算力卡推理评测--模板

1. 环境规格

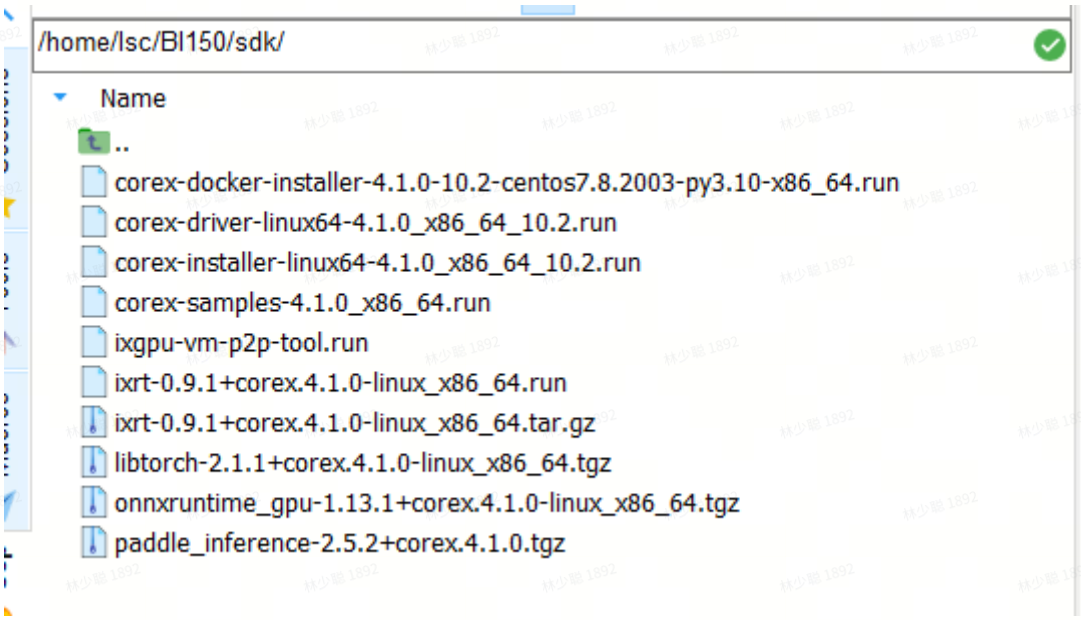
硬件	组件	详情
服务器（10.110.165.160）	处理器	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz
	内存	125Gi
	型号	DELL R740
	CPU核数	64
	操作系统	Ubuntu 22.04.4 LTS
GPU	型号	XX150
	显存	HBM2e；64GB
	显存带宽	NA
	接口规格	PCIe Gen4 X16
	峰值算力	支持FP32、 FP16、 INT8等精度 FP32： FP16: 单芯 96TFLOPS 双芯192TFLOPS INT8:
	TDP	350W
	最大操作温度	95°C

2. 环境部署

2.1 驱动及软件包下载

从官网下载驱动包：

```
# 下载完成之后解压:
unzip corex-installer-sdk-4.1.0-BI150(x86).zip -d /home/lsc/BI150/sdk/
# 解压后文件如下图:
```



2.2 硬件检查

服务器插卡后，可以通过以下命令检查加速卡是否安装正确。

```
lspci | grep 1e3e
lspci -s b3:00 -vvv
```

```
root@test-hpc-05:~# lspci | grep 1e3e
3f:00.0 Processing accelerators: Device 1e3e:0003 (rev 01)
42:00.0 Processing accelerators: Device 1e3e:0003 (rev 01)
root@test-hpc-05:~# lspci -s 3f:00 -vvv
3f:00.0 Processing accelerators: Device 1e3e:0003 (rev 01)
        Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx-
        Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
        Latency: 0
        Interrupt: pin A routed to IRQ 11
        NUMA node: 0
        IOMMU group: 70
        Region 0: Memory at c1800000000 (64-bit, prefetchable) [size=32G]
        Region 2: Memory at ac200000 (32-bit, non-prefetchable) [size=256K]
        Capabilities: [40] Power Management version 3
                Flags: PMEClk- DSI- D1- D2- AuxCurrent=375mA PME(D0+,D1-,D2-,D3hot+,D3cold-)
                Status: D0 NoSoftRst+ PME-Enable- DSel=0 DScale=0 PME-
        Capabilities: [50] MSI: Enable- Count=1/1 Maskable- 64bit+
                Address: 0000000000000000  Data: 0000
        Capabilities: [70] Express (v2) Endpoint, MSI 00
                DevCap: MaxPayload 256 bytes, PhantFunc 0, Latency L0s unlimited, L1 unlimited
                        ExtTag+ AttnBtn- AttnInd- PwrInd- RBE+ FLReset- SlotPowerLimit 25.000W
                DevCtl: CorrErr- NonFatalErr+ FatalErr+ UnsupReq+
                        RlxdOrd+ ExtTag+ PhantFunc- AuxPwr- NoSnoop+
                        MaxPayload 256 bytes, MaxReadReq 512 bytes
                DevSta: CorrErr- NonFatalErr- FatalErr- UnsupReq- AuxPwr- TransPend-
                LnkCap: Port #0, Speed 16GT/s, Width x16, ASPM L0s L1, Exit Latency L0s <4us, L1 <64us
                        ClockPM- Surprise- LLActRep- BwNot- ASPMOptComp+
                LnkCtl: ASPM Disabled; RCB 64 bytes, Disabled- CommClk-
                        ExtSynch- ClockPM- AutWidDis- BWInt- AutBWInt-
                LnkSta: Speed 16GT/s (ok), Width x16 (ok)
                        TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
                DevCap2: Completion Timeout: Not Supported, TimeoutDis+ NROPrPrP- LTR-
                         10BitTagComp+ 10BitTagReq- OBFF Not Supported, ExtFmt- EETLPPrefix-
                         EmergencyPowerReduction Not Supported, EmergencyPowerReductionInit-
                         FRS- TPHComp- ExtTPHComp-
                         AtomicOpsCap: 32bit+ 64bit- 128bitCAS-
```

