第四次作业教程

2025 人工智能导论

目的：本地推理、非量化

注意 ：

此平台 CPU 资源 1 小时未使用会自动释放，配置的环境会被清空！

大模型一次最好只下载一个调配，否则会存储不足！

A. 平台搭建

一、 注册&登录 ModelScope

进入 ModelScope（https://www.modelscope.cn/home），右上角点击完成新用户的注册；

二、 获取计算资源

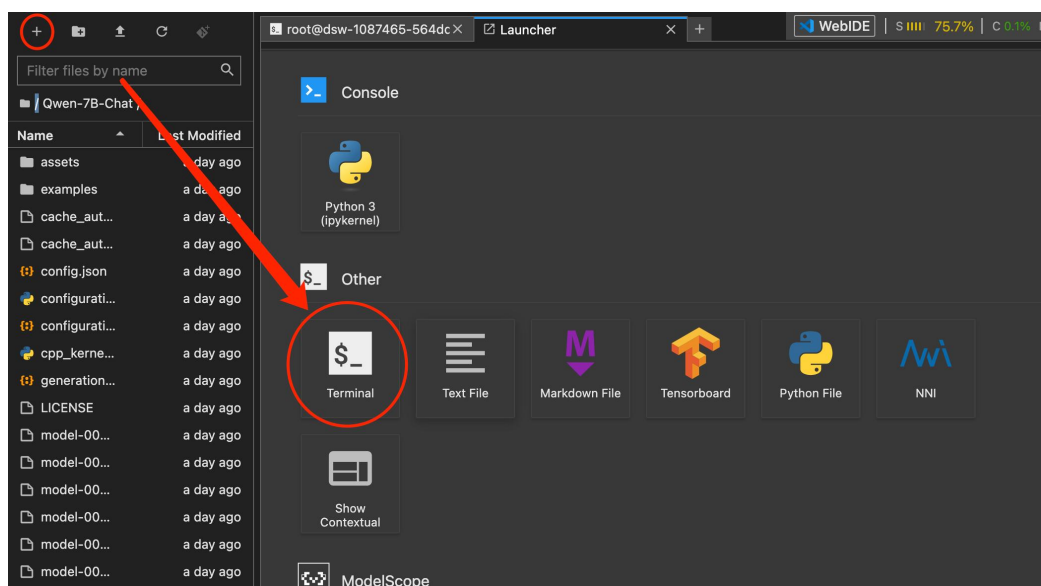注册完成后，登录 ModelScope，进入首页，请确定已经绑定阿里云账 号，并已获得免费的配套云计算资源。启动 CPU 服务器：

三、 启动 Notebook 准备

环境配置

一、 点击 Terminal 图标，打开终端命令行环境

二、 环境搭建

以下根据是否选择在 conda 环境中运行实例自行选择操作步骤

- conda 环境：步骤 1、2

- root 直接操作：2.2

1. CPU 版本发现并没有 conda,手动下载：

```
cd /opt/conda/envs
#问题：没有那个文件或目录
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-
x86_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
```

激活环境：

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

2. 基础依赖

(1) 基础环境：

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

(2) 基础依赖

在执行前加一条命令检查 pip 是否能正常联网

```
pip install -U pip setuptools wheel
```

```
# 安装基础依赖（兼容 transformers 4.33.3 和 neuralchat）
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"


# 安装 fschat（需要启用 PEP517 构建）
pip install fschat --use-pep517
```

3. 可选：安装 tqdm、huggingface-hub 等增强体验

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

B. 大模型实践

一、 下载大模型到本地

1. 切换到数据目录

```
cd /mnt/data
```

2. 下载对应大模型

供参考的大模型，可自行调研配置

一次最好只跑下载一个大模型，否则会存储不足。

```
git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
git clone https://www.modelscope.cn/qwen/Qwen-7B-Chat.git
```

```
Qwen-7B-Chat/
├── config.json
├── generation_config.json
├── pytorch_model-00001-of-00002.bin
├── pytorch_model-00002-of-00002.bin
├── tokenizer_config.json
├── tokenizer.model or tokenizer.json
├── model_index.json  （如是 ModelScope 格式）
├── README.md
```

