实验二 基于 chatglm3-6b 模型的 lora 方法微调

一、实验内容

ChatGLM3 是智谱 AI 和清华大学 KEG 实验室联合发布的对话预训练模型。ChatGLM3-6B 是 ChatGLM3 系列中的开源模型，在保留了前两代模型对话流畅、部署门槛低等众多优秀特性的基础上，ChatGLM3-6B 引入了更强大的基础模型；更完整的功能支持；更全面的开源序列，并且对学术研究完全开放。

LoRA（Low-Rank Adaptation of Large Language Models，大型语言模型的低秩适应）是微软研究员提出的一种新颖技术，旨在解决微调大型语言模型的问题。LoRA 提议冻结预训练模型的权重，并在每个 Transformer 块中注入可训练层（称为秩分解矩阵）。这大大减少了可训练参数的数量和 GPU 内存需求，因为大部分模型权重不需要计算梯度。通过专注于大型语言模型的 Transformer 注意力块，LoRA 的微调质量与完整模型的微调相当，同时速度更快，计算需求更低。

本次实验参考 ChatGLM3 官方提供的代码以及微调手册，带领大家学习 Lora 微调的方法流程，为大家学习大语言模型打好基础。

参考文件：

ChatGLM3 官方地址：THUDM/ChatGLM3: ChatGLM3 series: Open Bilingual Chat LLMs | 开源双语对话语言模型 (github.com)
Lora 原文：https://arxiv.org/abs/2106.09685

Step1 打开终端，下载代码

打开 **Terminal** 终端，把 ChatGLM3.zip 文件上传到服务器，并解压

```
unzip ChatGLM3.zip
cd ChatGLM3
```

Step2 创建虚拟环境并激活(2 分)

虚拟环境命名使用：姓名首字母小写+学号，如:cqf1190202318

```
conda create -n cqf1190202318 python=3.10
conda activate cqf1190202318
```

创建成功后给出截图。如：

```
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
#
# To activate this environment, use
#
#     $ conda activate cqf1190202318
#
# To deactivate an active environment, use
#
#     $ conda deactivate

root@dsw-346519-b567d758-fw7cz:/mnt/workspace/ChatGLM3# conda activate cqf1190202318
(cqf1190202318) root@dsw-346519-b567d758-fw7cz:/mnt/workspace/ChatGLM3#
```

如果出现以下报错，可能是因为 conda 配置中的某些通道无法访问或无效。可以尝试执行以下步骤来更换通道，再次创建：

```
root@dsw-338136-6fc869d6fd-vpdpt:/mnt/workspace/ChatGLM3# conda create -n chatglm3-6b python=3.10
Retrieving notices: ...working... done
Collecting package metadata (current_repodata.json): failed

UnavailableInvalidChannel: HTTP 404 NOT FOUND for channel anaconda/pkgs/main <http://mirrors.aliyun.com/anaconda/pkgs/main>

The channel is not accessible or is invalid.

You will need to adjust your conda configuration to proceed.
Use `conda config --show channels` to view your configuration's current state,
and use `conda config --show-sources` to view config file locations.
```

conda config --show channels

conda config --remove channels defaults

conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free/

conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/main/

conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud//pytorch/

conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/conda-forge/

Step3 使用 pip 安装模型所需依赖

pip install -r requirements.txt

安装成功如下图：

Step4 ChatGLM3-6b 模型下载

阿里云不支持 huggingface，可以到 modelscope 下载,把模型放在根目录下。

cd /mnt/workspace

git lfs install

git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git

Step5 验证模型命令行对话 Demo (3 分)

