1 创建环境

```
bash

# 创建环境
conda create --name personal_assistant python=3.10 -y

# 激活环境
conda activate personal_assistant

cd /root/
mkdir /root/personal_assistant && cd /root/personal_assistant

# 拉取0.1.9的版本源码
git clone -b v0.1.9  https://github.com/InternLM/xtuner

# 进入源码目录
cd xtuner

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

2 数据准备

在 `data` 目录下创建一个json文件 `personal_assistant.json` 作为本次微调所使用的数据集。

```
mkdir /root/personal_assistant/data && cd /root/personal_assistant/data

touch personal_assistant.json
```

personal_assistant.json内容如下：复制多次数据增强

```
[
    {
        "conversation": [
            {
                "input": "请介绍一下你自己",
                "output": "我是星辰的小助手，内在是上海AI实验室书生·浦语的7B大模型哦"
            }
        ]
    },
    {
        "conversation": [
            {
                "input": "请做一下自我介绍",
                "output": "我是星辰的小助手，内在是上海AI实验室书生·浦语的7B大模型哦"
            }
        ]
    }
```

```
]
```

3 配置准备

```
# 下载模型InternLM-chat-7B
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

# 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

# 列出所有内置配置
xtuner list-cfg

# 拷贝一个配置文件到当前目录：xtuner copy-cfg ${CONFIG_NAME} ${SAVE_PATH}
xtuner copy-cfg internlm_chat_7b_qlora_oasst1_e3 .
```

```
(personal_assistant) root@intern-studio:~/personal_assistant/xtuner# mkdir /root/personal_assistant/data && cd /root/personal_assistant/data
(personal_assistant) root@intern-studio:~/personal_assistant/data# vim personal_assistant.json
(personal_assistant) root@intern-studio:~/personal_assistant/data# mkdir -p /root/personal_assistant/model/Shanghai_AI_Laboratory
(personal_assistant) root@intern-studio:~/personal_assistant/data# cp -r /root/share/temp/model_repos/internlm-chat-7b /root/personal_assistant/model/Shanghai_AI_Laboratory
(personal_assistant) root@intern-studio:~/personal_assistant/data# mkdir /root/personal_assistant/config && cd /root/personal_assistant/config
(personal_assistant) root@intern-studio:~/personal_assistant/config# xtuner copy-cfg internlm_chat_7b_qlora_oasst1_e3 .
[2024-01-10 21:17:52,534] [INFO] [real_accelerator.py:161:get_accelerator] Setting ds_accelerator to cuda (auto detect)
[2024-01-10 21:19:10,456] [INFO] [real_accelerator.py:161:get_accelerator] Setting ds_accelerator to cuda (auto detect)
Copy to ./internlm_chat_7b_qlora_oasst1_e3_copy.py
```

修改拷贝后的文件internlm_chat_7b_qlora_oasst1_e3_copy.py，修改下述位置：

```
# PART 1 中
# 预训练模型存放的位置
pretrained_model_name_or_path =
'/root/personal_assistant/model/Shanghai_AI_Laboratory/internlm-chat-7b'

# 微调数据存放的位置
data_path = '/root/personal_assistant/data/personal_assistant.json'

# 训练中最大的文本长度
max_length = 512

# 每一批训练样本的大小
batch_size = 2

# 最大训练轮数
max_epochs = 3

# 验证的频率
evaluation_freq = 90

# 用于评估输出内容的问题（用于评估的问题尽量与数据集的question保持一致）
evaluation_inputs = [ '请介绍一下你自己', '请做一下自我介绍' ]


# PART 3 中
```

```
dataset=dict(type=load_dataset, path='json',
data_files=dict(train=data_path))
dataset_map_fn=None
```

PART1 部分

root@intern-studio: ~/per × | Launcher × | internlm_chat_7b_qlora_oa ×

```python
from xtuner.model import SupervisedFinetune
from xtuner.utils import PROMPT_TEMPLATE

######################################################################
#                          PART 1  Settings                          #
######################################################################
# Model
pretrained_model_name_or_path = '/root/personal_assistant/model/Shanghai_AI_Laboratory/internlm-chat-7b'

# Data
data_path = '/root/personal_assistant/data/personal_assistant.json'
prompt_template = PROMPT_TEMPLATE.internlm_chat
max_length = 512
pack_to_max_length = True

