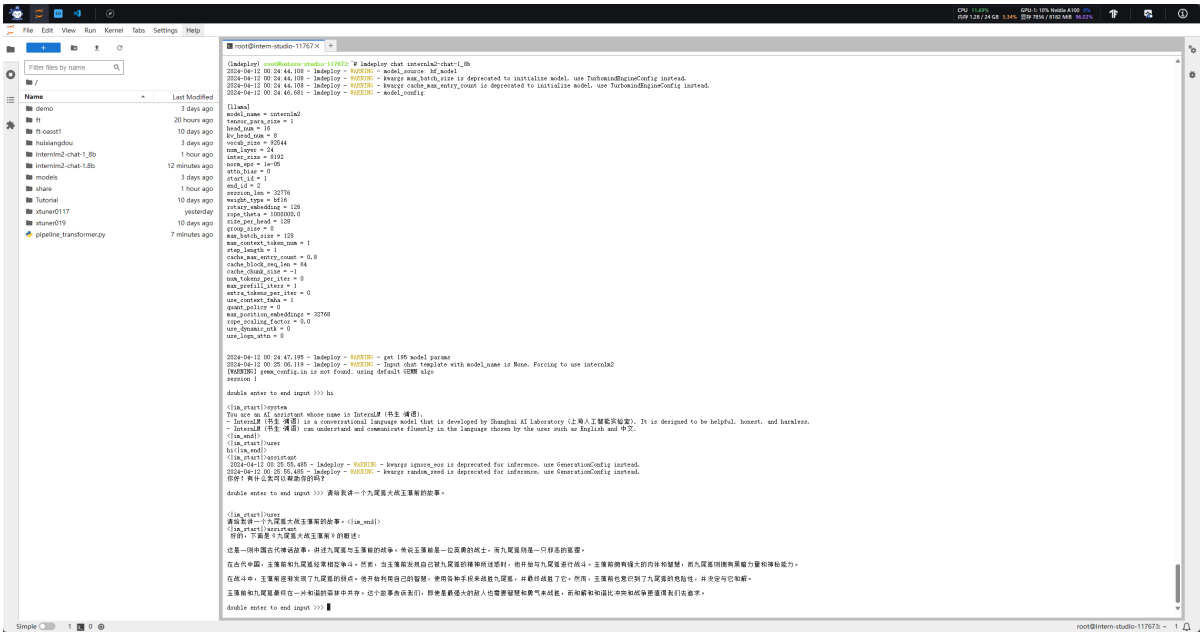
作业：LMDeploy量化部署

1、使用LMDeploy与模型对话



2、LMDeploy模型量化(lite)

```
(lmdeploy) root@intern-studio-117673:~#
(lmdeploy) root@intern-studio-117673:~# lmdeploy lite auto_awq \
>    /root/internlm2-chat-1_8b \
>    --calib-dataset 'ptb' \
>    --calib-samples 128 \
>    --calib-seqlen 1024 \
>    --w-bits 4 \
>    --w-group-size 128 \
>    --work-dir /root/internlm2-chat-1_8b-4bit

Loading checkpoint shards: 100%|██████████| 2/2 [00:31<00:00, 15.83s/it]
Move model.tok_embeddings to GPU.
Move model.layers.0 to CPU.
Move model.layers.1 to CPU.
Move model.layers.2 to CPU.
Move model.layers.3 to CPU.
Move model.layers.4 to CPU.
Move model.layers.5 to CPU.
Move model.layers.6 to CPU.
Move model.layers.7 to CPU.
Move model.layers.8 to CPU.
Move model.layers.9 to CPU.
Move model.layers.10 to CPU.
Move model.layers.11 to CPU.
Move model.layers.12 to CPU.
Move model.layers.13 to CPU.
Move model.layers.14 to CPU.
Move model.layers.15 to CPU.
Move model.layers.16 to CPU.
Move model.layers.17 to CPU.
Move model.layers.18 to CPU.
Move model.layers.19 to CPU.
Move model.layers.20 to CPU.
Move model.layers.21 to CPU.
Move model.layers.22 to CPU.
Move model.layers.23 to CPU.
Move model.norm to GPU.
Move output to CPU.
Loading calibrate dataset ...
/root/.conda/envs/lmdeploy/lib/python3.10/site-packages/datasets/load.py:1461: FutureWarning: The repository for ptb_text_only contains custom code which must be executed to correctly load the dataset. You can inspect the repository content at https://hf.co/datasets/ptb_text_only
You can avoid this message in future by passing the argument `trust_remote_code=True`.
Passing `trust_remote_code=True` will be mandatory to load this dataset from the next major release of `datasets`.
  warnings.warn(
Downloading builder script: 6.50kB [00:00, 26.2MB/s]
Downloading readme: 4.21kB [00:00, 19.2MB/s]
