用代码方式微调的方案，仅使用于本招投标项目

笔记本:	我的第一个笔记本		
创建时间:	2024/11/20 13:48	**更新时间:**	2024/11/20 14:35
作者:	153klixx022		
URL:	file:///D:/科大讯飞实习/模型微调/国产大模型chatglm3微调闭坑总结/国产大模型chat...		

一. 模型训练

从github上面下载我的项目：

```
git clone https://github.com/JieGeng1998-10-13/fine_tune_chatglm3.git
```

找到这个路径的文件夹： ChatGLM3代码微调/finetune_demo

找到这个笔记本lora_finetune.ipynb

按照笔记本的指示阅读代码并运行即可

注意：

ChatGLM3代码微调/finetune_demo/data/AdvertiseGen 存放的是训练集和验证集
分别是data.json和dev.json 根据需要上传自己的数据集

微调结果存放在output文件夹中， 以checkpoint的形式存在，每隔500个step存储一次

configs文件夹中存放着lora.yaml,是lora微调的配置文件，有硬件条件也可以选择另外两种微调方式比如ptuning，sft

我对配置进行了一定修改，保证训练的效果，配置如下：

```
data_config:
  train_file: train.json
  val_file: dev.json
  test_file: dev.json
  num_proc: 16
max_input_length: 256
```

```
max_output_length: 512
training_args:
  # see `transformers.Seq2SeqTrainingArguments`
  output_dir: ./output
  max_steps: 3000
  # needed to be fit for the dataset
  learning_rate: 5e-5
  # settings for data loading
  per_device_train_batch_size: 4
  dataloader_num_workers: 16
  remove_unused_columns: false
  # settings for saving checkpoints
  save_strategy: steps
  save_steps: 500
  # settings for logging
  log_level: info
  logging_strategy: steps
  logging_steps: 10
  # settings for evaluation
  per_device_eval_batch_size: 16
  evaluation_strategy: steps
  eval_steps: 500
  # settings for optimizer
  # adam_epsilon: 1e-6
  # uncomment the following line to detect nan or inf values
  # debug: underflow_overflow
  predict_with_generate: true
  # see `transformers.GenerationConfig`
  generation_config:
    max_new_tokens: 512
  # set your absolute deepspeed path here
  #deepspeed: ds_zero_2.json
  # set to true if train with cpu.
  use_cpu: false
peft_config:
  peft_type: LORA
  task_type: CAUSAL_LM
  r: 16
  lora_alpha: 32
  lora_dropout: 0.1
```

二. lora模型转化

训练完成后找到一个放置llama.cpp的文件夹（已经安装编译的可以略过）

这一步是为了使用ollama支持的格式进行本地自定义大模型的构建。

先在workspace下载llama.cpp作为转化工具

```
git clone https://github.com/ggerganov/llama.cpp.git
```

然后用make -j进行编译：

```
cd llama.cpp
make -j
```

新建一个环境，比如

```
conda create -n llamacpp python=3.10
```

注意：尽可能选择3.10的python版本，否则容易报错

然后激活环境，安装所需要的包

```
conda activate llamacpp
pip install -r requirements.txt
```

找到python脚本所在位置，在llama.cpp的文件夹中找到

这个脚本： convert_lora_to_gguf.py

对Lora模型进行转化：

```
/mnt/workspace/llama.cpp# python convert_lora_to_gguf.py --base
/mnt/workspace/chatglm3-6b --outfile /mnt/workspace/models
/mnt/workspace/ChatGLM3代码微调/finetune_demo/output/checkpoint-3000
```

在models文件夹中可以找到models/checkpoint-3000-F16-LoRA.gguf 这个文件

我把它的名字改为models2.gguf, 存放在自己能够找到的位置

注意：

/mnt/workspace/chatglm3-6b指的是本体模型所在路径

/mnt/workspace/models指的是输出目录

后面的是想要转化的checkpoint的目录

三. 转化本体模型

```
python convert_hf_to_gguf.py /mnt/workspace/chatglm3-6b
```

注意：

/mnt/workspace/chatglm3-6b指的是本体模型所在路径

四. Ollama部署合并模型

根据ollama官方文档进行modelfile的编写

```
https://github.com/ollama/ollama/blob/main/docs/modelfile.md
```