2.3 驱动+软件栈安装--docker方式(推荐)

2.3.1 安装docker(系统已安装docker，可跳过该步骤)

```
# 更新软件源列表
sudo apt-get update
# 安装软件依赖包
sudo apt-get install apt-transport-https ca-certificates curl software-properties-common
# 在系统中添加 Docker 的官方密钥
curl -fsSL http://mirrors.aliyun.com/docker-ce/linux/ubuntu/gpg | sudo apt-key add -
# 安装 Docker 存储库，选择 stable 长期稳定版:
sudo add-apt-repository "deb [arch=amd64] http://mirrors.aliyun.com/docker-ce/linux/ubuntu $(lsb_release -cs) stable"
安装最新版本 Docker
sudo apt install docker-ce --fix-missing
# 查看安装的 Docker 版本
docker -v
# 启动 Docker 服务、
sudo systemctl start docker
# 查看 Docker 开启状态，显示绿点表示服务正常启动
sudo systemctl status docker
```

```
root@test-hpc-05:/home/lsc/BI150# docker -v
Docker version 24.0.7, build 24.0.7-0ubuntu2~22.04.1
```

```
root@test-hpc-05:/home/lsc/BI150# sudo systemctl status docker
● docker.service - Docker Application Container Engine
     Loaded: loaded (/lib/systemd/system/docker.service; enabled; vendor preset: enabled)
     Active: active (running) since Thu 2024-08-29 02:51:42 UTC; 1h 40min ago
TriggeredBy: ● docker.socket
       Docs: https://docs.docker.com
   Main PID: 2129 (dockerd)
      Tasks: 75
     Memory: 147.1M
        CPU: 3min 34.747s
     CGroup: /system.slice/docker.service
             └─2129 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/containerd.sock

Aug 29 02:53:56 test-hpc-05 dockerd[2129]: time="2024-08-29T02:53:56.867596723Z" level=info msg="ignoring event" container=ab253ee7b
Aug 29 02:55:08 test-hpc-05 dockerd[2129]: time="2024-08-29T02:55:08.295346542Z" level=info msg="ignoring event" container=bf62a889e
Aug 29 02:55:12 test-hpc-05 dockerd[2129]: time="2024-08-29T02:55:12.207905376Z" level=info msg="ignoring event" container=772bfaefd
Aug 29 02:57:01 test-hpc-05 dockerd[2129]: time="2024-08-29T02:57:01.611769154Z" level=info msg="ignoring event" container=d89aa7081
Aug 29 02:57:51 test-hpc-05 dockerd[2129]: time="2024-08-29T02:57:51.471183674Z" level=info msg="ignoring event" container=fe66c7249
Aug 29 03:00:09 test-hpc-05 dockerd[2129]: time="2024-08-29T03:00:09.804513244Z" level=info msg="ignoring event" container=20e4ae6f9
Aug 29 03:03:00 test-hpc-05 dockerd[2129]: time="2024-08-29T03:03:00.344099918Z" level=info msg="ignoring event" container=daf71b8f3
Aug 29 03:05:38 test-hpc-05 dockerd[2129]: time="2024-08-29T03:05:38.011998774Z" level=info msg="ignoring event" container=4e5f29242
Aug 29 04:14:35 test-hpc-05 dockerd[2129]: time="2024-08-29T04:14:35.192042646Z" level=info msg="ignoring event" container=21afa3b64
Aug 29 04:14:59 test-hpc-05 dockerd[2129]: time="2024-08-29T04:14:59.259408376Z" level=info msg="ignoring event" container=f4ab71c6c
lines 1-22/22 (END)
```

2.3.2 安装cuda头文件

```
1 bash install-cuda-header.sh
```

```
root@test-hpc-05:/home/lsc/BI150# bash install-cuda-header.sh
列出新安装的cuda头文件
lrwxrwxrwx  1 root root   20 Aug 29 04:30 cuda -> /usr/local/cuda-10.2
drwxrwxrwx  5 root root 4.0K Aug 29 03:10 cuda-10.2
```

2.3.3 生成 Docker 镜像

```
1 sudo bash corex-docker-installer-4.1.0-10.2-centos7.8.2003-py3.10-x86_64.run
```

