CSIE5431 Applied Deep Learning

Homework 3 Report

Name: 高榮浩 **ID:** R12922127

1. LLM Tuning

- **Describe**

■ Training Data Volume Analysis

While I specified the number of epochs as 2, I only utilized 7/12 of the training data, approximately 5800 instances. This decision was based on the observation that the optimal adapter checkpoint consistently emerged around the 7/12 epoch mark. By doing so, each piece of training data is encountered either once or not at all, effectively mitigating overfitting and reducing overall training time.

■ Model Tuning Methodology

I employed the yentinglin/Taiwan-LLM-7B-v2.0-chat as the base model, a fully parameter fine-tuned model derived from Meta/LLaMa-2 tailored for Traditional Mandarin applications. This base model underwent pretraining on over 30 billion tokens and instruction-tuning on more than 1 million instruction-following conversations in Traditional Mandarin.

For fine-tuning, I utilized the QLoRA method, which optimizes memory usage in LLMs by employing 4-bit quantization for weight representation, thereby compressing the model. Simultaneously, computations were conducted using 16-bit float precision.

Quantization involves reducing the number of bits used to represent each weight in the model, effectively lowering memory requirements. In this instance, 4-bit quantization was implemented, signifying that each weight is represented using only 4 bits.

Moreover, the use of 16-bit float precision strikes a balance between model performance and memory efficiency. While 16-bit precision reduces the memory footprint compared to the standard 32-bit precision, it still maintains sufficient numerical precision for the model's computations.

This tuning methodology, a combination of quantization and low-rank approximation (QLoRA), enhances the model's memory efficiency, making it particularly well-suited for applications where memory constraints are a critical consideration.

Hyperparameter Configuration Details

I've used the hyperparameters outlined in the table below.

Model	yentinglin/ Taiwan-LLM-7B-v2.0-chat
Optimizer	AdamW
Epoch	2
Learning Rate	2e-4
Batch Size	64
`per_device_train_batch_size`	4
`gradient_accumulation_steps`	16
`max_seq_length`	512
`peft_lora_r`	64
`peft_lora_alpha`	16
`peft_lora_dropout`	0.05
`fp16`	True

Initially, all hyperparameters are set to their default values. Subsequently, through a series of experiments involving meticulous performance monitoring, I manually fine-tune these hyperparameters to enhance performance. The hyperparameters listed in the table above represent the final configuration that yielded the desired results.

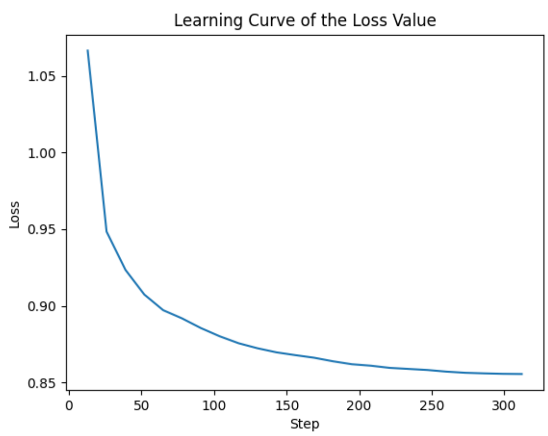
Performance

Model Performance on Public Testing Set

The table below illustrates the final performance on the public testing set.

	Mean Perplexity
LoRA	3.8255

Public Testing Set Learning Curve Analysis



2. LLM Inference Strategies

- **Zero-Shot**

  - **Prompt Design and Experimental Settings**

I employed `ppl.py` directly without loading LoRA, utilizing the default prompt. The prompt I utilized is as follows: `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。USER: {instruction} ASSISTANT:`.

- **Few-Shot (In-context Learning)**

  - **Prompt Design and Experimental Settings**

I utilized `ppl.py` directly without loading LoRA and made adjustments to the default prompt by incorporating examples through the revision of the `get_prompt` function in `utils.py`. The modified prompt is as follows: `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。USER: 翻譯成文言文：\n 雅裏惱怒地說： 從前在福山田獵時，你誣陷獵官，現在又說這種話。\n 答案： ASSISTANT: 雅裏怒曰： 昔畋於福山，卿誣獵官，今復有此言。USER: 辛未，命吳堅為左丞相兼樞密使，常楙參知政事。\n 把這句話翻譯成現代文。 ASSISTANT: 初五，命令吳堅為左承相兼樞密使，常增為參知政事。USER: {instruction} ASSISTANT:`.

  - **In-Context Examples Selection and Utilization**

I included two examples in the default prompt to enhance the model's performance. The first example is: `USER: 翻譯成文言文：\n 雅裏惱怒地說： 從前在福山田獵時，你誣陷獵官，現在又說這種話。\n 答案： ASSISTANT: 雅裏怒曰： 昔畋於福山，卿誣獵官，今復有此言。`. This example involves translating modern Chinese into classical Chinese. The second example is: `USER: 辛未，命吳堅為左丞相兼樞密使，常楙參知政事。\n 把這句話翻譯成現代文。 ASSISTANT: 初五，命令吳堅為左承相兼樞密使，常增為參知政事。`. In this case, the example requires translating classical Chinese into modern Chinese. I selected these two examples intentionally to cover both directions of translation: from classical to modern and from modern to classical Chinese. This diverse set of examples is designed to improve the model's overall performance in handling various translation tasks.

- **Comparison**

Comparative Analysis: Zero-shot, Few-shot, and LoRA Results

The table below presents the performance on the public testing set, employing zero-shot, few-shot, and LoRA strategies.

	Mean Perplexity
Zero-shot	5.4607
Few-shot	4.7259
LoRA	3.8255

In the context of zero-shot learning, the model is tasked with generating results without any prior exposure to examples, presenting a notably more challenging scenario compared to few-shot learning. In a few-shot setting, the model is afforded the opportunity to examine a limited number of examples before generating results. Leveraging the LoRA strategy allows the model to assimilate more data than in a few-shot scenario, consequently yielding superior performance. In this instance, I utilized 7/12 of the training data to train the adapter. To sum up, the results presented in the table align with expectations.