LLM实验一：大模型部署,qwen1.5-0.5b

1.模型部署

1.1 模型下载

模型简介，请访问https://www.modelscope.cn/models/qwen/Qwen1.5-0.5B-Chat-GGUF/summary。

下载代码使用download_model.py

```
from modelscope.hub.file_download import model_file_download

model_dir = model_file_download(model_id='qwen/Qwen1.5-0.5B-Chat-GGUF',file_path='qwen1_5-0_5b-chat-q5_k_m.gguf',revision='master',cache_dir='path/to/local/dir')
```

1.2 下载llama.cpp

使用git命令克隆llama.cpp项目

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

编译llama.cpp项目

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

1.3 加载模型，并执行

在llama.cpp目录中执行命令。

```
./main -m /path/to/local/dir/qwen/Qwen1.5-0.5B-Chat-GGUF/qwen1_5-0_5b-chat-q5_k_m.gguf -n 512 --color -i -cml
```

上述命令中的"/path/to/local/dir"在执行时需要替换为实际的本地目录。

2.基于OpenVINO的模型量化实践

2.1 安装基本环境

（1）创建目录qwen-ov

具体依赖的requirements.txt, 参见 https://github.com/OpenVINO-dev-contest/Qwen2.openvino。

创建Python虚拟环境：

```
python -m venv qwenVenv
source qwenVenv/bin/activate
```

安装依赖的包：

```
pip install wheel setuptools
pip install -r requirements.txt
```

2.2 下载模型

export HF_ENDPOINT=https://hf-mirror.com

huggingface-cli download --resume-download --local-dir-use-symlinks False Qwen/Qwen1.5-0.5B-Chat --local-dir {your_path}/Qwen1.5-0.5B-Chat

2.3 转换模型

python3 convert.py --model_id Qwen/Qwen1.5-0.5B-Chat --precision int4 --output {your_path}/Qwen1.5-0.5B-Chat-ov

2.4 加载模型并执行

python3 chat.py --model_path {your_path}/Qwen1.5-0.5B-Chat-ov --max_sequence_length 4096 --device CPU

3.任务要求

3.1 使用魔搭上的Intel CPU资源

通过登录并使用魔搭平台及注册时关联阿里云账号获得的免费CPU云计算资源，通过启动Jupyter Notebook或相应环境镜像进入相应的项目部署环境。

3.2 部署

根据相应模型的部署文档，确认或完成运行环境的准备，并在导入Notebook及相应模型后，完成模型的直接部署，或根据部署文档中的模型优化方式在完成模型低精度量化或格式转换后进行部署。

3.3 进行问答测试

部署后，可以针对相应模型进行一些应用场景如问答等的测试。

3.4 部署典型的开源LLM，进行比较

部署/实验多个项目中列明的模型，建议部署3-4个以开展不同模型的横向比对并提交符合要求的项目报告。

推荐的开源LLM：

- 通义千问(Qwen1.5-0.5B-Chat, Qwen-7B-Chat)
- 智谱 ChatGLM3-6B
- 百川2-7B-对话模型
- Neural-Chat
- ...

3.5 形成最终的报告

要求将模型部署过程中的步骤加以截图，图文并茂加以说明总结：

- git相关的截图，以及实验完成的相关截图
- 以Markdown文档形式编写报告
- 将部署中涉及的代码上传到个人的github或gitee的public账户，将相关链接放到报告里
- 将LLM部署过程中形成的经验认真撰写成博客文章，发布到csdn等技术博客平台，将相关链接放到报告里