Q and A in completion HW1

1. Q: bpe 算法的代码是否需要粘贴至 report。A: 应该只需要呈现核心的代码和思路
2. 上一学期part 3是先全量微调，然后 lora 微调。本次作业中先调试 lora 的最优超参数，然后更换数据集，再使用 lora 微调。

Whole Implementation stack

Part 1：Tokenizer

1. Youtube video: [Let’s build the GPT Tokenizer](https://www.youtube.com/watch?v=zduSFxRajkE)
2. Learn code in github repository like: <https://github.com/openai/tiktoken>
3. Finish API about tokenizer class using bpe.
4. Train a tokenizer.
5. Answer questions.

Part 2：Pretrain (Implementation LLM)

1. YouTube Video: [Let’s reproduce GPT-2 (124M)](https://www.youtube.com/watch?v=l8pRSuU81PU), Take notes(needed in report)
2. Research paper or blogs like: [The Annotated Transformer](https://nlp.seas.harvard.edu/annotated-transformer/), [Attention Is All You Need](https://arxiv.org/abs/1706.03762), [Language Models are Unsupervised Multitask Learners (GPT-2 paper)](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)
3. Init a git to control edition, organize “commit” to show implementation story.(also needed in report)

Part 3：Fine-tuning

1. Learn foundation knowledge on Hugging Face, like [Transformers](https://huggingface.co/docs/transformers/index), [LoRA paper](https://arxiv.org/abs/2106.09685).
2. Download code, train and change super parameter.
3. Commit running recordings, form: ipynb.
4. Implement lora.py code.
5. Change dataset to AIpaca and fine-tuning using LoRA.
6. Report loss and output of the model.