勾选 Install driver,禁用dkms，如下Install,如下图所示：

```
Corex v4.1.0 Docker Install
  [*] Install driver
  [ ] Disable dkms        Disable use of dkms
      Module Secret Key   Private Key File to sign driver module
      Module Public Key   Public Key File to sign driver module
      Set Cuda Path
      Set Image Name
      Install
```

```
root@test-hpc-05:/home/lsc/BI150/sdk# sudo bash corex-docker-installer-4.1.0-10.2-centos7.8.2003-py3.10-x86_64.run
Verifying archive integrity...  100%   All good.
Uncompressing Corex Centos Docker Installer  100%

Start to install the Corex Driver.

Start to build image corex:4.1.0

It may take some minutes to build image, please wait...
Error response from daemon: No such image: installer:4.1.0-centos-7.8.2003-py3.10-x86_64-10.2

Driver:        Installed
For the Corex Driver uninstallation, please run command:
    sudo /usr/local/corex-4.1.0/bin/corex-driver-uninstaller
Logfile is /var/log/iluvatarcorex/driver_installer.log


Docker image corex:4.1.0 is ready, load corex container as following example:
    docker run -it -v /usr/src:/usr/src -v /lib/modules:/lib/modules -v /dev:/dev \
        --privileged --cap-add=ALL --pid=host corex:4.1.0
Logfile is /var/log/iluvatarcorex/docker_installer.log
```

2.3.4 启动容器

```bash
# 创建容器
sudo docker run -it --name BI150_test --network=host \
-v /usr/src:/usr/src \
-v /lib/modules:/lib/modules -v /dev:/dev \
-v /home/r740/lsc/:/home/r740/lsc/ \
--shm-size="32g" \
--privileged --cap-add=ALL --pid=host corex:4.1.0 /bin/bash

# 退出容器
exit

# 重新启动容器
docker start BI150_test

# 使用 exec 命令进入一个正在运行的容器
docker exec -it BI150_test /bin/bash
```

2.4 驱动+软件栈安装--宿主机方式

1. 安装Driver+Toolkit

备注：Driver和Toolkit可选安装，按下Enter键选择安装对应选项。

```bash
sudo bash corex-installer-linux64-4.1.0_x86_64_10.2.run
```


```
root@test-hpc-05:/home/lsc/BI150/sdk# sudo bash corex-installer-linux64-4.1.0_x86_64_10.2.run
Verifying archive integrity...  100%   All good.
Uncompressing Corex Installer  100%

Start to install the Corex Toolkit at /usr/local/corex-4.1.0/...

===========
= Summary =
===========

Driver:       Not Selected
Toolkit:      Installed at location '/usr/local/corex-4.1.0/'

Please make sure that
- PATH includes /usr/local/corex-4.1.0/bin
- LD_LIBRARY_PATH includes /usr/local/corex-4.1.0/lib

For the Corex Toolkit uninstallation, please run command:
    sudo /usr/local/corex-4.1.0/bin/corex-uninstaller
Logfile is /var/log/iluvatarcorex/corex_installer.log
```

2. 设置环境变量

宿主机上安装软件栈后，您需要修改 PATH 和 LD_LIBRARY_PATH 环境变量才能正常使用软件栈。以软件栈默认安

装路径 /usr/local/corex-{v.r.m}/ 为例，您需要：

- 为 PATH 环境变量加上 /usr/local/corex-4.1.0/bin
- 为 LD_LIBRARY_PATH 环境变量加上 /usr/local/corex-4.1.0/lib

2.5 安装深度学习框架和推理框架--宿主机方式

1. **创建虚拟环境**

```
python3 -m venv bi150_venv
```

这里 bi150_env 是你虚拟环境的名称，你可以根据需要更改。

2. **激活虚拟环境**

```
source bi150_venv/bin/activate
```

激活后，你会看到命令提示符前面有 `(bi150_venv)`，表示你已进入虚拟环境。

3. **安装包**

在虚拟环境中，你可以使用 `pip` 安装所需的包。例如：

```
pip3 install <one_whl_file>
```

4. **退出虚拟环境**

当你完成工作后，可以使用以下命令退出虚拟环境：

```
deactivate
```

5. 当前提供的天数智芯适配版深度学习框架和推理框架主要包含：

- TensorFlow v2.12.0
- PyTorch v2.1.1
- torchaudio 领域库 v2.1.0
- torchvision 领域库 v0.16.0
- PaddlePaddle v2.5.2
- Horovod v0.27.0
- ONNXRuntime_gpu v1.13.1
- DeepSpeed 大模型训练框架 v0.14.3
- Megatron-DeepSpeed 大模型训练框架 v2.4.1
- Megatron-LM 大模型训练框架 v0.6.0
- Triton 训练框架 v2.1.0
- IxFormer 大模型推理框架 v0.4.0
- IGIE 推理框架 v0.9.1
- IxRT 推理引擎 v0.9.1
- vLLM 推理框架 v0.3.3
- Apex 加速库 v0.1 (支持 PyTorch)
- DALI 加速库 v1.21.0 (支持 PyTorch)
- cluster 加速库 v1.6.0 (支持 PyTorch)

- quiver 加速库 v0.1.0 (支持 PyTorch)
- scatter 加速库 v2.1.0 (支持 PyTorch)
- sparse 加速库 v0.6.16 (支持 PyTorch)
- FlashAttention 加速库 v2.5.8 (支持 PyTorch)
- TorchDebug 精度调试工具 v0.1.0
- ixTE 大模型训练加速库 v0.2.0

