

项目部署

1. 总览

本项目为人工智能大模型部署作业的部署脚本仓库与提问仓库

2. 部署

2.1 在服务器平台上进行部署

本项目的在线服务器版本采用魔搭平台进行部署。

2.1.0 服务器环境

服务器获取方式：绑定阿里云后免费获取

服务器类型：PAI-DSW免费示例，CPU环境，CPU为8核，内存32GB，预装 ModelScope Library

操作系统版本：ubuntu22.04-py311-torch2.3.1-1.26.0

2.1.1 python依赖项的部署

(1) . Conda

运行：

```
cd /opt/conda/envs
```

若无此目录，说明服务器端未配置Conda，需进行安装

执行以下安装指令：

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linuxx86_64.sh
bash Miniconda3-latest-Linux-x86_64.sh -b -p /opt/conda
echo 'export PATH="/opt/conda/bin:$PATH"' >> ~/.bashrc
source ~/.bashrc
conda --version
```

安装完成后，激活Conda环境

执行以下代码：

```
conda create -n qwen_env python=3.10 -y
source /opt/conda/etc/profile.d/conda.sh
conda activate qwen_env
```

(2) . 基础环境

首先安装pytorch环境，如已经安装，则进行版本控制

执行以下代码：

```
pip install \
torch==2.3.0+cpu \
torchvision==0.18.0+cpu \
--index-url https://download.pytorch.org/whl/cpu
```

(3) 基础依赖

执行以下代码：

```
pip install \
"intel-extension-for-transformers==1.4.2" \
"neural-compressor==2.5" \
"transformers==4.33.3" \
"modelscope==1.9.5" \
"pydantic==1.10.13" \
"sentencepiece" \
"tiktoken" \
"einops" \
"transformers_stream_generator" \
"uvicorn" \
"fastapi" \
"yacs" \
"setuptools_scm"
```

依赖项安装完成后，应安装fs chat便于后续进行对话

执行以下代码：

```
pip install fschat --use-pep517
```

若想增加体验，也可选择安装tqdm、huggingface-hub 等

执行以下代码：

```
pip install tqdm huggingface-hub
```

2.1.2 大模型实践

(1) 模型部署

首先，在魔搭服务器中将路径切换至数据目录下，以确保数据可持久化

执行以下代码：

```
cd /mnt/data
```

在该目录下，部署指定的大模型，此处以清华智谱的6b模型为例

执行以下代码：

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

(2) 构建示例并进行测试

首先，在魔搭服务器中将路径切换至数据目录下

执行以下代码：

```
cd /mnt/workspace
```

在该路径下，编写对应脚本，本次测样使用的脚本位于 `test` 文件夹中

编写完成后，执行以下脚本，进行测试

```
python 脚本名.py
```

2.2 在本地进行部署

本项目的本地版本采用Ollama进行部署

2.2.1 Ollama的安装

访问以下网址以获取Windows下的Ollama安装包

```
https://ollama.com/download/OllamaSetup.exe
```

下载完成后，双击运行，并完成后续安装操作

2.2.2 利用Ollama部署大模型

Ollama支持一键大模型部署，仅需执行部署指令即可

本项目共部署deepseek R1-7b, deepseek R1-32b, qwen-32b

在命令提示符中执行以下命令

```
ollama pull deepseek-r1:7b
ollama pull deepseek-r1:32b
ollama pull qwen3:32b
```

执行后，等待模型下载并部署结束即可

2.2.3 运行并进行相关测试

Ollama支持直接在命令提示符中使用命令

```
ollama run 指定模型名
```

进行对话，但是出于美观和便捷因素，本项目部署了Page Assist作为图形化界面进行测试

Page Assist是一款基于Chrome的浏览器插件，在Chrome的浏览器插件商店中搜索并直接安装即可使用

启动前，应确保Ollama处于运行状态

启动后，可直接进入进入对话界面

Ollama正在运行

输入一条消息...

提交

左上角可进行选择模型，保存对话记录等操作，提问无需编写推理脚本，直接在下方文本框中输入相关问题并点击提交即可

3. 模型对比与分析

此部分内容详见同目录下的课程报告部分