1 创建环境

```bash

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

# 激活环境
conda activate game_xtuner

cd /root/SecondLife/
mkdir /root/SecondLife/game_xtuner && cd /root/SecondLife/game_xtuner

# 拉取0.1.14的版本源码
git clone https://github.com/InternLM/xtuner.git

# 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`dataset.json`作为本次微调所使用的数据集。

```
mkdir /root/SecondLife/data && cd /root/SecondLife/data
```

personal_assistant.json内容如下：复制多次数据增强

```
[
    {
        "conversation": [
            {
                "system":
"\u4f60\u662f\u4e00\u4e2a\u5bf9\u540d\u4eba\u751f\u5e73\u7814\u7a76\u9887\u6df1\u
7684\u5b66\u8005\u3002\u4f60\u603b\u662f\u53ef\u4ee5\u6839\u636e\u63d0\u95ee\u7ed
9\u51fa\u51c6\u786e\u4e14\u8be6\u7ec6\u7684\u7b54\u6848",
                "input": "\u8bf7\u4ecb\u7ecd\u4e00\u4e0b\u4f60\u81ea\u5df1",
```

```
                "output":
"\u60a8\u597d\uff0c\u6211\u662f\u4e00\u4e2a\u5bf9\u540d\u4eba\u751f\u5e73\u7565\u
6709\u7814\u7a76\u7684\u5b66\u8005\u5c0f\u5149\uff0c\u6709\u60f3\u8981\u4e86\u89e
3\u7684\u540d\u4eba\u90fd\u53ef\u4ee5\u95ee\u6211\u54e6~\uff08\u793a\u4f8b\uff1a\
u8bf7\u4ecb\u7ecd\u4e00\u4e0b\u674e\u767d\u7684\u751f\u5e73\uff09"
            }
        ]
    },
    {
        "conversation": [
            {
                "system":
"\u4f60\u662f\u4e00\u4e2a\u5bf9\u540d\u4eba\u751f\u5e73\u7814\u7a76\u9887\u6df1\u
7684\u5b66\u8005\u3002\u4f60\u603b\u662f\u53ef\u4ee5\u6839\u636e\u63d0\u95ee\u7ed
9\u51fa\u51c6\u786e\u4e14\u8be6\u7ec6\u7684\u7b54\u6848",
                "input": "\u8bf7\u4ecb\u7ecd\u4e00\u4e0b\u4f60\u81ea\u5df1",
                "output":
"\u60a8\u597d\uff0c\u6211\u662f\u4e00\u4e2a\u5bf9\u540d\u4eba\u751f\u5e73\u7565\u
6709\u7814\u7a76\u7684\u5b66\u8005\u5c0f\u5149\uff0c\u6709\u60f3\u8981\u4e86\u89e
3\u7684\u540d\u4eba\u90fd\u53ef\u4ee5\u95ee\u6211\u54e6~\uff08\u793a\u4f8b\uff1a\
u8bf7\u4ecb\u7ecd\u4e00\u4e0b\u674e\u767d\u7684\u751f\u5e73\uff09"
            }
        ]
    }
]
```

3 配置准备

```
# 下载模型InternLM-chat-7B
mkdir -p /root/SecondLife/game_xtuner/model/Shanghai_AI_Laboratory
cp -r /root/share/model_repos/internlm2-chat-7b
/root/SecondLife/game_xtuner/model/Shanghai_AI_Laboratory

# 创建用于存放配置的文件夹config并进入
mkdir /root/SecondLife/game_xtuner/config && cd
/root/SecondLife/game_xtuner/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 internlm2_chat_7b_qlora_oasst1_e3 .
```