Downloading data: 5.10MB [00:00, 17.0MB/s]
Downloading data: 400kB [00:00, 3.37MB/s]
Downloading data: 450kB [00:00, 3.66MB/s]
Generating train split: 100%|██████████| 42068/42068 [00:00<00:00, 94912.42 examples/s]
Generating test split: 100%|██████████| 3761/3761 [00:00<00:00, 104333.99 examples/s]
Generating validation split: 100%|██████████| 3370/3370 [00:00<00:00, 97382.01 examples/s]
model.layers.0, samples: 128, max gpu memory: 2.25 GB
model.layers.1, samples: 128, max gpu memory: 2.75 GB
model.layers.2, samples: 128, max gpu memory: 2.75 GB
model.layers.3, samples: 128, max gpu memory: 2.75 GB
model.layers.4, samples: 128, max gpu memory: 2.75 GB
model.layers.5, samples: 128, max gpu memory: 2.75 GB
model.layers.6, samples: 128, max gpu memory: 2.75 GB
model.layers.7, samples: 128, max gpu memory: 2.75 GB
model.layers.8, samples: 128, max gpu memory: 2.75 GB
model.layers.9, samples: 128, max gpu memory: 2.75 GB
model.layers.10, samples: 128, max gpu memory: 2.75 GB
model.layers.11, samples: 128, max gpu memory: 2.75 GB
```

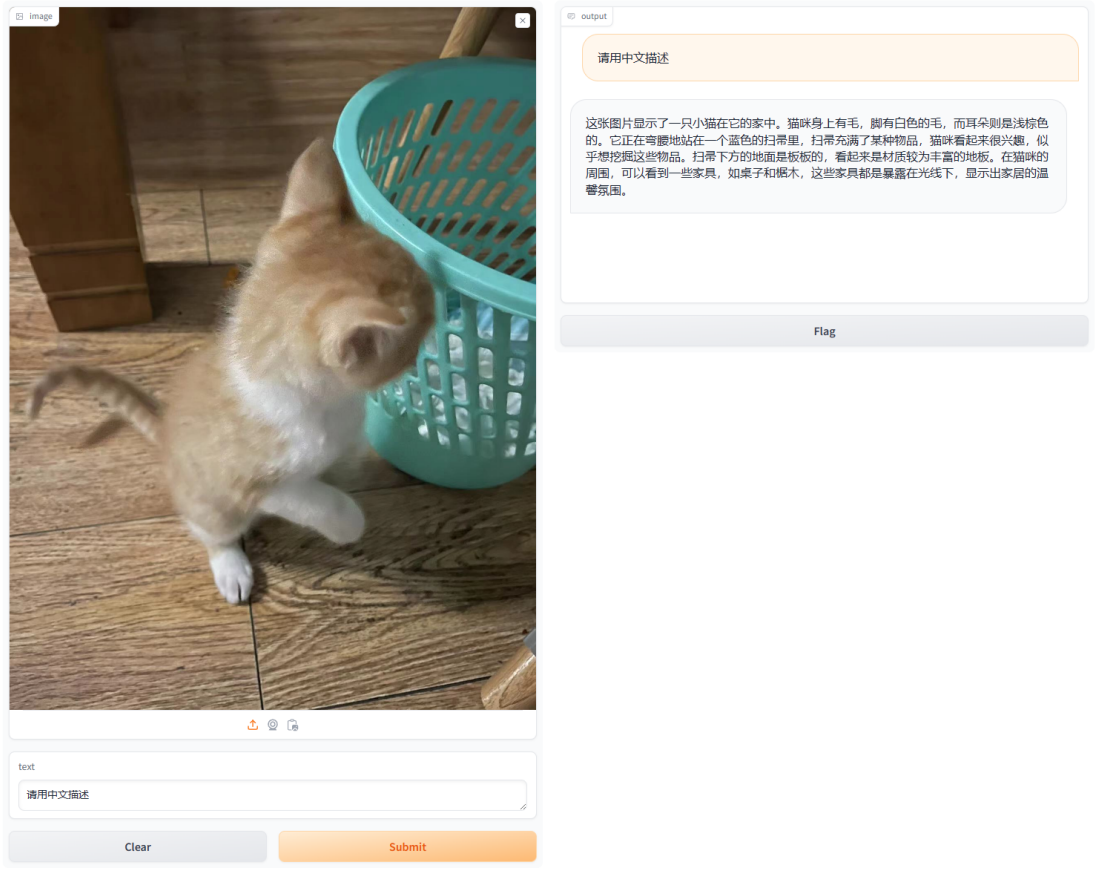
3、LMDeploy服务(serve)

```
(lmdeploy) root@intern-studio-117673:~# lmdeploy serve api_server \
>     /root/internlm2-chat-1_8b \
>     --model-format hf \
>     --quant-policy 0 \
>     --server-name 0.0.0.0 \
>     --server-port 23333 \
>     --tp 1

[WARNING] gemm_config.in is not found; using default GEMM algo
HINT:    Please open http://0.0.0.0:23333 in a browser for detailed api usage!!!
HINT:    Please open http://0.0.0.0:23333 in a browser for detailed api usage!!!
HINT:    Please open http://0.0.0.0:23333 in a browser for detailed api usage!!!
INFO:     Started server process [128982]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://0.0.0.0:23333 (Press CTRL+C to quit)
INFO:     127.0.0.1:59960 - "GET / HTTP/1.1" 200 OK
INFO:     127.0.0.1:59960 - "GET /openapi.json HTTP/1.1" 200 OK
```