打开 ChatGLM3/basic_demo/cli_demo.py 文件，修改模型路径 MODEL_PATH 为刚才下载模型的绝对路径。

python cli_demo.py

出现下列对话功能代表运行成功，需截图。

```
(chatglm3-6b) root@dsw-337413-5b4dbc6b9-14sdm:/mnt/workspace/ChatGLM3/basic_demo# python cli_demo.py
Setting eos_token is not supported, use the default one.
Setting pad_token is not supported, use the default one.
Setting unk_token is not supported, use the default one.
Loading checkpoint shards: 100%|                                                | 7/7 [00:04<00:00,  1.60it/s]
欢迎使用 ChatGLM3-6B 模型，输入内容即可进行对话，clear 清空对话历史，stop 终止程序

用户：11

ChatGLM：I'm sorry, but I'm not sure what you mean by "11." Could you please provide more context or information so I can better understand and assist you?

用户：你好

ChatGLM：你好！有什么我可以帮助你的吗？

用户：谁是最帅的男人

ChatGLM：关于最帅的男人，这是一个主观的问题，每个人的审美观和喜好都不尽相同。因此，没有一个确切的答案。在不同的文化、地区和时代，人们对于帅气的定义也会有所不同。对于某些人来说，帅可能指的是外表的英俊，而对于其他人来说，帅可能指的是内在的品质和才能。所以，谁是最帅的男人，这是一个因人而异的问题，没有固定的答案。
```

Step6 安装微调所需的依赖

本实验使用官方提供的微调代码，按照如下指令安装依赖。

```
cd ../
cd finetune_demo/
conda install gcc_linux-64
pip install -r requirements.txt
pip install nltk   # 补充安装的库，不然后续训练会报错
```

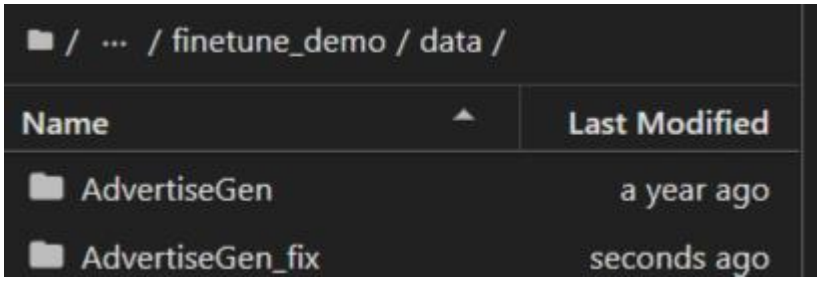
Step7 下载并处理 AdvertiseGen 数据集（2 分）

```
wget https://cloud.tsinghua.edu.cn/f/b3f119a008264b1cabd1/?dl=1
mv index.html?dl=1 AdvertiseGen.tar.gz
mkdir data
tar -xzf AdvertiseGen.tar.gz -C /mnt/workspace/ChatGLM3/finetune_demo/data
```

然后需要对解压后的 AdvertiseGen 数据集进行处理，以符合模型训练所需格式，数据需转化到如下格式：

{"conversations": [{"role": "user", "content": "类型#裙*裙长#半身裙"}, {"role": "assistant", "content": "这款百搭时尚的仙女半身裙……。"}]}

数据可以自行处理，也可以按照官方提供的 lora_fineyune.ipynb 处理数据集，切割后的数据需保存在 finetune_demo/data/AdvertiseGen_fix 文件下。数据处理成功后，data 文件夹会出现以下两个文件。



报告部分需提供你处理数据所使用的代码截图和运行成功后的截图。

Step8 训练模型（4 分）

运行下面指令训练模型：

python finetune_hf.py 处理后的数据集路径 下载的 chatglm3-6b 模型路径 configs/lora.yaml

如果出现报错，按照提示安装所需函数库

模型成功训练需截图。

Step9 验证广告生成功能（4 分）

```
python inference_hf.py your_finetune_path --prompt your prompt
```

可以使用如下提示词验证模型：

--prompt "类型#裙*版型#显瘦*材质#网纱*风格#性感*裙型#百褶*裙下摆#压褶*裙长#连衣裙*裙衣门襟#拉链*裙衣门襟#套头*裙款式#拼接*裙款式#拉链*裙款式#木耳边*裙款式#抽褶*裙款式#不规则"

--prompt "类型#裤*材质#羊毛*裤长#九分裤*裤口#微喇裤"

广告成功生成需截图