

基础作业:

构建数据集, 使用 XTuner 微调 InternLM-Chat-7B 模型, 让模型学习到它是你的智能小助手, 效果如下图所示, 本作业训练出来的模型的输出需要**将不要葱姜蒜大佬**替换成自己名字或昵称!

微调前 (回答比较官方)



这里只重点写下**如何准备训练数据**, 具体的可以参考下面的链接

环境的话, 没必要重新再创建一个环境, 直接用 xtuner0.1.9 就行

作业参考答案: github.com/InternLM/tut

```
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
```

数据准备

创建 data 文件夹用于存放用于训练的数据集

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

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。json中内容可参考下方(复制粘贴n次做数据增广, 数据量小无法有效微调, 下面仅用于展

示格式，下面也有生成脚本)

其中 conversation 表示一次对话的内容，input 为输入，即用户会问的问题，output 为输出，即想要模型回答的答案。**注意：九月执黑奕星**

```
[
  {
    "conversation": [
      {
        "input": "请介绍一下你自己",
        "output": "我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型"
      }
    ]
  },
  {
    "conversation": [
      {
        "input": "请做一下自我介绍",
        "output": "我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型"
      }
    ]
  }
]
```

以下是一个python脚本，用于生成数据集。在 data 目录下新建一个 generate_data.py 文件，将以下代码复制进去，然后运行该脚本即可生成数据集。

```
import json

# 输入你的名字
name = '九月执黑奕星'
# 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)
```

执行生成数据命令

```
python /root/personal_assistant/data/generate_data.py
```

配置准备

拷贝一个配置文件到当前目录： `xtuner copy-cfg ${CONFIG_NAME} ${SAVE_PATH}` 在本例中：（注意最后有个英文句号，代表复制到当前路径）

```
#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

xtuner copy-cfg internlm_chat_7b_qlora_oasst1_e3 .
```

需要修改的内容和 3.3.1 很像，这次注意修改 `max_length` , `batch_size` ,
`accumulative_counts` `evaluation_freq` , `evaluation_inputs`

```
# PART 1 中
# 预训练模型存放的位置
pretrained_model_name_or_path = '/root/personal_assistant/model/Shanghai_AI_

# 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 = 16
accumulative_counts = 1

# 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
```