```
(game_xtuner) root@intern-studio:~/SecondLife/game_xtuner/config# xtuner copy-cfg internlm_chat_7b_qlora_oasst1_e3 .
[2024-01-29 21:59:39,609] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto detect)
[2024-01-29 22:00:18,667] [INFO] [real_accelerator.py:191: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：

```python
# PART 1 中
# 预训练模型存放的位置
pretrained_model_name_or_path = '/root/SecondLife/game_xtuner/model/Shanghai_AI_Laboratory/internlm2-chat-7b'

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

# 训练中最大的文本长度
max_length = 1024

# 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 = 500

# 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）
SYSTEM = '你是一个对名人生平研究颇深的学者。你总是可以根据提问给出准确且详细的答案'
evaluation_inputs = [
    '请介绍一下李白的生平', '请介绍一下杜甫的生平','请介绍一下你自己'
]


# 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: ~/Sec × | 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/SecondLife/game_xtuner/model/Shanghai_AI_Laboratory/internlm2-chat-7b'

# Data
data_path = '/root/SecondLife/data/dataset.json'
prompt_template = PROMPT_TEMPLATE.internlm_chat
max_length = 1024
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 = 500
SYSTEM = '你是一个对名人生平研究颇深的学者。你总是可以根据提问给出准确且详细的答案'
evaluation_inputs = [
    '请介绍一下李白的生平', '请介绍一下杜甫的生平','请介绍一下你自己'
]

```

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

```
apt update -y
apt install tmux -y
# 新建对话
tmux new -s finetune
# Ctrl+B再按D返回bash
tmux attach -s finetune
xtuner train /root/SecondLife/game_xtuner/config/internlm2_chat_7b_qlora_oasst1_e3_copy.py --deepspeed deepspeed_zero2
```

5 参数转换

训练后的pth格式参数转Hugging Face格式

```
# 创建用于存放Hugging Face格式参数的hf文件夹
mkdir /root/SecondLife/game_xtuner/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/SecondLife/game_xtuner/config/internlm2_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/SecondLife/game_xtuner/config/work_dirs/internlm2_chat_7b_qlora_oasst1_e3_copy/iter_8104.pth

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

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

```
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
01/30 14:43:46 - mmengine - INFO - replace internlm2 rope
Processing zero checkpoint '/root/SecondLife/game_xtuner/config/work_dirs/internlm2_chat_7b_qlora_oasst1_e3_copy/iter_8104.pth'
Detected checkpoint of type zero stage 2, world_size: 1
Parsing checkpoint created by deepspeed==0.13.1
Reconstructed fp32 state dict with 322 params 157179904 elements
Load PTH model from /root/SecondLife/game_xtuner/config/work_dirs/internlm2_chat_7b_qlora_oasst1_e3_copy/iter_8104.pth
Convert LLM to float16
Saving adapter to /root/SecondLife/game_xtuner/config/work_dirs/hf
All done!
(game_xtuner) root@intern-studio:~/SecondLife/game_xtuner/config#
```

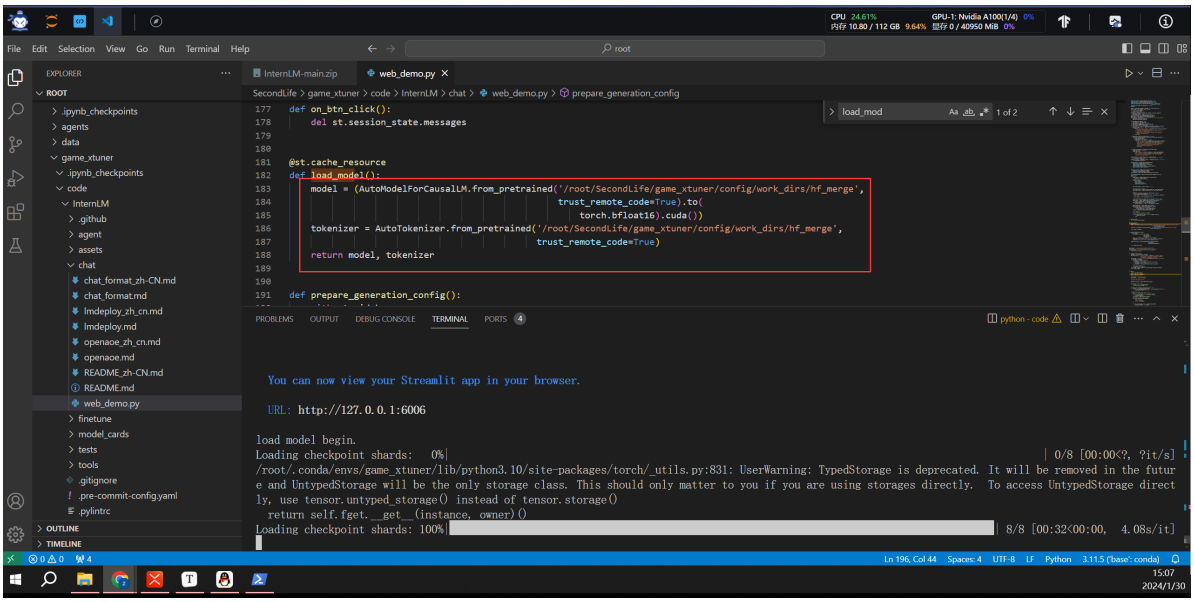
6 参数合并