或者可以直接参考我写的chatglm-SQLprompt.Modelfile，放在配置文件中，修改路径提示词部分即可

具体内容如下：

```
FROM /mnt/workspace/models/chatglm3-6B-F16.gguf
TEMPLATE "
[gMASK]<sop>{{ if .System }}<|system|>
{{ .System }}{{ end }}{{ if .Prompt }}<|user|>
{{ .Prompt }}{{ end }}<|assistant|>
{{ .Response }}
"
SYSTEM """
当你需要执行SQL查询相关的任务时，你要遵守下面的表格结构，
CREATE TABLE "中标结果" (
    "中标人名称" TEXT,
    "价格（万）" TEXT,
    "主要标的信息" TEXT,
```

```
    "页面网址" TEXT,
    "项目编号" TEXT,
    "项目名称" TEXT,
    "日期" TEXT,
    "类别" TEXT,
    "爱企查网址" TEXT,
    "公司官网" TEXT,
    "修改后的网址" TEXT
)


/*
3 rows from 中标结果 table:
中标人名称  价格（万）  主要标的信息  页面网址  项目编号  项目名称  日
期  类别  爱企查网址  公司官网  修改后的网址
上海沥新市政工程有限公司  34.8339  对大芦东路(潮和路-渔港路)，实施路段起点桩号
K3+086，止点桩号K4+402，长度约1.316公里预养护，实施将恢复路面表层沥青油质，防止路面面层
结构骨料的进一步散失、脱落，从而延长道路使用寿
命。  https://www.shggzy.com/szjtzbjggs/8041363  JTHY20240918006  2024年农村
公路预养护大芦东路(潮和路-渔港路)服务项目-2024年农村公路预养护大芦东路(潮和路-渔港路)服
务项目成交结果公告  2024年9月30日  公路项
目  https://aiqicha.baidu.com/detail/compinfo?
pid=44872237309270&rq=ef&pd=ee&from=ps&query=上海沥新市政工程有限公司  暂无网
址  https://aiqicha.baidu.com/company_detail_44872237309270
上海城建审图咨询有限公司  156.28  中标金额156.28万
元。  https://www.shggzy.com/szjtzbjggs/8040530  JTHY20240822001  S32公路
（G1503公路-G15公路）改建工程--施工图设计文件审查-S32公路（G1503公路-G15公路）改建工程
施工图设计文件审查成交结果公告  2024年9月30日  公路项
目  https://aiqicha.baidu.com/detail/compinfo?
pid=31676829900729&rq=ef&pd=ee&from=ps&query=上海城建审图咨询有限公司  暂无网
址  https://aiqicha.baidu.com/company_detail_31676829900729
上海申龙道路设施工程有限公司  95.964525  依据磋商文件、工程量清单、施工图图纸所显示
的波形护栏、机非隔离护栏等所有交通安全设施的施工、验收、保修等各阶段的所有工作内容以及发
包人要求的其他工作内容，所有按工程规范要求以及为取得政府主管部门批准
所  https://www.shggzy.com/szjtzbjggs/8039824  JTHY20240911001  奉贤区神州路
（奉云东路-G1503)交通安全设施工程-奉贤区神州路（奉云东路-G1503）交通安全设施工程成交结果
公告  2024年9月29日  公路项目  https://aiqicha.baidu.com/detail/compinfo?
pid=31949245434116&rq=ef&pd=ee&from=ps&query=上海申龙道路设施工程有限公  暂无网
址  https://aiqicha.baidu.com/company_detail_31949245434116
*/



CREATE TABLE "变更和异常" (
    "项目编号" TEXT,
    "内容" TEXT,
    "页面网址" TEXT,
    "类别" TEXT,
    "状态" TEXT
)


/*
3 rows from 变更和异常 table:
项目编号  内容  页面网址  类别  状态
JTHY20240906004  变更公告
奉贤区10米级氢能源城市客车采购项目更正公告
原招标文件中支付方式为：招标人在提车时且收到公交车辆购置专项资金后支付车款的20%，3个月后
支付车款的40%，6个月后支付车款35%，车款的
5%  https://www.shggzy.com/szjtbggg/8023975  公路项目  变更
JTHY20240718002  变更公告
G1503地面道路（牡丹江路-富长路）绿化还建工程
施工补充公告
招标项目编码：  JTHY20240718002
投标文件提交的截止时间(投标截止时间，下同)变更为2024年8月16日
15  https://www.shggzy.com/szjtbggg/7983993  公路项目  变更
JTHY20240530012  变更公告
变更公告
项目名称：排堵保畅工程--勘察
```

```
开标时间（响应文件递交截止时间）延期至：2024-07-01 10:00
招标需求调整详见采购澄清文件
招标人：上海市奉贤区交通建设管理中心
地址：    https://www.shggzy.com/szjtbggg/7919418    公路项目    变更
