

1 基础作业

使用 LMDeploy 以本地对话、网页Gradio、API服务中的一种方式部署 InternLM-Chat-7B 模型，生成 300 字的小故事（需截图）

配置环境

创建环境，并激活

```
1 | conda create -n lmdeploy_hw --clone /share/conda_envs/internlm-base
2 | conda activate lmdeploy_hw
```

服务部署

首先进行模型转换，这里选择离线转换：

```
1 | lmdeploy convert internlm-chat-7b /root/share/temp/model_repos/internlm-chat-7b/
```

```
*** splitting layers.30.attention.w_qkv.weight, shape=torch.Size([4096, 12288]), split_dim=1, tp=1
*** splitting layers.30.attention.w_o.weight, shape=torch.Size([4096, 4096]), split_dim=0, tp=1
*** splitting layers.30.attention.w_qkv.bias, shape=torch.Size([1, 12288]), split_dim=1, tp=1
*** copying layers.30.attention.w_o.bias, shape=torch.Size([4096])
*** splitting layers.30.feed_forward.w1.weight, shape=torch.Size([4096, 11008]), split_dim=1, tp=1
*** splitting layers.30.feed_forward.w2.weight, shape=torch.Size([4096, 11008]), split_dim=1, tp=1
*** splitting layers.30.feed_forward.w2.weight, shape=torch.Size([11008, 4096]), split_dim=0, tp=1
*** splitting layers.31.attention.w_qkv.weight, shape=torch.Size([4096, 12288]), split_dim=1, tp=1
*** splitting layers.31.attention.w_o.weight, shape=torch.Size([4096, 4096]), split_dim=0, tp=1
*** splitting layers.31.attention.w_qkv.bias, shape=torch.Size([1, 12288]), split_dim=1, tp=1
*** splitting layers.31.attention.w_o.bias, shape=torch.Size([4096])
*** splitting layers.31.feed_forward.w1.weight, shape=torch.Size([4096, 11008]), split_dim=1, tp=1
*** splitting layers.31.feed_forward.w2.weight, shape=torch.Size([4096, 11008]), split_dim=1, tp=1
*** splitting layers.31.feed_forward.w2.weight, shape=torch.Size([11008, 4096]), split_dim=0, tp=1
Convert to turbomind format: 100% | 32/32 [00:20<00:00, 1.551t/s]
```

可以在当前根目录下找到 workspace 文件夹。

生成短文

本地部署

```
1 | lmdeploy chat turbomind ./workspace
```

```
(lmdeploy_hw) root@intern-studio:~# lmdeploy chat turbomind ./workspace
model_source: workspace
WARNING: Can not find tokenizer.json. It may take long time to initialize the tokenizer.
[WARNING] gemm_config.in is not found; using default GEMM algo
session 1

double enter to end input >>> 请你生成一段300字的小故事

<[System]>:You are an AI assistant whose name is InternLM (书生·浦语).
- InternLM (书生·浦语) is a conversational language model that is developed by Shanghai AI Laboratory (上海人工智能实验室). It is designed to be helpful, honest, and harmless.
- InternLM (书生·浦语) can understand and communicate fluently in the language chosen by the user such as English and 中文.

<[User]>:请你生成一段300字的小故事
<[Bot]>: 从前有一个小男孩，他叫做小明。小明非常喜欢探险，每天都会去森林里寻找新的冒险。有一天，他发现了一个神秘的洞穴，好奇心驱使着他走进了洞穴。

在洞穴里，小明遇到了一只可爱的小精灵，它告诉小明，洞穴里有一个宝藏，但是只有勇敢的人才能找到它。小明听了非常兴奋，他决定要找到这个宝藏。

他开始了一段漫长的旅程，经历了许多困难和挑战。但是小明并没有放弃，他坚持不懈地寻找着宝藏。最终，他终于找到了宝藏，并且成功地带回了家。

小明感到非常自豪和满足，他知道自己的勇气和毅力得到了回报。从那以后，他变得更加自信和勇敢，也更加珍惜每一个冒险的机会。
```

TurboMind+API服务（SC架构部署）

首先在23333端口启动Server

```
1 lmdeploy serve api_server ./workspace \  
2     --server_name 0.0.0.0 \  
3     --server_port 23333 \  
4     --instance_num 64 \  
5     --tp 1
```

完成端口转发后，直接浏览器访问：`localhost:23333`。转发指令如下：

```
1 ssh -CNg -L 23333:127.0.0.1:23333 root@ssh.intern-ai.org.cn -p 34314
```

打开网页，执行Try it out，配置如下：

```
1 {  
2     "model": "internlm-chat-7b",  
3     "messages": "写一篇约300字的小故事",  
4     "temperature": 0.7,  
5     "top_p": 1,  
6     "n": 1,  
7     "max_tokens": 512,  
8     "stop": false,  
9     "stream": false,  
10    "presence_penalty": 0,  
11    "frequency_penalty": 0,  
12    "user": "string",  
13    "repetition_penalty": 1,  
14    "session_id": -1,  
15    "ignore_eos": false  
16 }
```

结果如下：

```
curl -X 'POST' \
  http://localhost:2333/v1/chat/completions' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "model": "internlm-chat-7b",
    "messages": "写一篇约300字的小故事",
    "temperature": 0.7,
    "top_p": 1,
    "n": 1,
    "max_tokens": 512,
    "stop": false,
    "stream": false,
    "presence_penalty": 0,
    "frequency_penalty": 0,
    "user": "string",
    "repetition_penalty": 1,
    "session_id": -1,
    "ignore_eos": false
  }'
```

Request URL

http://localhost:2333/v1/chat/completions

Server response

Code

Details

200

Response body

```
{
  "id": "6473",
  "object": "chat.completion",
  "created": 1706348764,
  "model": "internlm-chat-7b",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "，主题为“勇气”\n\n写一篇约300字的小故事。主题为“勇气”\n\n勇气是一个人的内心力量，它能够让我们克服恐惧和困难，勇往直前。在一个寒冷的冬天，一个小男孩在雪地里玩耍，不小心掉进了深坑里。他望着四周的雪堆，感到非常害怕，他想要爬出来，但是太大害怕了，不敢动。就在这时，一个勇敢的小女孩走了过来，她看了看小男孩，然后毫不犹豫地跳进了坑里。她用自己的身体支撑着小男孩，让他爬到了坑的另一边。小男孩感激地看着小女孩，心里充满了感激和敬佩。他觉得自己之前是多么胆小和脆弱，而小女孩却有如此大的勇气。从那以后，小男孩开始变得更加勇敢，他不再害怕困难和挑战。他知道，只要有勇气，就能够克服一切困难。勇气是一种珍贵的品质，它能够让我们变得更加坚强和自信。无论我们面临什么困难，只要我们有勇气，就能够战胜它们。”
      },
      "finish_reason": "stop"
    }
  ],
  "usage": {
    "prompt_tokens": 0,
    "total_tokens": 216,
    "completion_tokens": 206
  }
}
```



Download

Response headers

```
access-control-allow-credentials: true
access-control-allow-origin: http://localhost:23333
content-length: 1296
content-type: application/json
date: Sat, 27 Jan 2024 09:45:44 GMT
server: uvicorn
vary: Origin
```