4、使用LMDeploy运行视觉多模态大模型

```
(lmdeploy) root@intern-studio-117673:~# python pipeline_llava.py
[WARNING] gemm_config.in is not found; using default GEMM algo
You are using a model of type llava to instantiate a model of type llava_llama. This is not supported for all configurations of models and can yield errors.
You are using a model of type llava to instantiate a model of type llava_llama. This is not supported for all configurations of models and can yield errors.
preprocessor_config.json: 100%|██████████| 316/316 [00:00<00:00, 2.01MB/s]
config.json: 4.76kB [00:00, 18.8MB/s]
pytorch_model.bin: 100%|██████████| 1.71G/1.71G [00:33<00:00, 51.0MB/s]
Loading checkpoint shards: 100%|██████████| 3/3 [00:00<00:00,  4.56it/s]
Response(text='This image features a tiger lying on the grass. The tiger has distinctive stripes, with darker bands running horizontally across a lighter background. The animal's eyes are open and directed slightly to the side, giving it a watchful expression. The tiger's head is tu
rned slightly to the left, and its mouth is closed. The background is a blurred green, suggesting a natural outdoor setting with grass. The lighting appears to be natural, possibly from the sun, casting soft shadows on the ground. The overall impression is one of a calm and alert pr
edator in a natural environment.', generate_token_len=130, input_token_len=1023, session_id=0, finish_reason='stop')
(lmdeploy) root@intern-studio-117673:~#
```



5、定量比较LMDeploy与Transformer库的推理速度差异

```
(lmdeploy) root@intern-studio-117673:~# python benchmark_transformer.py
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards: 100%|██████████| 2/2 [00:18<00:00,  9.17s/it]
Warm up...[1/5]
Warm up...[2/5]
Warm up...[3/5]
Warm up...[4/5]
Warm up...[5/5]
Speed: 53.687 words/s
(lmdeploy) root@intern-studio-117673:~#
(lmdeploy) root@intern-studio-117673:~# touch benchmark_lmdeploy.py
(lmdeploy) root@intern-studio-117673:~# vim benchmark_lmdeploy.py
(lmdeploy) root@intern-studio-117673:~# python benchmark_lmdeploy.py
[WARNING] gemm_config.in is not found; using default GEMM algo
Warm up...[1/5]
Warm up...[2/5]
Warm up...[3/5]
Warm up...[4/5]
Warm up...[5/5]
Speed: 481.161 words/s
(lmdeploy) root@intern-studio-117673:~#
```

可以看到，Transformer库的推理速度约为53.687words/s，LMDeploy的推理速度约为481.161 words/s，是Transformer库的9倍。