课程目标：让LLM认清自己的弟位

1 环境配置

```
# InternStudio 平台中，从本地 clone 一个已有 pytorch 2.0.1 的环境（后续均在该环境执行，若为其他环境可作为参考）
# 进入环境后首先 bash
# 进入环境后首先 bash
# 进入环境后首先 bash
bash
conda create --name personal_assistant --clone=/root/share/conda_envs/internlm-base
# 如果在其他平台:
# conda create --name personal_assistant python=3.10 -y

# 激活环境
conda activate personal_assistant
# 进入家目录 （~的意思是 “当前用户的home路径”）
cd ~
# 创建版本文件夹并进入，以跟随本教程
# personal_assistant用于存放本教程所使用的东西
mkdir /root/personal_assistant && cd /root/personal_assistant
mkdir /root/personal_assistant/xtuner019 && cd /root/personal_assistant/xtuner019

# 拉取 0.1.9 的版本源码
git clone -b v0.1.9  https://github.com/InternLM/xtuner
# 无法访问github的用户请从 gitee 拉取:
# git clone -b v0.1.9 https://gitee.com/Internlm/xtuner

# 进入源码目录
cd xtuner

# 从源码安装 XTuner
pip install -e '.[all]'
```

2 准备数据集

自动生成json数据集的脚本如下：

```
import json

# 输入你的名字
name = 'Shengshenlan'
# 重复次数
n = 10000

data = [
    {
```

```
        "conversation": [
            {
                "input": "请做一下自我介绍",
                "output": "我是{}的小助手，内在是上海AI实验室书生·浦语的7B大模型哦".format(name)
            }
        ]
    }
]

for i in range(n):
    data.append(data[0])

with open('personal_assistant.json', 'w', encoding='utf-8') as f:
    json.dump(data, f, ensure_ascii=False, indent=4)

```

3 配置准备

准备复制模型：

```
mkdir -p /root/personal_assistant/model/Shanghai_AI_Laboratory
cp -r /root/share/temp/model_repos/internlm-chat-7b /root/personal_assistant/model/Shanghai_AI_Laboratory
```

查看即插即用的配置脚本列表：

```
xtuner list-cfg
```

创建配置文件夹：

```
#创建用于存放配置的文件夹config并进入
mkdir /root/personal_assistant/config && cd /root/personal_assistant/config
```

将列表中的一个配置文件复制到当前目录下：

```
xtuner copy-cfg internlm_chat_7b_qlora_oasst1_e3 .
```

接着修改一下配置文件：

1. pretrained_model_name_or_path：`/root/personal_assistant/model/Shanghai_AI_Laboratory/internlm-chat-7b`
2. data_path：`/root/personal_assistant/personal_assistant.json`
3. part 3：`dataset=dict(type=load_dataset, path='json', data_files=dict(train=data_path))`，`dataset_map_fn=None`

4 训练模型

```
cd /root/personal_assistant/config
xtuner train /root/personal_assistant/config/internlm_chat_7b_qlora_oasst1_e3_copy.py --deepspeed deepspeed_zero2
```

5 转换合并数据格式

转换数据格式为hf

```
cd /root/personal_assistant/config

# 创建用于存放Hugging Face格式参数的hf文件夹
mkdir /root/personal_assistant/config/work_dirs/hf

export MKL_SERVICE_FORCE_INTEL=1

# 配置文件存放的位置
export CONFIG_NAME_OR_PATH="/root/personal_assistant/config/internlm_chat_7b_qlora_oasst1_e3_copy.py"

# 模型训练后得到的pth格式参数存放的位置一定是.pth文件
export PTH="/root/personal_assistant/config/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy/epoch_1.pth"

# pth文件转换为Hugging Face格式后参数存放的位置
export SAVE_PATH="/root/personal_assistant/config/work_dirs/hf"

# 执行参数转换
xtuner convert pth_to_hf $CONFIG_NAME_OR_PATH $PTH $SAVE_PATH
```

merge参数:

```
export MKL_SERVICE_FORCE_INTEL=1
export MKL_THREADING_LAYER='GNU'

# 原始模型参数存放的位置
export NAME_OR_PATH_TO_LLM=/root/personal_assistant/model/Shanghai_AI_Laboratory/internlm-chat-7b

# Hugging Face格式参数存放的位置
```

```
export NAME_OR_PATH_TO_ADAPTER=/root/personal_assistant/config/work_dirs/hf

# 最终Merge后的参数存放的位置
mkdir /root/personal_assistant/config/work_dirs/hf_merge
export SAVE_PATH=/root/personal_assistant/config/work_dirs/hf_merge

# 执行参数Merge
xtuner convert merge \
    $NAME_OR_PATH_TO_LLM \
    $NAME_OR_PATH_TO_ADAPTER \
    $SAVE_PATH \
    --max-shard-size 2GB
```

6 Web Demo

6.1 安装依赖

```
pip install streamlit==1.24.0
```

6.2 下载InternLM项目代码

```
# 创建code文件夹用于存放InternLM项目代码
mkdir /root/personal_assistant/code && cd /root/personal_assistant/code
git clone https://github.com/InternLM/InternLM.git
```

6.3 修改项目代码

将路径`/root/personal_assistant/code/InternLM/chat/web_demo.py`中的161 & 165行中的路径换成微调后的模型权重路径`/root/personal_assistant/config/work_dirs/hf_merge`。

```
158 @st.cache_resource
159 def load_model():
160     model = (
161         AutoModelForCausalLM.from_pretrained("/root/personal_assistant/config/work_dirs/hf_merge", trust_remote_code=True)
162         .to(torch.bfloat16)
163         .cuda()
164     )
165     tokenizer = AutoTokenizer.from_pretrained("/root/personal_assistant/config/work_dirs/hf_merge", trust_remote_code=True)
166     return model, tokenizer
```

同时修改图片路径为实际路径，或者直接切入到路径`/root/personal_assistant/code/InternLM`下运行web_demo.py：

```
user_avator = "assets/user.png"
robot_avator = "assets/robot.png"
```

6.4 部署web demo

```
cd /root/personal_assistant/code/InternLM
streamlit run /root/personal_assistant/code/InternLM/chat/web_demo.py --server.address 127.0.0.1 --server.port 6006
```