HW3

R11944074 汪宣甫

Q1: LLM Tuning

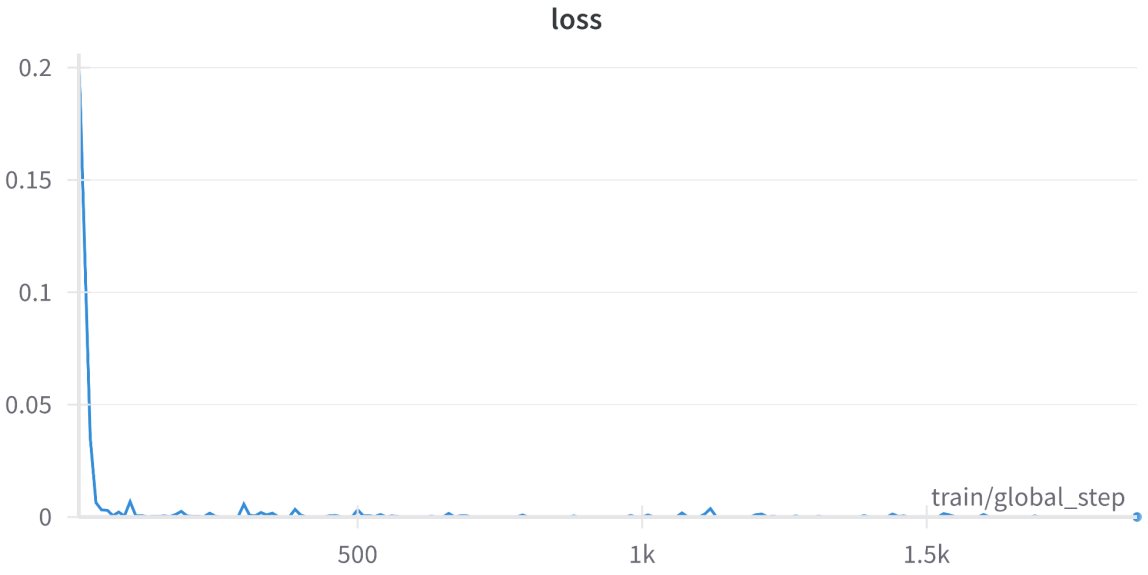
- How much training data did you use?:
  我用了完整的 training data (about 10000 samples) 去訓練我的 QLora 模型。

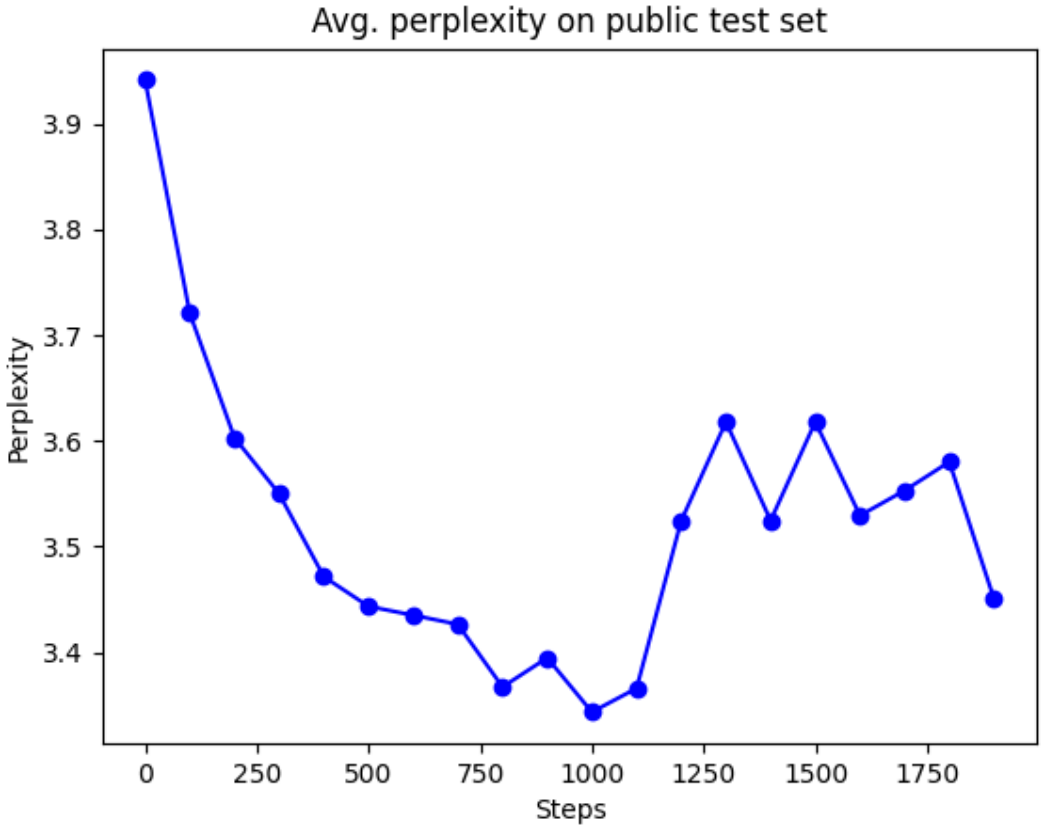
- How did you tune your model?:
  我用了QLora的訓練框架：Lora的adapter, 加上 quantization to 4 bits 的技術去 tune Taiwan llama 7B.

