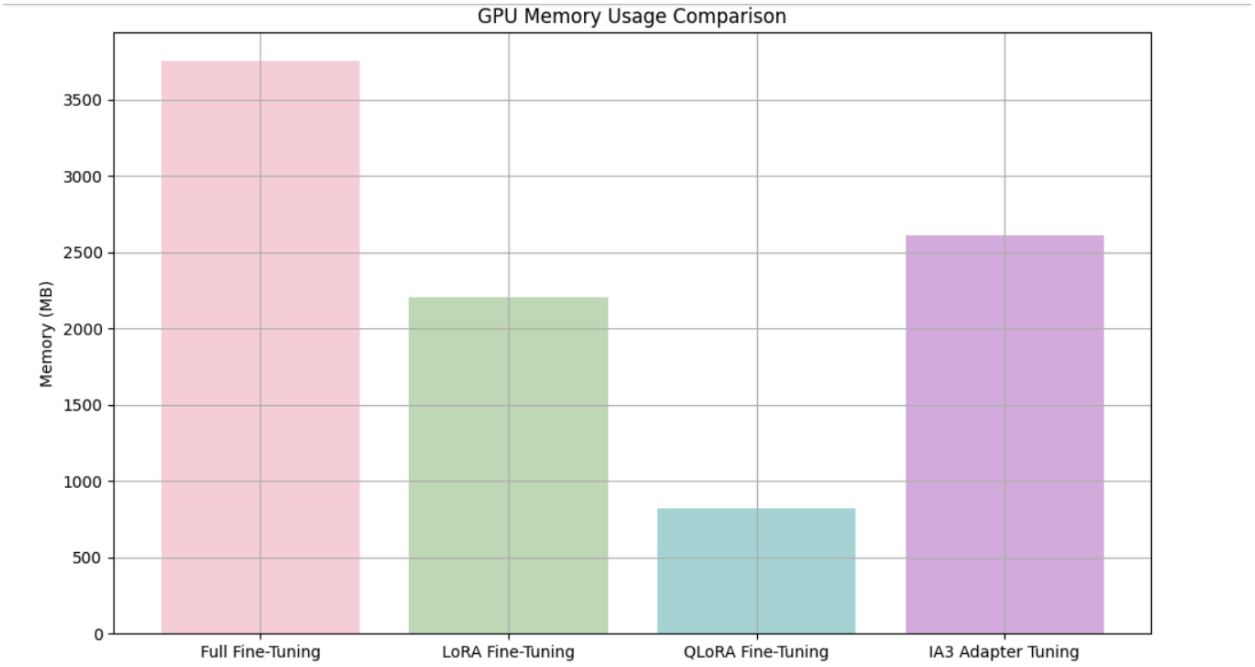
Trainable Parameters Comparison :



Training Time Comparison



GPU Memory



Conclusion

- **Full Fine-Tuning** is best for maximum accuracy but is computationally expensive.
- **LoRA** and **QLoRA** are efficient alternatives that achieve strong performance with a fraction of the parameters.
- **IA3** is the most lightweight but shows a clear trade-off in performance.
  The differences arise from how much of the model is being updated, how model precision is managed (as in QLoRA), and where the task-specific learning occurs (adapters vs. full model layers).