LoRA Fine-Tuning Lab

Aim

Fine-tune GPT-2 efficiently using low-rank adapters (LoRA).

Theory

LoRA trains only low-rank matrices (A,B) while freezing base weights. $\Delta W = BA$, reducing VRAM usage.

Algorithm

1. Load GPT-2 tokenizer & model.

2. Tokenize dataset.

3. Apply LoRA config (r, alpha, dropout).

4. Train with Trainer.

5. Generate samples.

Pseudocode

```
model = get_peft_model(base, lora_cfg)
trainer.train()
out = model.generate()
```

Results

Small adapter file; model adapts to dataset style.

Conclusion

LoRA gives high-quality fine-tuning with minimal compute.