2.6 资源监控

安装驱动+工具之后，在Host 环境下可以查看GPU 信息。

```
1 ixsmi
```

```
root@test-hpc-05:/home/lsc/BI150/sdk# ixsmi
Timestamp    Thu Aug 29 17:09:05 2024
+-----------------------------------------------------------------------------+
|  IX-ML: 4.1.0       Driver Version: 4.1.0       CUDA Version: 10.2          |
|-------------------------------+----------------------+----------------------|
| GPU  Name                     | Bus-Id               | Clock-SM  Clock-Mem  |
| Fan  Temp  Perf  Pwr:Usage/Cap|      Memory-Usage    | GPU-Util  Compute M. |
|===============================+======================+======================|
| 0    Iluvatar BI-V150         | 00000000:3F:00.0     | 1500MHz   1600MHz    |
| 0%   46C   P0    N/A / N/A    | 114MiB / 32768MiB    | 0%        Default    |
+-------------------------------+----------------------+----------------------+
| 1    Iluvatar BI-V150         | 00000000:42:00.0     | 1500MHz   1600MHz    |
| 0%   44C   P0    101W / 350W  | 114MiB / 32768MiB    | 0%        Default    |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU        PID      Process name                                Usage(MiB) |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

```
1 ixsmi dmon
```

```
root@test-hpc-05:/home/lsc/BI150/sdk# ixsmi dmon
# gpu    pwr  gtemp  mtemp     sm    mem    enc    dec   mclk   pclk
# Idx      W      C      C      %      %      %      %    MHz    MHz
    0      -     46      -      0      1      0      0   1600   1500
    1    101     44      -      0      1      0      0   1600   1500
    0      -     46      -      0      1      0      0   1600   1500
    1    101     44      -      0      1      0      0   1600   1500
    0      -     46      -      0      1      0      0   1600   1500
    1    101     44      -      0      1      0      0   1600   1500
    0      -     46      -      0      1      0      0   1600   1500
    1    101     44      -      0      1      0      0   1600   1500
    0      -     46      -      0      1      0      0   1600   1500
    1    101     44      -      0      1      0      0   1600   1500
```

3. Qwen2-7B-Instruct推理用例测试

3.1 测试需求：

1. 输入参数：
   a. batch size = 1, 2, 4, 8
   b. Input seq length = 8K
   c. Output seq length = 256
   d. FP16/INT4
2. 指标:
   a. TTFT
   b. TPOT
   c. TPS

3.2 基于vLLM在线推理功能测试

1. 进入容器，并启动一个基于 OpenAI 的 API 服务器。

```
# 安装了 outlines 模块（第一次运行需要安装）
pip3 install --upgrade outlines

# FP16
python3 -m vllm.entrypoints.openai.api_server \
        --model /home/r740/lsc/models/Qwen2-7B-Instruct/ \
        --device 'auto' \
        --host 0.0.0.0 \
        --trust-remote-code \
        --port 8080 \
```

```
        --enforce-eager

# 或者 GPTQ INT4
python3 -m vllm.entrypoints.openai.api_server \
        --model /home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/ \
        --device 'auto' \
        --host 0.0.0.0 \
        --trust-remote-code \
        --port 8080 \
        --enforce-eager
```

2. 打开另外一个终端，发送客户端请求。

注：下面IP需要替换为对应服务器IP

```
# FP16
curl http://10.110.165.160:8080/v1/completions \
-H "Content-Type: application/json" \
-d '{
"model": "/home/r740/lsc/models/Qwen2-7B-Instruct/",
"prompt": "如何制作月饼",
"max_tokens": 256,
"temperature": 0.01
}'

# 或者 GPTQ INT4
curl http://10.110.165.160:8080/v1/completions \
-H "Content-Type: application/json" \
-d '{
"model": "/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/",
"prompt": "如何制作月饼",
"max_tokens": 256,
"temperature": 0.01
}'
```

FP16

```
r740@r740:~/lsc$ curl http://10.110.165.160:8080/v1/completions \
-H "Content-Type: application/json" \
-d '{
"model": "/home/r740/lsc/models/Qwen2-7B-Instruct/",
"prompt": "如何制作月饼",
"max_tokens": 256,
"temperature": 0.01
}'
{"id":"cmpl-f7debe3ff1e54f1182f3606df5ad6c07","object":"text_completion","created":7623,"model":"/home/r740/lsc/models/Qwen2-7B-Instruct/","choices":[{"index":0,"text":"皮和馅料？\n制作月饼皮和馅料需要以下材料和步骤：\n月饼皮：\n材料：\n- 中筋面粉 200克\n- 糖浆 100克\n- 猪油 100克\n- 水 50克\n步骤：\n1. 将中筋面粉过筛，放入大碗中。\n2. 将糖浆、猪油和水混合均匀，倒入面粉中。\n3. 用手揉成面团，揉至表面光滑，盖上湿布静置30分钟。\n4. 将面团分成小份，每个约30克，搓成圆形。\n5. 将面团压扁，包入馅料，收口捏紧，搓成圆形。\n6. 将月饼放入烤盘中，用叉子在表面轻轻划几下，防止膨胀。\n7. 将烤盘放入预热至180度的烤箱中，烤约15-20分钟，至表面金黄即可。\n月饼馅料：\n材料：\n- 熟糯米粉 100克\n- 熟花生粉 100克\n- 熟","logprobs":null,"finish_reason":"length"}],"usage":{"prompt_tokens":3,"total_tokens":259,"compl
```

GPTQ INT4

