ADL HW3 Report

NTU CSIE, R12922051
資工碩一 陳韋傑

Q1: LLM Tuning

Parameters

- Bits: 8
- Lora Rank: 64
- Learning Rate: 3e-5
- Source Max Length: 384
- Target Max Length: 128
- Batch Size: 16
- Max Step: 1000 (1.8 Epoch)
- Train Datapoints: 9900
- Validation Datapoints: 100
- Optimizer: paged_adamw_32bit

I use QLoRA, a PEFT method, to tune Taiwan-LLaMA. For every linear layers, I add a LoRA layer and only finetune them.

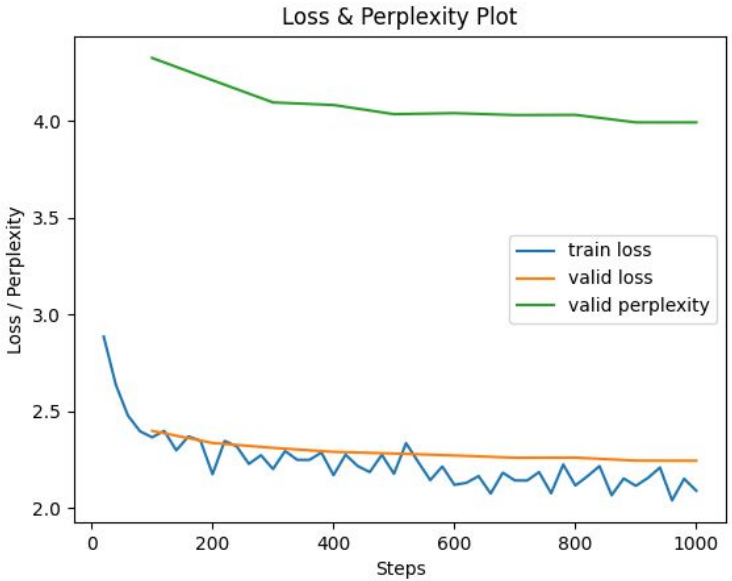
Q1: LLM Tuning

Total Steps: 1000

Epoch: 1.8

At Step 1000:

- Train Loss: 2.0906
- Valid Loss: 2.2455
- Perplexity: 3.9941



Q2: LLM Inference Strategies

Setting	Perplexity
Zero-shot on Taiwan-LLaMA	5.0744
Few-shot on Taiwan-LLaMA (Shot=1)	1.1323
Few-shot on Taiwan-LLaMA (Shot=3)	1.1756
Few-shot on Taiwan-LLaMA (Shot=5)	1.1775
QLoRA on Taiwan-LLaMA	3.9941

Q2: LLM Inference Strategies

Settings:

- Zero-Shot
  - Same prompt as QLoRA (Since there are no example provided)
- Few-Shot
  - Refined prompt for model understanding and examples
  - Different numbers of example are provided, example are chosen in the relative position in the dataset (previous data pair)

Comparison

Between three methods, Few-Shot in-context learning performs the best (no matter how many examples are used), while QLoRA performs worse than Few-Shot but better then Zero-Shot.

Q2: LLM Inference Strategies - Prompt

- Zero-Shot & QLoRA
  你是一個精通現代中文與文言文的大師, 以下是用戶和你之間的對話。你的目標是對用戶的問題提供有用、精確且簡潔的回答。USER: {instruction} ASSISTANT:

- Few-Shot
  你是一個精通現代中文與文言文的大師, 以下是用戶和你之間的對話。你的目標是對用戶的問題提供有用、精確且簡潔的回答。以下為幾個翻譯的正確例子\n (USER:{example} ASSISTANT:{answer}\n) x N 以下為你需要翻譯的句子, 請根據前面提供的正確結果進行翻譯\n USER:{instruction} ASSISTANT:

Reference

- https://github.com/MiuLab/Taiwan-LLaMa
- https://github.com/artidoro/qlora