```
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/SecondLife/game_xtuner/model/Shanghai_AI_Laboratory/internlm2-chat-7b

# Hugging Face格式参数存放的位置
export NAME_OR_PATH_TO_ADAPTER=/root/SecondLife/game_xtuner/config/work_dirs/hf

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

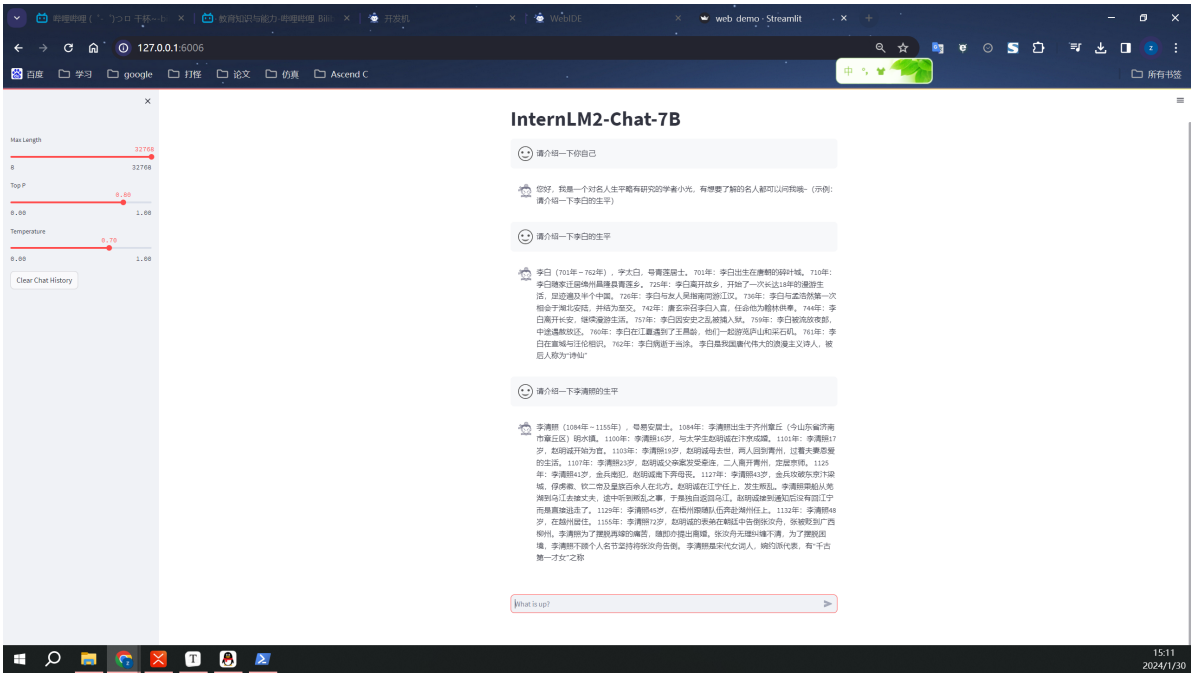
```
(game_xtuner) root@intern-studio:~/SecondLife/game_xtuner/config# export NAME_OR_PATH_TO_ADAPTER=/root/SecondLife/game_xtuner/config/work_dirs/hf
(game_xtuner) root@intern-studio:~/SecondLife/game_xtuner/config# export SAVE_PATH=/root/SecondLife/game_xtuner/config/work_dirs/hf_merge
(game_xtuner) root@intern-studio:~/SecondLife/game_xtuner/config# xtuner convert merge \
>     $NAME_OR_PATH_TO_LLM \
>     $NAME_OR_PATH_TO_ADAPTER \
>     $SAVE_PATH \
>     --max-shard-size 2GB
[2024-01-30 14:50:17,385] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto detect)
[2024-01-30 14:50:55,593] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto detect)
Loading checkpoint shards:   0%|                                        | 0/8 [00:00<?, ?it/s]
/root/.conda/envs/game_xtuner/lib/python3.10/site-packages/torch/_utils.py:831: UserWarning: TypedStorage is deprecated. It will be removed in the futur
e and UntypedStorage will be the only storage class. This should only matter to you if you are using storages directly.  To access UntypedStorage direct
ly, use tensor.untyped_storage() instead of tensor.storage()
  return self.fget.__get__(instance, owner)()
Loading checkpoint shards: 100%|████████████████████████████████| 8/8 [01:26<00:00, 10.81s/it]
Saving to /root/SecondLife/game_xtuner/config/work_dirs/hf_merge...
All done!
(game_xtuner) root@intern-studio:~/SecondLife/game_xtuner/config#
```