```
code":404}[root@r740 inference_scripts]# curl http://10.110.165.160:8080/v1/completions \
> -H "Content-Type: application/json" \
> -d '{
> "model": "/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/",
> "prompt": "如何制作月饼",
> "max_tokens": 256,
> "temperature": 0.01
> }'
{"id":"cmpl-3a493836cf314a71a16d07fab2788f80","object":"text_completion","created":255126,"model":"/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/","choices":[{"index":0,"text":"？请提供详细的步骤和所需材料。\n制作月饼需要以下材料：\n- 月饼皮：糯米粉 150克、糖粉 50克、猪油 50克、水 60克\n- 月饼馅：豆沙馅 200克、莲蓉馅 200克、五仁馅 200克（任选一种或多种）\n制作步骤如下：\n1. 将糯米粉、糖粉、猪油和水混合，揉成面团，放置15分钟。\n2. 将豆沙馅、莲蓉馅或五仁馅分成20克一个的小球，放置备用。\n3. 将面团分成20克一个的小球，压扁，包入馅料，搓圆。\n4. 将月饼放入模具中，轻轻压紧，然后倒出。\n5. 将月饼放入预热至180度的烤箱中，烤15-20分钟即可。\n注意事项：\n1. 糖粉和猪油的比例可以根据个人口味调整。\n2. 面团和馅料的比例可以根据个人口味调整。\n3. 烤箱的温度和时间可以根据实际情况调整","logprobs":null,"finish_reason":
```

3.3 基于vLLM在线推理性能测试

1. 用以下文件替换/home/r740/lsc/vllm/benchmarks/benchmark_serving.py文件

benchmark_serving.py

2. 继续沿用上面容器,启动服务端。

```
# 启动server端服务
cd /home/r740/lsc/inference_scripts/
./run_openai_api_server_xn.sh
```

附录：run_openai_api_server_xn.sh

run_openai_api_server_xn.sh

3. 打开另一个终端，并进入同个容器，启动客户端发送请求。

```
# 重新启动容器
docker start docker_ruijie_test

```

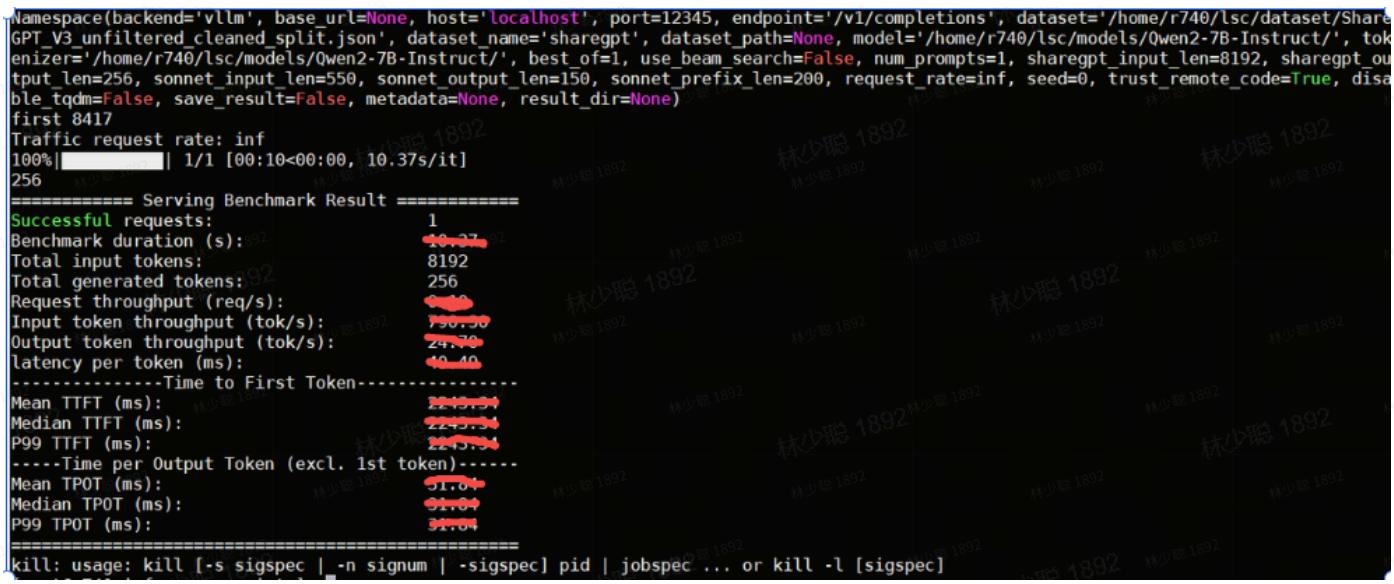
```
# 使用 exec 命令进入一个正在运行的容器
docker exec -it docker_ruijie_test /bin/bash

