Fine-Tuning a Language Model with LoRA (PEFT)

Author: Touseef Ahmed

Contact: touseefahmed00710@gmail.com

1. Model Used

We used the 'google/flan-t5-small' model from Hugging Face, which is a fine-tuned variant of T5 (Text-to-Text Transfer Transformer). This model is suitable for sequence-to-sequence tasks and is lightweight enough to be trained on a Google Colab T4 GPU (~15GB VRAM).

(Advantage: Small size, faster training).

! Limitation: Not as powerful as 7B+ models like Falcon or Mistral

2. Dataset Used

1. We fine-tuned the model on the Alpaca Instruction-Tuning Dataset, originally released by Stanford and hosted on Hugging Face.
   Source: https://huggingface.co/datasets/tatsu-lab/alpaca

   Dataset Fields:
   - instruction: what the user wants the model to do
   - input: optional context
   - output: expected response

3. LoRA Configuration Used

We applied Parameter-Efficient Fine-Tuning (PEFT) using the LoRA (Low-Rank Adaptation) technique via the `peft` library.

Parameter	Value
R	8
lora_alpha	32
lora_dropout	0.1
target_modules	["q", "v"]
Bias	"none"
task_type	SEQ_2_SEQ_LM

Only adapter layers are trained → Reduces GPU usage & speeds up training

➔ Base model weights remain frozen

4. Sample Test Inputs & Model Outputs

After training for 3 epochs, the fine-tuned model was tested on a few instruction-based prompts.

Input Prompt	Model Output
Translate English to French: I love machine learning.	J'aime l'apprentissage automatique
What is the capital of Pakistan?	Islamabad
Summarize the following: Large language models are transforming AI development.	Large models are revolutionizing AI.

➔ Outputs were mostly coherent and aligned with the instruction

➔Slight grammatical oddities on long prompts, due to model size

5. Challenges Faced

Tokenizer mismatch: The SFTTrainer class threw an error due to the unsupported tokenizer argument.

➔ Fixed by removing the argument when using SFTTrainer.

- Limited GPU Memory: Larger models like Falcon-7B or Mistral could not run on Colab T4.

➔ Resolved by selecting flan-t5-small for compatibility.

- Manual preprocessing: The dataset required careful concatenation of instruction and input before tokenization.

➔ Final Remarks

This activity demonstrates how instruction-tuned datasets and PEFT techniques like LoRA can be combined to fine-tune LLMs on consumer-grade GPUs. Despite hardware limitations, meaningful outputs were achieved using efficient methods.