ALD2023-HW3 Report

Q1: LLM Tuning

Describe

How much training data did you use?

All of them (10000).

How did you tune your model?

I use Axolotl, and I found that the performance of default setting is enough to pass base line, but I still conducted some experiment of tuning hyperparameter.

Batch size: smaller batch size leads to better performance.

Lora_r: the performances of setting lora_r as 4, 8, 16, 32 are almost similar, but the smaller the lora_r is, the less time it need for training.

Epoch: I found that more epochs don't lead to better performance, it will overfit.

What hyper-parameters did you use?

Lora_r = 4

Sequence_length = 2048

Batch = 4

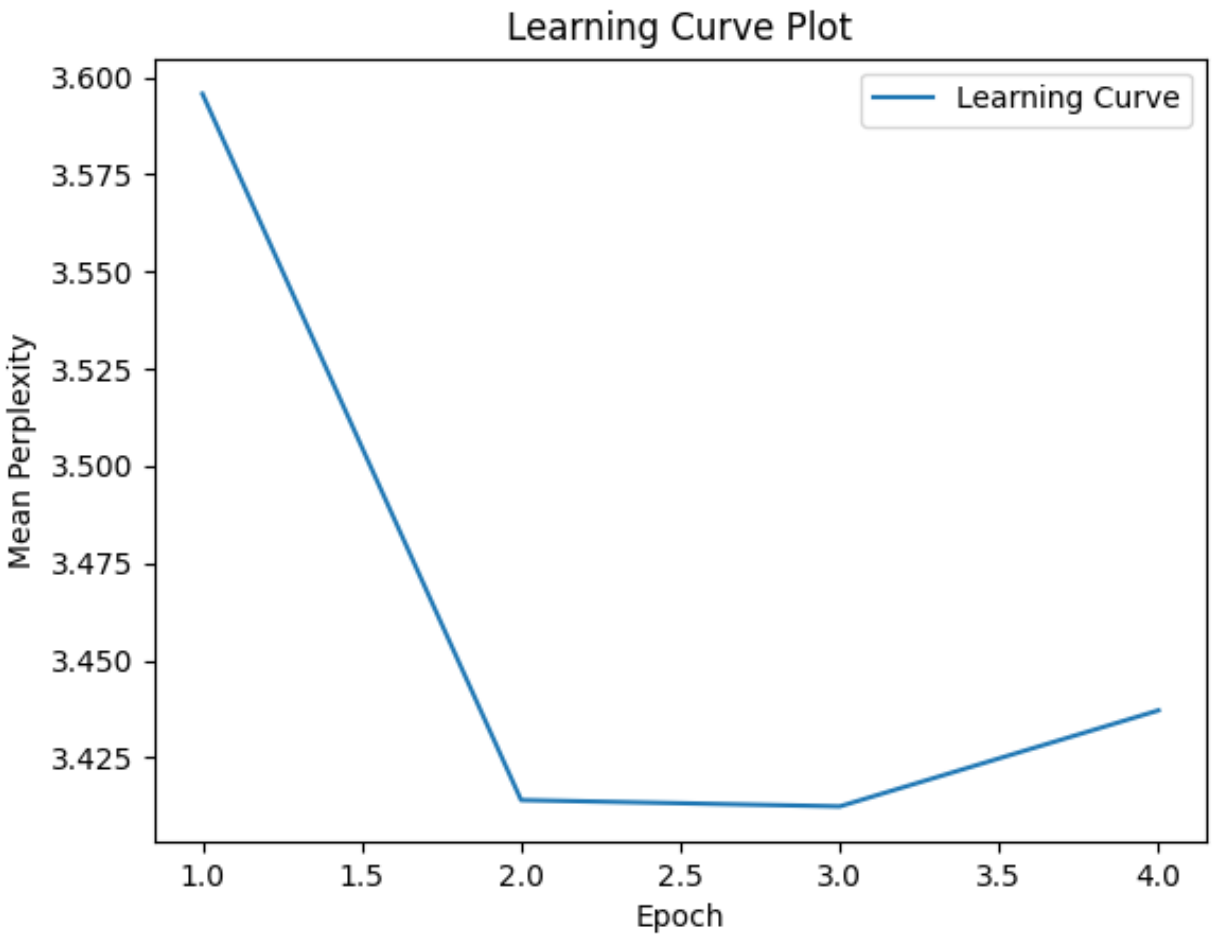
Epoch = 4

Show your performance:

What is the final performance of your model on the public testing set?

Mean perplexity: 3.4371703333854677(this is one of the result), parameters are shown above.

Plot the learning curve on the public testing set



Q2

Zero-Shot

Prompt: "你是一名中文系教授，以下是學生對文言文或現代文的問題。請提供他們有用且精確無誤的答案。學生: {instruction} 教授:"

Mean perplexity: 4.997839138984681

Few-Shot

你是一名中文系教授，以下是學生對文言文或現代文的問題。請提供他們有用且精確無誤的答案。\
學生一：「第二年召迴朝廷，改任著作佐郎，直史館，改任左拾遺。翻譯成文言文：」\
教授：明年召還，改著作佐郎，直史館，改左拾遺。\
學生二：「文言文翻譯：中宗與庶人嘗因正月十五日夜幸其第，賜賚不可勝數。」\
教授： 唐中宗與韋庶人曾經在正月十五日夜到韋安石的宅第，並賜賞給他不可勝數的財物。\
學生：「{instruction}」 \
教授："

Mean perplexity: 5.240043347835541

I select 2 examples, one is translation from classic chinese to plain chinese, one is plain chinese to classic chinese. I think this giving it examples of both tasks may lead to better performance.

Comparison

LORA: 3.4371703333854677

Zero-Shot: 4.997839138984681

Few-Shot: 5.240043347835541

LORA is finetuned on 10000 data to perform transformation between classic chinese and plain chinese. Zero-Shot is a model without finetune, so I think the performance would be the worst. Few-Shot has a few examples, maybe the model can learn something from the examples, so the performance may be better than Zero-Shot.

The performance of LORA is the best, this is imaginable.

But the performance of Zero-Shot is better than Few-Shot, this is out of my imagination.

I've discussed with others, and it seemed that the performance of my Zero-Shot is way too good.

Q3: Bonus

Use Lora to tune model.

The setting is same as qlora:

Lora_r = 4

Sequence_length = 2048

Batch = 4

Epoch = 4

Mean perplexity: 3.408150182723999

The training time of lora is longer than qlora, but the performance is slightly better.

I think the reason is that lora is the original version and qlora is a light version which consumes less time and memory but performance is almost the same.