# 执行client请求
cd home/workspace/inference_scripts
./run_openai_api_client_xn.sh
```

附录：run_openai_api_client_xn.sh

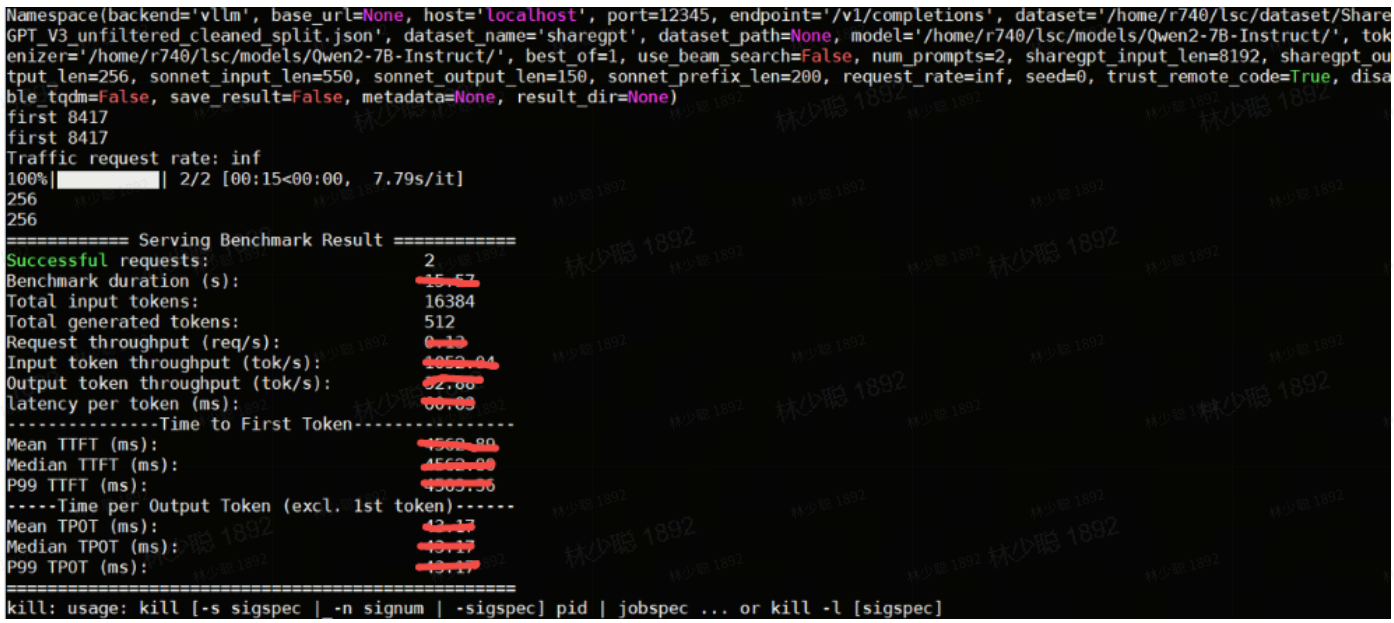
run_openai_api_client_xn.sh

4. 测试结果截图及记录：

a. batch size = 1；input_len = 8192；output_len = 256; FP16



b. batch size = 2；input_len = 8192；output_len = 256; FP16



c. batch size = 4； input_len = 8192； output_len = 256; FP16

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/home/r740/lsc/dataset/Share
GPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/home/r740/lsc/models/Qwen2-7B-Instruct/', tok
enizer='/home/r740/lsc/models/Qwen2-7B-Instruct/', best_of=1, use_beam_search=False, num_prompts=4, sharegpt_input_len=8192, sharegpt_ou
tput_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_remote_code=True, disa
ble_tqdm=False, save_result=False, metadata=None, result_dir=None)
first 8417
first 8417
Traffic request rate: inf
100%|          | 4/4 [00:22<00:00,  5.53s/it]
256
256
256
256
============ Serving Benchmark Result ============
Successful requests:                     4
Benchmark duration (s):                  22.12
Total input tokens:                      32768
Total generated tokens:                  1024
Request throughput (req/s):              [redacted]
Input token throughput (tok/s):          [redacted]
Output token throughput (tok/s):         [redacted]
latency per token (ms):                  [redacted]
---------------Time to First Token----------------
Mean TTFT (ms):                          [redacted]
Median TTFT (ms):                        [redacted]
P99 TTFT (ms):                           [redacted]
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          [redacted]
Median TPOT (ms):                        [redacted]
P99 TPOT (ms):                           [redacted]
==================================================
kill: usage: kill [-s sigspec | -n signum | -sigspec] pid | jobspec ... or kill -l [sigspec]
```

