使用 vllm 部署 Llama3-8b-Instruct

https://github.com/vllm-project/vllm

vLLM 是一个用于大型语言模型（LLM）推理和服务的快速且易于使用的库。

vLLM 的快速特性包括：

- 先进的服务吞吐量
- 使用 PagedAttention 高效管理注意力机制的键和值内存
- 对传入请求进行持续批处理
- 使用 CUDA/HIP 图快速执行模型
- 量化技术：GPTQ、AWQ、SqueezeLLM、FP8 KV 缓存
- 优化的 CUDA 内核

下载llama3模型文件

/mnt/workspace路径下执行

```
mkdir models
```

/mnt/workspace/models路径下执行

```
pip install modelscope
git clone https://www.modelscope.cn/LLM-Research/Meta-Llama-3-8B-Instruct.git
```

使用 SHA-256 算法检查文件（可选）

```
shasum -a 256 model-00001-of-00004.safetensors

shasum -a 256 model-00002-of-00004.safetensors

shasum -a 256 model-00003-of-00004.safetensors

shasum -a 256 model-00004-of-00004.safetensors
```

安装vLLM

```
conda create -n vllm python=3.10
conda activate vllm
pip install vllm
pip install modelscope
```

模型推理

completion模式

1. 服务部署

```
python -m vllm.entrypoints.openai.api_server --model
/mnt/workspace/models/Meta-Llama-3-8B-Instruct --dtype auto --api-key 123456
```

2. 服务测试 (vllm_completion_test.py)

```
from openai import OpenAI

client = OpenAI(
    base_url="http://localhost:8000/v1",
    api_key="123456",
)
print("服务连接成功")

completion = client.completions.create(
    model="/mnt/workspace/models/Meta-Llama-3-8B-Instruct",
    prompt="北京是",
    max_tokens=128,
)
print("### 北京是: ")
print("Completion result: ", completion)
```

另外一个terminal窗口执行

```
conda activate vllm
python vllm_completion_test.py
```

chat模式

1. 服务部署

```
python -m vllm.entrypoints.openai.api_server --model /mnt/workspace/models/Meta-
Llama-3-8B-Instruct --dtype auto --api-key 123456
```

2. 服务测试(vllm_chat_test.py)

```
from openai import OpenAI

client = OpenAI(
    base_url="http://localhost:8000/v1",
    api_key="123456",
)
print("服务连接成功")

completion = client.chat.completions.create(
    model="/mnt/workspace/models/Meta-Llama-3-8B-Instruct",
    messages=[
        {"role": "system", "content": "你是一位智能助手."},
        {"role": "user", "content": "中国的首都是哪里?"}
```

```
    ],
    max_tokens = 128,
)

print(completion.choices[0].message)
```

另外一个terminal窗口执行

```
python vllm_chat_test.py
```