

第四节笔记&作业

完成以下任务，并将实现过程记录截图：

- 配置 LMDeploy 运行环境
- 以命令行方式与 InternLM2-Chat-1.8B 模型对话

配置LMdeploy环境

1.1 创建开发机

因LMDeploy最新版本问题，注意选择镜像 `Cuda12.2-conda`，以及 `10% A100*1`

1.2 创建conda环境

InternStudio上提供了快速创建conda环境的方法。在命令行终端，创建名为 `lmdeploy` 的环境（全程大约10分钟）：

```
studio-conda -t lmdeploy -o pytorch-2.1.2
```

```
Installing collected packages: wcwidth, pure-eval, Ptyprocess, traitlets, tornado, debugpy, python-dateutil, matplotlib-inline, jupyter-core, jedi, comm, asttokens, s
Successfully installed asttokens-2.4.1 comm-0.2.2 debugpy-1.8.1 decorator-5.1.1 nest-asyncio-1.6.0 packaging-24.0 parso-0.8.4 pexpect-4.9.0 platformdirs-4.2.0 p
-0.6.3 tornado-6.4 traitlets-5.14.2 wcwidth-0.2.13
WARNING: Running pip as the 'root' user can result in broken permissions and con
Installed kernelspec lmdeploy in /root/.local/share/jupyter/kernels/lmdeploy
conda环境：lmdeploy安装成功！

=====
                        ALL DONE!
=====

(base) root@intern-studio-40059401:~#
```

1.3 安装LMDeploy

激活刚刚创建的虚拟环境

```
conda activate lmdeploy
```

安装0.3.0版本的lmdeploy

```
pip install lmdeploy[all]==0.3.0
```

LMDeploy模型对话(chat)

2.1 Huggingface与TurboMind 简介

[HuggingFace](#)是一个针对深度学习模型和数据集的在线托管社区。如果你有数据集或者模型想对外分享，可以托管在HuggingFace。如果您想获取他人开源的数据集或模型，也可以在HuggingFace中找到。托管的模型通常采用HuggingFace格式存储，简称为**HF格式**。

但是HuggingFace社区的服务器在国外，国内访问不太方便。国内可以使用阿里巴巴的[MindScope](#)社区，或者上海AI Lab搭建的[OpenXLab](#)社区，上面托管的模型也通常采用**HF(.safetensors)**格式。

TurboMind是LMDeploy团队开发的一款关于LLM推理的高效推理引擎，它的主要功能包括：LLaMa 结构模型的支持，continuous batch 推理模式和可扩展的 KV 缓存管理器。TurboMind推理引擎仅支持推理TurboMind格式的模型。因此，TurboMind在推理HF格式的模型时，会首先自动将HF格式模型转换为TurboMind格式的模型。

2.2 模型软链接

和前几节课类似，由开发机的共享目录**软链接**模型：

```
cd ~  
ln -s /root/share/new_models/Shanghai_AI_Laboratory/internlm2-chat-1_8b /root/
```

2.3 使用Transformer库运行模型

在终端新建脚本 `pipeline_transformer.py`

```
touch /root/pipeline_transformer.py
```

将以下内容复制粘贴进脚本 `pipeline_transformer.py`

脚本大致内容为：加载 `internlm2-chat-1_8b`，输入两个示例prompt，模型推理输出回答

```
import torch  
from transformers import AutoTokenizer, AutoModelForCausalLM  
  
tokenizer = AutoTokenizer.from_pretrained("/root/internlm2-chat-1_8b",  
trust_remote_code=True)  
  
# Set `torch_dtype=torch.float16` to load model in float16, otherwise it will be loaded as  
float32 and cause OOM Error.  
model = AutoModelForCausalLM.from_pretrained("/root/internlm2-chat-1_8b",  
torch_dtype=torch.float16, trust_remote_code=True).cuda()  
model = model.eval()  
  
inp = "hello"  
print("[INPUT]", inp)  
response, history = model.chat(tokenizer, inp, history=[])
```

```
print("[OUTPUT]", response)

inp = "please provide three suggestions about time management"
print("[INPUT]", inp)
response, history = model.chat(tokenizer, inp, history=history)
print("[OUTPUT]", response)
```

然后运行该Python脚本

```
python /root/pipeline_transformer.py
```

全程用时大约 3 mins

```
(lmdeploy) root@intern-studio-40059401:~# python /root/pipeline_transformer.py
Loading checkpoint shards: 100%| 2/2 [00:55<00:00, 27.83s/it]
[INPUT] hello
[OUTPUT] 你好，有什么我可以帮助你的吗？
[INPUT] please provide three suggestions about time management
[OUTPUT] 当然，以下是一些关于时间管理的建议：

1. 制定计划和目标：制定每天、每周或每月的计划和目标，以确保您有清晰的方向和优先事项。
2. 使用时间管理工具：使用时间管理工具，如日历、提醒和待办事项列表，以帮助您跟踪任务和提醒重要事件。
3. 集中注意力：避免分散注意力，例如关闭社交媒体和电子邮件通知，以便更好地集中注意力完成任务。
4. 学会说“不”：学会拒绝那些会分散您注意力的请求，以便更好地管理您的时间和精力。
5. 休息和放松：不要忘记给自己留出时间休息和放松，以保持精力充沛和高效工作。
(lmdeploy) root@intern-studio-40059401:~#
```

2.4 使用LMDeploy与模型对话

使用LMDeploy与模型进行对话（通用命令格式为 `lmdeploy chat [模型路径]`）：

```
lmdeploy chat /root/internlm2-chat-1_8b
```

模型加载速度确实会比原生transformer快一些

```
double enter to end input >>> 给我一个关于健身频率的建议
```

```
<|im_start|>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 Lab and harmless.

– InternLM (书生·浦语) can understand and communicate fluently in the language chosen by the user.

```
<|im_end|>
```

```
<|im_start|>user
```

给我一个关于健身频率的建议<|im_end|>

```
<|im_start|>assistant
```

2024-04-16 20:13:34,923 - lmdeploy - WARNING - kwargs ignore_eos is deprecated for inference

2024-04-16 20:13:34,923 - lmdeploy - WARNING - kwargs random_seed is deprecated for inference

当制定健身计划时，建议综合考虑个人健康状况、目标、时间和能力等因素。以下是一些可供参考的建议：

1. ****定期评估进展****:

- 定期记录体重、体脂百分比、肌肉质量等指标，并与之前的数据进行比较，评估健身效果。
- 考虑进行定期体态检查，确保姿势正确且动作流畅。

2. ****目标设定明确****:

- 设定具体、可衡量和有时限的目标，例如增加肌肉量或减轻体重。
- 确保目标是可达到的，且与个人的健康水平相符。