d. batch size = 8； input_len = 8192； output_len = 256; FP16

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/home/r740/lsc/dataset/Share
GPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/home/r740/lsc/models/Qwen2-7B-Instruct/', tok
enizer='/home/r740/lsc/models/Qwen2-7B-Instruct/', best_of=1, use_beam_search=False, num_prompts=8, sharegpt_input_len=8192, sharegpt_ou
tput_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_remote_code=True, disa
ble_tqdm=False, save_result=False, metadata=None, result_dir=None)
first 8417
first 8417
Traffic request rate: inf
100%|          | 8/8 [00:35<00:00,  4.44s/it]
256
256
256
256
256
256
256
256
============ Serving Benchmark Result ============
Successful requests:                     8
Benchmark duration (s):                  35.50
Total input tokens:                      65536
Total generated tokens:                  2048
Request throughput (req/s):              [redacted]
Input token throughput (tok/s):          [redacted]
Output token throughput (tok/s):         [redacted]
latency per token (ms):                  [redacted]
---------------Time to First Token----------------
Mean TTFT (ms):                          [redacted]
Median TTFT (ms):                        [redacted]
P99 TTFT (ms):                           [redacted]
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          [redacted]
Median TPOT (ms):                        [redacted]
P99 TPOT (ms):                           [redacted]
==================================================
kill: usage: kill [-s sigspec | -n signum | -sigspec] pid | jobspec ... or kill -l [sigspec]
```

e. batch size = 1； input_len = 8192； output_len = 256; INT4

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/home/r740/lsc/dataset/Shar
GPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_I
nt4/', tokenizer='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/', best_of=1, use_beam_search=False, num_prompts=1, sharegpt_input_l
en=8192, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_re
mote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
first 8417
Traffic request rate: inf
100%|          | 1/1 [00:02<00:00,  2.69s/it]
16
============ Serving Benchmark Result ============
Successful requests:                     1
Benchmark duration (s):                  2.69
Total input tokens:                      8192
Total generated tokens:                  16
Request throughput (req/s):              [redacted]
Input token throughput (tok/s):          [redacted]
Output token throughput (tok/s):         [redacted]
latency per token (ms):                  [redacted]
---------------Time to First Token----------------
Mean TTFT (ms):                          [redacted]
Median TTFT (ms):                        [redacted]
P99 TTFT (ms):                           [redacted]
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          [redacted]
Median TPOT (ms):                        [redacted]
P99 TPOT (ms):                           [redacted]
==================================================
kill: usage: kill [-s sigspec | -n signum | -sigspec] pid | jobspec ... or kill -l [sigspec]
```

f. batch size = 2； input_len = 8192； output_len = 256; INT4

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/home/r740/lsc/dataset/Shar
GPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_I
nt4/', tokenizer='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/', best_of=1, use_beam_search=False, num_prompts=2, sharegpt_input_l
en=8192, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_re
mote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
first 8417
first 8417
Traffic request rate: inf
100%|          | 2/2 [00:05<00:00,  2.61s/it]
16
16
============ Serving Benchmark Result ============
Successful requests:                     2
Benchmark duration (s):                  [redacted]
Total input tokens:                      16384
Total generated tokens:                  32
Request throughput (req/s):              [redacted]
Input token throughput (tok/s):          [redacted]
Output token throughput (tok/s):         [redacted]
latency per token (ms):                  [redacted]
---------------Time to First Token----------------
Mean TTFT (ms):                          [redacted]
Median TTFT (ms):                        [redacted]
P99 TTFT (ms):                           [redacted]
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          [redacted]
Median TPOT (ms):                        [redacted]
P99 TPOT (ms):                           [redacted]
==================================================
kill: usage: kill [-s sigspec | -n signum | -sigspec] pid | jobspec ... or kill -l [sigspec]
```

g. batch size = 4； input_len = 8192； output_len = 256; INT4

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/home/r740/lsc/dataset/Share
GPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_I
nt4/', tokenizer='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/', best_of=1, use_beam_search=False, num_prompts=4, sharegpt_input_l
en=8192, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_re
mote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
first 8417
first 8417
Traffic request rate: inf
100%|██████████| 4/4 [00:16<00:00,  4.07s/it]
16
16
256
256
============ Serving Benchmark Result ============
Successful requests:                     4
Benchmark duration (s):                  [illegible]
Total input tokens:                      32768
Total generated tokens:                  544
Request throughput (req/s):              [illegible]
Input token throughput (tok/s):          [illegible]
Output token throughput (tok/s):         [illegible]
latency per token (ms):                  [illegible]
---------------Time to First Token----------------
Mean TTFT (ms):                          [illegible]
Median TTFT (ms):                        [illegible]
P99 TTFT (ms):                           [illegible]
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          [illegible]
Median TPOT (ms):                        [illegible]
P99 TPOT (ms):                           [illegible]
==================================================
kill: usage: kill [-s sigspec | -n signum | -sigspec] pid | jobspec ... or kill -l [sigspec]
```

h. batch size = 8； input_len = 8192； output_len = 256; INT4

```
Namespace(backend='vllm', base_url=None, host='localhost', port=12345, endpoint='/v1/completions', dataset='/home/r740/lsc/dataset/Share
GPT_V3_unfiltered_cleaned_split.json', dataset_name='sharegpt', dataset_path=None, model='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_I
nt4/', tokenizer='/home/r740/lsc/models/Qwen2-7B-Instruct_GPTQ_Int4/', best_of=1, use_beam_search=False, num_prompts=8, sharegpt_input_l
en=8192, sharegpt_output_len=256, sonnet_input_len=550, sonnet_output_len=150, sonnet_prefix_len=200, request_rate=inf, seed=0, trust_re
mote_code=True, disable_tqdm=False, save_result=False, metadata=None, result_dir=None)
first 8417
first 8417
Traffic request rate: inf
100%|██████████| 8/8 [00:32<00:00,  4.03s/it]
16
16
256
256
256
256
256
256
============ Serving Benchmark Result ============
Successful requests:                     8
Benchmark duration (s):                  [illegible]
Total input tokens:                      65536
Total generated tokens:                  1568
Request throughput (req/s):              [illegible]
Input token throughput (tok/s):          [illegible]
Output token throughput (tok/s):         [illegible]
latency per token (ms):                  [illegible]
---------------Time to First Token----------------
Mean TTFT (ms):                          [illegible]
Median TTFT (ms):                        [illegible]
P99 TTFT (ms):                           [illegible]
-----Time per Output Token (excl. 1st token)------
Mean TPOT (ms):                          [illegible]
Median TPOT (ms):                        [illegible]
P99 TPOT (ms):                           [illegible]
==================================================
kill: usage: kill [-s sigspec | -n signum | -sigspec] pid | jobspec ... or kill -l [sigspec]
```