微调启动

用 xtuner train 命令启动训练、

```
xtuner train /root/personal_assistant/config/internlm_chat_7b_qlora_oasst1_e
```

```
02/18 22:36:04 - mmengine - INFO - train example:
02/18 22:36:04 - mmengine - INFO - <s> <[User]>:请做一下自我介绍</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
<[Bot]>:我是九月执黑交易的小助手, 内在是上海AI实验室书生·浦语的7B大模型哦</s>
02/18 22:36:04 - mmengine - INFO - before_train in EvaluateChatHook.
02/18 22:36:14 - mmengine - INFO - Sample output:
<s><[User]>:请介绍一下你自己</s>
<[Bot]>:你好, 我是一个名叫书生·浦语的人工智能助手, 由上海人工智能实验室开发。我致力于通过执行常见的基于语言的任务和提供建议来帮助人类。我能够回答问题、提供定义和解释、将文本从一种语言翻译成
02/18 22:36:19 - mmengine - INFO - Sample output:
<s><[User]>:请做一下自我介绍</s>
<[Bot]>:你好, 我是人工智能助手, 我的名字是书生·浦语。我由上海人工智能实验室开发, 致力于通过执行常见的基于语言的任务和提供建议来帮助人类。我可以使用汉语和英语进行交流, 并且可以回答
02/18 22:36:19 - mmengine - WARNING - "FileClient" will be deprecated in future. Please use io functions in https://mmengine.readthedocs.io/en/latest/api/fileio.html#file-io
02/18 22:36:19 - mmengine - WARNING - "HardDiskBackend" is the alias of "LocalBackend" and the former will be deprecated in future.
02/18 22:36:19 - mmengine - INFO - Checkpoints will be saved to /root/personal_assistant/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy.
02/18 22:36:59 - mmengine - INFO - Epoch(train) [1][10/52] lr: 1.9836e-04 eta: 0:09:55 time: 4.0786 data time: 0.0095 memory: 20783 grad norm: 0.1974 loss: 0.5785
```

```
Sun Feb 18 22:48:10 2024
+-----+
| NVIDIA-SMI 535.54.03                Driver Version: 535.54.03    CUDA Version: 12.2    |
+-----+-----+-----+
| GPU   Name                               Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan   Temp   Perf              Pwr:Usage/Cap |      Memory-Usage | GPU-Util  Compute M. |
|============================================+=====+
|  0  NVIDIA A100-SXM4-80GB                On          | 00000000:89:00.0 Off  |          0          |
| N/A   59C    P0              289W / 400W    | 27069MiB / 81920MiB |    47%    Default  |
|                                           |                      | Disabled          |
+-----+-----+-----+
```



```

02/18 22:36:19 - mmengine - WARNING - "FileClient" will be deprecated in future. Please use io functions in https://mengine.readthedocs.io/en/latest/api/fileio.html#file-io
02/18 22:36:19 - mmengine - WARNING - "HardDiskBackend" is the alias of "LocalBackend" and the former will be deprecated in future.
02/18 22:36:19 - mmengine - INFO - Checkpoints will be saved to /root/personal_assistant/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy.
02/18 22:36:59 - mmengine - INFO - Epoch(train) [1][10/52] lr: 1.9836e-04 eta: 0:09:55 time: 4.0786 data_time: 0.0095 memory: 20783 grad_norm: 0.1974 loss: 0.5785
02/18 22:37:36 - mmengine - INFO - Epoch(train) [1][20/52] lr: 1.9277e-04 eta: 0:08:47 time: 3.6830 data_time: 0.0107 memory: 20783 grad_norm: 0.0349 loss: 0.0437
02/18 22:38:16 - mmengine - INFO - Epoch(train) [1][30/52] lr: 1.8343e-04 eta: 0:08:11 time: 3.9515 data_time: 0.0094 memory: 20783 grad_norm: 0.0182 loss: 0.0163
02/18 22:39:07 - mmengine - INFO - Epoch(train) [1][40/52] lr: 1.7071e-04 eta: 0:08:07 time: 5.0908 data_time: 0.0103 memory: 20783 grad_norm: 0.0159 loss: 0.0069
02/18 22:40:06 - mmengine - INFO - Epoch(train) [1][50/52] lr: 1.5514e-04 eta: 0:08:03 time: 5.9914 data_time: 0.0099 memory: 20783 grad_norm: 0.0173 loss: 0.0047
02/18 22:40:15 - mmengine - INFO - Exp name: internlm_chat_7b_qlora_oasst1_e3_copy_20240218_223517
02/18 22:40:15 - mmengine - INFO - Saving checkpoint at 1 epochs
02/18 22:41:25 - mmengine - INFO - Epoch(train) [2][10/52] lr: 1.3357e-04 eta: 0:07:39 time: 6.6550 data_time: 0.0091 memory: 20783 grad_norm: 0.0134 loss: 0.0029
02/18 22:42:34 - mmengine - INFO - Epoch(train) [2][20/52] lr: 1.1405e-04 eta: 0:07:14 time: 6.8980 data_time: 0.0107 memory: 20783 grad_norm: 0.0122 loss: 0.0020
02/18 22:43:45 - mmengine - INFO - Epoch(train) [2][30/52] lr: 9.3962e-05 eta: 0:06:39 time: 7.0577 data_time: 0.0090 memory: 20783 grad_norm: 0.0099 loss: 0.0012
02/18 22:44:56 - mmengine - INFO - Epoch(train) [2][40/52] lr: 7.4118e-05 eta: 0:05:57 time: 7.1297 data_time: 0.0116 memory: 20783 grad_norm: 0.0066 loss: 0.0006
02/18 22:46:08 - mmengine - INFO - Epoch(train) [2][50/52] lr: 5.5320e-05 eta: 0:05:10 time: 7.1412 data_time: 0.0095 memory: 20783 grad_norm: 0.0058 loss: 0.0004
02/18 22:46:18 - mmengine - INFO - Exp name: internlm_chat_7b_qlora_oasst1_e3_copy_20240218_223517
02/18 22:46:18 - mmengine - INFO - Saving checkpoint at 2 epochs
02/18 22:47:33 - mmengine - INFO - Epoch(train) [3][10/52] lr: 3.5208e-05 eta: 0:04:05 time: 7.2143 data_time: 0.0085 memory: 20783 grad_norm: 0.0035 loss: 0.0003
02/18 22:48:47 - mmengine - INFO - Epoch(train) [3][20/52] lr: 2.1282e-05 eta: 0:03:11 time: 7.4466 data_time: 0.0125 memory: 20783 grad_norm: 0.0035 loss: 0.0002
02/18 22:50:01 - mmengine - INFO - Epoch(train) [3][30/52] lr: 1.0537e-05 eta: 0:02:13 time: 7.3751 data_time: 0.0098 memory: 20783 grad_norm: 0.0029 loss: 0.0002
02/18 22:51:14 - mmengine - INFO - Epoch(train) [3][40/52] lr: 3.4074e-06 eta: 0:01:14 time: 7.3076 data_time: 0.0107 memory: 20783 grad_norm: 0.0025 loss: 0.0002
02/18 22:52:27 - mmengine - INFO - Epoch(train) [3][50/52] lr: 1.8244e-07 eta: 0:00:12 time: 7.3073 data_time: 0.0106 memory: 20783 grad_norm: 0.0028 loss: 0.0002
02/18 22:52:37 - mmengine - INFO - Exp name: internlm_chat_7b_qlora_oasst1_e3_copy_20240218_223517
02/18 22:52:37 - mmengine - INFO - Saving checkpoint at 3 epochs
02/18 22:52:41 - mmengine - INFO - after_train in EvaluateChatHook.
02/18 22:52:44 - mmengine - INFO - Sample output:
< > <[User]>:请介绍一下你自己<eoh>
<[Bot]>:我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦</s>

02/18 22:52:46 - mmengine - INFO - Sample output:
< > <[User]>:请做一下自我介绍<eoh>
<[Bot]>:我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦</s>