# Scheduler & Optimizer
batch_size = 2  # per_device
accumulative_counts = 16
dataloader_num_workers = 0
max_epochs = 3
optim_type = PagedAdamW32bit
lr = 2e-4
betas = (0.9, 0.999)
weight_decay = 0
max_norm = 1  # grad clip

# Evaluate the generation performance during the training
evaluation_freq = 90
SYSTEM = ''
evaluation_inputs = [ '请介绍一下你自己', '请做一下自我介绍' ]

```

PART3 部分

root@intern-studio: ~/per × | internlm_chat_7b_qlora_oa ×

```python
        lora_dropout=0.1,
        bias='none',
        task_type='CAUSAL_LM'))

#######################################################################
#                      PART 3  Dataset & Dataloader                   #
#######################################################################
train_dataset = dict(
    type=process_hf_dataset,
    dataset=dict(type=load_dataset, path='json', data_files=dict(train=data_path)),
    tokenizer=tokenizer,
    max_length=max_length,
    dataset_map_fn=None,
    template_map_fn=dict(
        type=template_map_fn_factory, template=prompt_template),
    remove_unused_columns=True,
    shuffle_before_pack=True,
    pack_to_max_length=pack_to_max_length)

train_dataloader = dict(
    batch_size=batch_size,
    num_workers=dataloader_num_workers,
    dataset=train_dataset,
    sampler=dict(type=DefaultSampler, shuffle=True),
    collate_fn=dict(type=default_collate_fn))

```

4 微调启动

```
xtuner train
/root/personal_assistant/config/internlm_chat_7b_qlora_oasst1_e3_copy.py
```

```
01/10 23:09:58 - mmengine - INFO - Epoch(train) [1][380/593]  lr: 5.7671e-05  eta: 0:08:47  time: 2.7894  data_time: 0.0052  memory: 10427  loss: 0.0309  grad_norm: 0.0224
01/10 23:10:25 - mmengine - INFO - Epoch(train) [1][390/593]  lr: 5.2933e-05  eta: 0:08:24  time: 2.7675  data_time: 0.0029  memory: 10427  loss: 0.0239  grad_norm: 0.0197
01/10 23:10:53 - mmengine - INFO - Epoch(train) [1][400/593]  lr: 4.8327e-05  eta: 0:08:00  time: 2.7355  data_time: 0.0061  memory: 10427  loss: 0.0318  grad_norm: 0.0188
01/10 23:11:20 - mmengine - INFO - Epoch(train) [1][410/593]  lr: 4.3866e-05  eta: 0:07:37  time: 2.7865  data_time: 0.0036  memory: 10427  loss: 0.0316  grad_norm: 0.0188
01/10 23:11:49 - mmengine - INFO - Epoch(train) [1][420/593]  lr: 3.9563e-05  eta: 0:07:13  time: 2.8984  data_time: 0.0079  memory: 10427  loss: 0.0351  grad_norm: 0.0182
01/10 23:12:16 - mmengine - INFO - Epoch(train) [1][430/593]  lr: 3.5429e-05  eta: 0:06:49  time: 2.6757  data_time: 0.0056  memory: 10427  loss: 0.0362  grad_norm: 0.0182
01/10 23:12:44 - mmengine - INFO - Epoch(train) [1][440/593]  lr: 3.1476e-05  eta: 0:06:25  time: 2.8308  data_time: 0.0067  memory: 10427  loss: 0.0313  grad_norm: 0.0184
01/10 23:13:13 - mmengine - INFO - after_train_iter in EvaluateChatHook.
01/10 23:13:16 - mmengine - INFO - Sample output:
<s><|User|>:请做一下自我介绍<eoh>
<|Bot|>:我是星辰的小助手，内在是上海AI实验室书生·浦语的7B大模型哦</s>