*/



CREATE TABLE "招标公告" (
    "项目编号" TEXT,
    "网址" TEXT,
    "类别" TEXT,
    "项目名称" TEXT,
    "发布日期" TIMESTAMP,
    "招标公告" TEXT
)


/*
3 rows from 招标公告 table:
项目编号    网址    类别    项目名称    发布日期    招标公告
JTHY20240507001    https://www.shggzy.com/szjtZbgg/7879395    市政交通    2024年上海市道路合杆整治工程（虹口区） 监理-2024年上海市道路合杆整治工程（虹口区） 监理采购公告    2024-05-07 00:00:00    项目2024年上海市道路合杆整治工程（虹口区）具体内容为对虹口区内30条道路实施合杆整治，整治道路长度为10.887公里，包括原有杆件、杆上废弃设施、箱体、基础、管线、手井等设施拆除；新建综合杆、箱体
JTHY20240507002    https://www.shggzy.com/szjtZbgg/7879416    市政交通    2024年上海市道路合杆整治工程（黄浦区） 监理-2024年上海市道路合杆整治工程（黄浦区） 监理采购公告    2024-05-07 00:00:00    项目2024年上海市道路合杆整治工程（黄浦区）具体内容为对黄浦区内28条道路实施合杆整治，整治道路长度为7.3公里，包括原有杆件、杆上废弃设施、箱体、基础、管线、手井等设施拆除；新建综合杆、箱体及配套
JTHY20240507003    https://www.shggzy.com/szjtZbgg/7879420    市政交通    2024年上海市道路合杆整治工程（宝山区） 监理-2024年上海市道路合杆整治工程（宝山区） 监理采购公告    2024-05-07 00:00:00    项目2024年上海市道路合杆整治工程（宝山区）具体内容为对宝山区内4条道路实施合杆整治，整治道路长度为2.839公里，整治内容有：原有杆件、杆上废弃设施、箱体、基础、管线、手井等设施拆除；新建综合杆、
*/



CREATE TABLE "项目分类" (
    "项目名称" TEXT,
    "项目编号" TEXT,
    "时间" TEXT,
    "交易分类" TEXT,
    "项目类型" TEXT,
    "公告类型" TEXT,
    "类型" TEXT,
    "阶段" TEXT
)


/*
3 rows from 项目分类 table:
项目名称    项目编号    时间    交易分类    项目类型    公告类型    类型    阶段
铁路杭州萧山机场站枢纽及接线工程第三方监测标招标公告    TLSG202410080342    2024-10-08    铁路建设    施工    招标公告和资格预审公告    None    None
阜阳站信号联锁等室内设备更新改造工程工程总承包招标公告    TLSG202409180333    2024-09-28    铁路建设    施工    招标公告和资格预审公告    None    None
芜湖东驼峰场驼峰自动控制系统等室内设备更新改造工程工程总承包招标公告    TLSG202409180334    2024-09-28    铁路建设    施工    招标公告和资格预审公告    None    None
*/
"""
ADAPTER /mnt/workspace/models/models2.gguf
PARAMETER stop <|system|>
PARAMETER stop <|user|>
PARAMETER stop <|assistant|>
```

```
PARAMETER temperature 0
```

在脚本所在文件夹运行，下面这个是伪代码，替换成自己的名称和路径

```
ollama create choose-a-model-name -f <location of the file e.g. ./Modelfile>'
```

我直接就是这么写，在当前文件夹就能找到chatglm-SQLprompt.Modelfile

```
ollama create jiesql_thousand -f chatglm-SQLprompt.Modelfile
```

运行ollama list

```
ollama list
```

就可以看到可以使用这个新的微调后的模型了

```
NAME                              ID              SIZE      MODIFIED
jiesql_thousand:latest            1782c8d46686    12 GB     22 seconds ago
jiesql:latest                     2ef3ea6a8cf9    12 GB     2 days ago
bge-m3:latest                     790764642607    1.2 GB    2 days ago
shaw/dmeta-embedding-zh:latest    55960d8a3a42    408 MB    3 weeks ago
EntropyYue/chatglm3:6b            254ec1286add    3.6 GB    3 weeks ago
milkey/m3e:latest                 1477f12451b0    650 MB    3 weeks ago
```

然后就可以在本地使用这个模型啦