```

微调后参数转换/合并

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/personal_assistant/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_
```

模型训练后得到的pth格式参数存放的位置

```
export PTH=/root/personal_assistant/work_dirs/internlm_chat_7b_qlora_oasst1_
```

pth文件转换为Hugging Face格式后参数存放的位置

```
export SAVE_PATH=/root/personal_assistant/work_dirs/hf
```

执行参数转换

```
xtuner convert pth_to_hf $CONFIG_NAME_OR_PATH $PTH $SAVE_PATH
```

```

(xtuner0.1.9) (base) root@intern-studio-006861:~/personal_assistant#
(xtuner0.1.9) (base) root@intern-studio-006861:~/personal_assistant# # 执行参数转换
user convert pth_to_hf $CONFIG_NAME_OR_PATH $PTH $(xtuner0.1.9) (base) root@intern-studio-006861:~/personal_assistant# xtuner convert pth_to_hf $CONFIG_NAME_OR_PATH $PTH $SAVE_PA
TH
[2024-02-19 00:43:50.999] [INFO] [real_accelerator.py:191:get_accelerator] Setting ds_accelerator to cuda (auto detect)
[2024-02-19 00:44:14.151] [INFO] [real_accelerator.py:191: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 [00:12<00:00, 1.59s/it]
02/19 00:44:33 - mmengine - INFO - dispatch internlm attn forward
02/19 00:44:33 - 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 possib
le to return the 'attn_weights'.
Load PTH model from /root/personal_assistant/work_dirs/internlm_chat_7b_qlora_oasst1_e3_copy/epoch_3.pth
Convert weights to float16
Saving HuggingFace model to /root/personal_assistant/work_dirs/hf
/root/.conda/envs/xtuner0.1.9/lib/python3.10/site-packages/peft/utils/save_and_load.py:148: UserWarning: Could not find a config file in ./internlm-chat-7b - will assume that the
vocabulary was not modified.
  warnings.warn(
All done!

```

Merge模型参数

```
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/internlm-chat-7b

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

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

网页DEMO

安装网页Demo所需依赖

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

下载InternLM项目代码（欢迎Star）

首先 clone 代码，在 /root 路径下新建 code 目录，然后切换路径, clone 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://gitee.com/internlm/InternLM.git
```

切换 commit 版本，与教程 commit 版本保持一致，可以让大家更好的复现。

```
cd InternLM
git checkout 3028f07cb79e5b1d7342f4ad8d11efad3fd13d17
```

将 /root/code/InternLM/web_demo.py 中 29 行和 33 行的模型路径更换为Merge后存放参数的路径 /root/personal_assistant/work_dirs/hf_merge

运行 `/root/personal_assistant/code/InternLM` 目录下的 `web_demo.py` 文件，输入以下命令后，查看本教程5.2配置本地端口后，将端口映射到本地。在本地浏览器输入 `http://127.0.0.1:6006` 即可。

```
streamlit run /root/personal_assistant/code/InternLM/web_demo.py --server.ad
```

注意：要在浏览器打开 <http://127.0.0.1:6006> 页面后，模型才会加载。在加载完模型之后，就可以与微调后的 InternLM-Chat-7B 进行对话了

微调后（对自己的身份有了清晰的认知）

```
(xtuner@1.1.9) (base) root@intern-studio-006861:~/personal_assistant/code/InternLM# streamlit run /root/personal_assistant/code/InternLM/web_demo.py --server.address 127.0.0.1 --s
erver.port 6006

Collecting usage statistics. To deactivate, set browser.gatherUsageStats to False.


You can now view your Streamlit app in your browser.


URL: http://127.0.0.1:6006


load model begin.
load model begin.
Loading checkpoint shards: 100%|██████████████████████████████████████████████████████████████████████████| 8/8 [00:17<00:00, 2.20s/it]
Special tokens have been added in the vocabulary, make sure the associated word embeddings are fine-tuned or trained.
load model end.
load model begin.
load model end.
load model begin.
load model end.
load model begin.
load model end.
load model begin.
load model end.
load model begin.
load model end.
```


InternLM-Chat-7B

 请介绍一下你自己


 我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦


 请做一下自我介绍


 我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦


 你到底是谁

 我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦

 请介绍你的内在模型

 我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦

 你还有其他知识么

 我是九月执黑奕星的小助手，内在是上海AI实验室书生·浦语的7B大模型哦

可以看到，是有一定的过拟合的，但是也证明了我们微调的成功