01/10 23:13:16 - mmengine - INFO - Epoch(train) [1][450/593]  lr: 2.7715e-05  eta: 0:06:01  time: 2.8588  data_time: 0.0034  memory: 10427  loss: 0.0370  grad_norm: 0.0181
01/10 23:13:46 - mmengine - INFO - Epoch(train) [1][460/593]  lr: 2.4158e-05  eta: 0:05:38  time: 3.3144  data_time: 0.2526  memory: 10427  loss: 0.0332  grad_norm: 0.0181
01/10 23:14:16 - mmengine - INFO - Epoch(train) [1][470/593]  lr: 2.0813e-05  eta: 0:05:13  time: 2.9742  data_time: 0.0116  memory: 10427  loss: 0.0325  grad_norm: 0.0182
01/10 23:14:44 - mmengine - INFO - Epoch(train) [1][480/593]  lr: 1.7690e-05  eta: 0:04:48  time: 2.7877  data_time: 0.0059  memory: 10427  loss: 0.0194  grad_norm: 0.0182
01/10 23:15:12 - mmengine - INFO - Epoch(train) [1][490/593]  lr: 1.4798e-05  eta: 0:04:23  time: 2.8110  data_time: 0.0047  memory: 10427  loss: 0.0248  grad_norm: 0.0182
01/10 23:15:41 - mmengine - INFO - Epoch(train) [1][500/593]  lr: 1.2146e-05  eta: 0:03:58  time: 2.8755  data_time: 0.0075  memory: 10427  loss: 0.0178  grad_norm: 0.0174
01/10 23:16:10 - mmengine - INFO - Epoch(train) [1][510/593]  lr: 9.7396e-06  eta: 0:03:33  time: 2.8841  data_time: 0.0057  memory: 10427  loss: 0.0304  grad_norm: 0.0174
01/10 23:16:38 - mmengine - INFO - Epoch(train) [1][520/593]  lr: 7.5867e-06  eta: 0:03:08  time: 2.8623  data_time: 0.0049  memory: 10427  loss: 0.0255  grad_norm: 0.0169
01/10 23:17:07 - mmengine - INFO - Epoch(train) [1][530/593]  lr: 5.6932e-06  eta: 0:02:42  time: 2.8814  data_time: 0.0057  memory: 10427  loss: 0.0253  grad_norm: 0.0168
01/10 23:17:37 - mmengine - INFO - after_train_iter in EvaluateChatHook.
01/10 23:17:39 - mmengine - INFO - Sample output:
<s><|User|>:请做一下自我介绍<eoh>
<|Bot|>:我是星辰的小助手，内在是上海AI实验室书生·浦语的7B大模型哦</s>

01/10 23:17:39 - mmengine - INFO - Epoch(train) [1][540/593]  lr: 4.0643e-06  eta: 0:02:17  time: 2.9816  data_time: 0.0039  memory: 10427  loss: 0.0233  grad_norm: 0.0168
01/10 23:18:09 - mmengine - INFO - Epoch(train) [1][550/593]  lr: 2.7046e-06  eta: 0:01:52  time: 3.2687  data_time: 0.2333  memory: 10427  loss: 0.0303  grad_norm: 0.0168
01/10 23:18:40 - mmengine - INFO - Epoch(train) [1][560/593]  lr: 1.6179e-06  eta: 0:01:26  time: 3.0718  data_time: 0.0045  memory: 10427  loss: 0.0340  grad_norm: 0.0172
01/10 23:19:11 - mmengine - INFO - Epoch(train) [1][570/593]  lr: 8.0723e-07  eta: 0:01:00  time: 3.0607  data_time: 0.0053  memory: 10427  loss: 0.0240  grad_norm: 0.0172
01/10 23:19:41 - mmengine - INFO - Epoch(train) [1][580/593]  lr: 2.7493e-07  eta: 0:00:34  time: 3.0106  data_time: 0.0050  memory: 10427  loss: 0.0295  grad_norm: 0.0168
01/10 23:20:11 - mmengine - INFO - Epoch(train) [1][590/593]  lr: 2.2452e-08  eta: 0:00:07  time: 2.9751  data_time: 0.0057  memory: 10427  loss: 0.0299  grad_norm: 0.0168
01/10 23:20:20 - mmengine - INFO - Exp name: internlm_chat_7b_qlora_oasst1_e3_copy_20240110_225248
01/10 23:20:20 - mmengine - INFO - Saving checkpoint at 1 epochs
01/10 23:20:23 - mmengine - INFO - after_train in EvaluateChatHook.
01/10 23:20:25 - mmengine - INFO - Sample output:
<s><|User|>:请做一下自我介绍<eoh>
<|Bot|>:我是星辰的小助手，内在是上海AI实验室书生·浦语的7B大模型哦</s>
```

5 参数转换

训练后的pth格式参数转Hugging Face格式

```bash
# 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
export PTH=/root/personal_assistant/config/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy/epoch_3.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
```