二、 构建实例

1. 切换工作目录

```
cd /mnt/workspace
```

2. 实例代码：

编写推理脚本 run_qwen_cpu.py

```python
from transformers import TextStreamer, AutoTokenizer, AutoModelForCausalLM

model_name = "/mnt/data/Qwen-7B-Chat"  # 本地路径
prompt = "请说出以下两句话区别在哪里？ 1、冬天：能穿多少穿多少 2、夏天：能穿多少穿多少"

tokenizer = AutoTokenizer.from_pretrained(
    model_name,
    trust_remote_code=True
)

model = AutoModelForCausalLM.from_pretrained(
    model_name,
    trust_remote_code=True,
    torch_dtype="auto"  # 自动选择 float32/float16（根据模型配置）
).eval()

inputs = tokenizer(prompt, return_tensors="pt").input_ids

streamer = TextStreamer(tokenizer)
outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)
```

三、 运行实例

```
python run_qwen_cpu.py
```

```
root@dsw-1087465-564dd9db5-6w2p8:/mnt/workspace/itrex# python run_llm.py
Loading checkpoint shards: 100%|
                                      | 8/8 [00:50<00:00,  6.31s/it]
2025-05-20 12:52:18.301355: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different numerical results due t
o floating-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-05-20 12:52:18.706124: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:485] Unable to register cuFFT factory: Attempting to register fac
tory for plugin cuFFT when one has already been registered
2025-05-20 12:52:18.854611: E external/local_xla/xla/stream_executor/cuda/cuda_dnn.cc:8454] Unable to register cuDNN factory: Attempting to register fa
ctory for plugin cuDNN when one has already been registered
2025-05-20 12:52:18.899960: E external/local_xla/xla/stream_executor/cuda/cuda_blas.cc:1452] Unable to register cuBLAS factory: Attempting to register
factory for plugin cuBLAS when one has already been registered
2025-05-20 12:52:19.197355: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions
in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
2025-05-20 12:52:20.908351: W tensorflow/compiler/tf2tensorrt/utils/py_utils.cc:38] TF-TRT Warning: Could not find TensorRT
Once upon a time, there existed a little

girl, who loved to explore the world around her. She would often venture out into the forest near her home, and spend hours playing in the streams and
discovering new creatures.

One day, as she was wandering through the forest, she stumbled upon a hidden cave. It was dark inside, but the little girl felt drawn to enter. As she
made her way deeper into the cave, she noticed that the air grew colder and the ground became slippery.

Suddenly, she heard a soft rustling noise coming from behind her. The little girl turned around to see a small, furry creature with big, bright eyes st
aring back at her. The creature seemed friendly enough, so the little girl reached out her hand to pet it.

To her surprise, the creature didn't hesitate to climb onto her lap and snuggle up against her. It seemed to be very content to stay there for hours, p
urring softly and occasionally nuzzling its nose against the little girl's ear.
```

其他大预言模型部署

在完成上述实验后，可以进一步尝试下载及加载其他大语言模型进行部署实验的体验。推荐体验的开源模型列表

清华智谱（ZhipuAI）

```
git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
git clone https://www.modelscope.cn/ZhipuAI/chatglm2-6b.git
git clone https://www.modelscope.cn/ZhipuAI/chatglm-6b.git
```

百川智能（Baichuan）

```
git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-7B-Base.git
git clone https://www.modelscope.cn/baichuan-inc/Baichuan2-13B-Base.git
git clone https://www.modelscope.cn/baichuan-inc/Baichuan-13B-Chat.git
```

- 深度求索（DeepSeek）
- 智源研究院（BLOOMZ-zh 微调版）

参考资料：

https://github.com/intel/intel-extension-for-transformers/tree/main

https://github.com/intel/intel-extension-for-transformers/blob/main/docs/installation.md