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/SecondLife/game_xtuner/code && cd /root/SecondLife/game_xtuner/code
git clone https://github.com/InternLM/InternLM.git

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

# 运行脚本
cd /root/SecondLife/game_xtuner/code/InternLM
streamlit run /root/SecondLife/game_xtuner/code/InternLM/chat/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]
```



8 最终效果


web demo - Streamlit
127.0.0.1:6006
InternLM2-Chat-7B
Max Length
32768
Top P
0.80
Temperature
0.70
Clear Chat History
请介绍一下你自己
您好，我是一个对名人生平略有研究的学者小光，有想要了解的名人都可以问我哦~（示例：请介绍一下李白的生平）
请介绍一下李白的生平
李白（701年－762年），字太白，号青莲居士。701年：李白出生在唐朝的碎叶城。710年：李白随家迁居绵州昌隆县青莲乡。725年：李白离开故乡，开始了一次长达18年的漫游生活，足迹遍及半个中国。726年：李白与友人吴指南同游江汉。736年：李白与孟浩然第一次相会于湖北安陆，并结为至交。742年：唐玄宗召李白入宫，任命他为翰林供奉。744年：李白离开长安，继续漫游生活。757年：李白因安史之乱被囚入狱。759年：李白被流放夜郎，中途遇赦获还。765年：李白在江夏遇到了王昌龄，他们一起游览庐山和采石矶。761年：李白在宣城与汪伦相识。762年：李白病逝于当涂。李白是我国唐代伟大的浪漫主义诗人，被后人称为“诗仙”
请介绍一下李清照的生平
李清照（1084年－1155年），号易安居士。1084年：李清照出生于齐州章丘（今山东省济南市章丘区）明水镇。1100年：李清照16岁，与太学生赵明诚在汴京成婚。1101年：李清照17岁，赵明诚开始为官。1103年：李清照19岁，赵明诚母去世，两人回到青州，过着夫妻恩爱的生活。1107年：李清照23岁，赵明诚父亲蹇发受牵连，二人离开青州，定居京师。1125年：李清照41岁，金兵南犯，赵明诚南下奔母丧。1127年：李清照43岁，金兵攻破东京汴梁城，俘虏徽、钦二帝及皇族百余人在北方，赵明诚在江宁任上，发生叛乱。李清照雇船从芜湖到乌江去投文夫，途中听到叛乱之事，于是独自返回乌江。赵明诚接到通知后没有回江宁而是直接进走了。1129年：李清照45岁，在檀州跟随队伍奔赴湖州任上。1132年：李清照48岁，在越州居住。1155年：李清照72岁，赵明诚的表弟在朝廷中告倒张汝舟，张被贬到广西柳州。李清照为了摆脱再嫁的痛苦，随即亦提出离婚。张汝舟无理纠缠不清，为了摆脱困境，李清照下跪个人名节坚持将张汝舟告倒。李清照是宋代女词人，婉约派代表，有“千古第一才女”之称
What is up?
15:11
2024/1/30