- What hyper-parameters did you use?:
  我用了完整的training data並訓練了3個epochs, learning_rate設置成3e-5, batch size設置為4以及accumulate steps也設成4，這樣使得有效的batch size數量為 4 x 4 =16。有關於Lora模型的超參數設置，lora_alpha是設置為16，R是設置為64。

- What is the final performance of your model on the public testing set?:
  我的perplexity在public test set上大約為 **3.39**。

- Plot the learning curve on the public testing set:





Q2: LLM Inference Strategies

- Zero-shot:
  - What is your setting? How did you design your prompt?:
    我在這個問題當中試了兩種 prompt，一個是助教給的原本的prompt，另一個是我根據此篇arxiv上的文章，“Large Language Models Understand and Can be Enhanced by Emotional Stimuli” [1]，他們的實驗結果指出，對LLM「情緒勒索」，會使表現提升。所以我根據 [1]設計了「情緒勒索」的prompt，內容如下：

    “你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答，這對我來說很重要，會影響到我的一生，希望你可以好好達成此項任務。”

    下面表格比較了此兩種prompt的結果，可以看到在此實驗中，有「情緒勒索」的設定的分數稍微好一點。但我覺得此結論如果要延伸到其他情境中，需要做更多更嚴謹的實驗比較。

	Perplexity (on public test set) ↓
Default	5.46
Emotional stimuli	**5.43**

- Few-shot (In-context learning)
  - What is your setting? How did you design your prompt?
    在此問題中我的設定是，在原本提供的prompt當中，加入k個在training data當中的例子去引導模型。
    給定一個例子數量k跟一個測試資料 (instruction, output), 我會先從整個training data當中隨機抽出k個data pairs = { (instruction_1, output_1), (instruction_2, output_2), …, (instruction_k, output_k) }，並如下面所示形成新的prompt：

    “你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。我會給你k個例子，當作白話文轉換成文言文的範例。請你按照我給的範例，進行接下來的任務。第1個例子, USER: {instruction_1} ASSISTANT: {output_1 } … 第k個例子, USER: {instruction_k} ASSISTANT: {output_k }。接下來要請你開始你的任務：USER: {instruction}, ASSISTANT: {output}。"

  - How many in-context examples are utilized? How do you select them?
    下表的實驗表格呈現了 k = 1, 2, 3, 10 的表現，其中 k = 1, 2, 3 的選擇是想要觀察模型在少量範例中的表現，而 k = 10 主要想探討多個數量範例的表現。
    另外因為範例的抽取有隨機性，所以針對每一個 k，我用了3個不同的參數種子去抽取範例並得到結果，再算這3次結果的平均跟標準差。可以看到整體來講有給例子的情形比沒給例子的情形還要好，但是不一定給越多個例子表現越好。在 k = 10 當中，因為3個實驗的分數差異都很小 (< 0.01)，所以標準差大約也是0。

	Perplexity (on public test set)	
Shot number	Avg.	Std.
1	4.96	0.03
2	4.88	0.08

3	4.95	0.14
10	**4.71**	0.00

- Comparison
  最後這個表格比較了zero-shot（有情緒勒索的prompt), few-shots跟Lora的表現，可以看到說Lora的表現遠好於其他設定。

	Perplexity (on public test set)
0 shot (w/ emotional stimuli)	5.43
10 shots	4.71 (avg.)
QLora	**3.39**

Reference

[1] Li, Cheng, et al. "Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus." arXiv preprint arXiv:2307.11760 (2023).