```
(personal_assistant) root@intern-studio:~/personal_assistant/config# mkdir /root/personal_assistant/config/work_dirs/hf
(personal_assistant) root@intern-studio:~/personal_assistant/config# export MKL_SERVICE_FORCE_INTEL=1
(personal_assistant) root@intern-studio:~/personal_assistant/config# export CONFIG_NAME_OR_PATH=/root/personal_assistant/config/internlm_chat_7b_qlora_oasst1_e3_copy.py
(personal_assistant) root@intern-studio:~/personal_assistant/config# export PTH=/root/personal_assistant/config/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy/epoch_3.pth
(personal_assistant) root@intern-studio:~/personal_assistant/config# export SAVE_PATH=/root/personal_assistant/config/work_dirs/hf
(personal_assistant) root@intern-studio:~/personal_assistant/config# xtuner convert pth_to_hf $CONFIG_NAME_OR_PATH $PTH $SAVE_PATH
[2024-01-10 21:35:17,304] [INFO] [real_accelerator.py:161:get_accelerator] Setting ds_accelerator to cuda (auto detect)
[2024-01-10 21:36:02,502] [INFO] [real_accelerator.py:161:get_accelerator] Setting ds_accelerator to cuda (auto detect)
quantization_config convert to <class 'transformers.utils.quantization_config.BitsAndBytesConfig'>
Loading checkpoint shards: 100% ████████████████████████████████████████ 8/8 [00:17<00:00,  2.15s/it]
01/10 21:36:33 - mmengine - INFO - dispatch internlm attn forward
01/10 21:36:33 - mmengine - WARNING - Due to the implementation of the PyTorch version of flash attention, even when the `output_attentions` flag is set to True, it is not possible to return the `attn_weights`.
Load PTH model from /root/personal_assistant/config/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy/epoch_3.pth
Convert weights to float16
Saving HuggingFace model to /root/personal_assistant/config/work_dirs/hf
All done!
```

6 参数合并

```
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
```

```
(personal_assistant) root@intern-studio:~/personal_assistant/config# export MKL_SERVICE_FORCE_INTEL=1
(personal_assistant) root@intern-studio:~/personal_assistant/config# export MKL_THREADING_LAYER='GNU'
(personal_assistant) root@intern-studio:~/personal_assistant/config# export NAME_OR_PATH_TO_LLM=/root/personal_assistant/model/Shanghai_AI_Laboratory/internlm-chat-7b
(personal_assistant) root@intern-studio:~/personal_assistant/config# export NAME_OR_PATH_TO_ADAPTER=/root/personal_assistant/config/work_dirs/hf
(personal_assistant) root@intern-studio:~/personal_assistant/config# mkdir /root/personal_assistant/config/work_dirs/hf_merge
(personal_assistant) root@intern-studio:~/personal_assistant/config# export SAVE_PATH=/root/personal_assistant/config/work_dirs/hf_merge
(personal_assistant) root@intern-studio:~/personal_assistant/config# xtuner convert merge \
>     $NAME_OR_PATH_TO_LLM \
>     $NAME_OR_PATH_TO_ADAPTER \
>     $SAVE_PATH \
>     --max-shard-size 2GB
[2024-01-10 23:26:35,978] [INFO] [real_accelerator.py:161:get_accelerator] Setting ds_accelerator to cuda (auto detect)
Loading checkpoint shards: 100%|████████████████████████████████████████████████████| 8/8 [00:15<00:00,  1.90s/it]
Saving to /root/personal_assistant/config/work_dirs/hf_merge...
All done!
```

7 Web部署

```
# 安装依赖
pip install streamlit==1.24.0

# 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

# 修改/root/code/InternLM/web_demo.py中的模型路径
修改为"/root/personal_assistant/config/work_dirs/hf_merge"

# 运行脚本
streamlit run /root/personal_assistant/code/InternLM/web_demo.py --
server.address 127.0.0.1 --server.port 6006

# powershell
ssh -CNg -L 6006:127.0.0.1:6006 root@ssh.intern-ai.org.cn -p [开发机端口号]
```

root@intern-studio: ~/per × | web_demo.py ×

```
    del st.session_state.messages


@st.cache_resource
def load_model():
    model = (
        AutoModelForCausalLM.from_pretrained("/root/personal_assistant/config/work_dirs/hf_merge", trust_remote_code=True)
        .to(torch.bfloat16)
        .cuda()
    )
    tokenizer = AutoTokenizer.from_pretrained("/root/personal_assistant/config/work_dirs/hf_merge", trust_remote_code=True)
    return model, tokenizer

```

8 最终效果

变成星辰的小助手啦