3.4 基于VLLM性能测试数据汇总

	Output tokens_per_second (TPS)(ms)	latency_per_token（TPOT）(ms)	first_token (TTFT)(ms)	Benchmark duration(s)
1. batch size = 1 2. input_len = 8192				

3. output_len = 256 4. FP16				
1. batch size = 2 2. input_len = 8192 3. output_len = 256 4. FP16				
1. batch size = 4 2. input_len = 8192 3. output_len = 256 4. FP16				
1. batch size = 8 2. input_len = 8192 3. output_len = 256 4. FP16				
1. batch size = 1 2. input_len = 8192 3. output_len = 256 4. INT4				
1. batch size = 2 2. input_len = 8192 3. output_len = 256 4. INT4				
1. batch size = 4 2. input_len = 8192				

3. output_len = 256 4. INT4				
1. batch size = 8 2. input_len = 8192 3. output_len = 256 4. INT4				

4. Qwen2-7B-Instruct 模型ceval gsm8k mmlu bbh基准评测

4.1 Qwen2-7B-Instruct原始模型ceval gsm8k mmlu bbh评测

1. 评测命令及结果

```
CUDA_VISIBLE_DEVCIES=0 swift eval \
    --model_type qwen2-7b-instruct \
    --model_id_or_path /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct \
    --eval_dataset ceval gsm8k mmlu bbh \
    --infer_backend vllm \
    --max_model_len=16000
```

```
[INFO:swift] result: {'ceval': 0.6657, 'gsm8k': 0.7915, 'bbh': 0.2439, 'mmlu': 0.6846}
[INFO:swift] save_result_path: /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct/eval_result.jsonl
[INFO:swift] End time of running main: 2024-06-27 23:29:35.336406
```

4.2 Qwen2-7B-Instruct模型gptq int4量化ceval gsm8k mmlu bbh评测

1. 评测命令及结果

```
CUDA_VISIBLE_DEVCIES=0 swift eval \
    --model_type qwen2-7b-instruct \
    --model_id_or_path /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct_GPTQ_Int4 \
    --eval_dataset ceval gsm8k mmlu bbh \
    --infer_backend vllm \
    --max_model_len=16000
```

```
[INFO:swift] result: {'ceval': 0.6738}
[INFO:swift] save_result_path: /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct_GPTQ_Int4/eval_result.jsonl
[INFO:swift] End time of running main: 2024-06-27 16:54:18.367103
```

```
[INFO:swift] result: {'gsm8k': 0.7763, 'mmlu': 0.6631}
[INFO:swift] save_result_path: /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct_GPTQ_Int4/eval_result.jsonl
[INFO:swift] End time of running main: 2024-06-27 18:57:24.105001
```

```
[INFO:swift] result: {'bbh': 0.2338}
[INFO:swift] save_result_path: /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct_GPTQ_Int4/eval_result.jsonl
[INFO:swift] End time of running main: 2024-06-28 11:06:56.101122
```

4.2.1 Qwen2-7B-Instruct模型awq int4量化ceval gsm8k mmlu bbh评测

```
CUDA_VISIBLE_DEVCIES=0 swift eval \
    --model_type qwen2-7b-instruct \
    --model_id_or_path /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct_AWQ_Int4 \
    --eval_dataset ceval gsm8k mmlu bbh \
    --infer_backend vllm \
    --max_model_len=16000
```

```
INFO:llmuses:** Report table:
 +---------+---------------------+--------------------+--------------------+
| Model   | ceval               | gsm8k              | mmlu               |
+=========+=====================+====================+====================+
|         | (ceval/acc) 0.6501 | (gsm8k/acc) 0.768 | (mmlu/acc) 0.6529 |
+---------+---------------------+--------------------+--------------------+
```

```
[INFO:swift] result: {'bbh': 0.2367}
[INFO:swift] save_result_path: /home/aigc/lsc/vllm_test/Qwen2-7B-Instruct_AWQ_Int4/eval_result.jsonl
[INFO:swift] End time of running main: 2024-06-28 19:12:09.927829
```

5. bge-large-zh-v1.5模型推理部署

5.1 FlagEmbedding方式

1. 环境准备--安装FlagEmbedding

```
pip3 install -U FlagEmbedding
```

2. 推理脚本--bge-large-zh-1.5_flagembedding.py

```
from FlagEmbedding import FlagModel
sentences_1 = ["样例数据-1", "样例数据-2"]
```

```
sentences_2 = ["样例数据-3", "样例数据-4"]
model = FlagModel('/home/r740/lsc/models/bge-large-zh-v1.5',
                  query_instruction_for_retrieval="为这个句子生成表示以用于检索相关文章：",
                  use_fp16=True) # Setting use_fp16 to True speeds up computation with a slight performance degradation
embeddings_1 = model.encode(sentences_1)
embeddings_2 = model.encode(sentences_2)
similarity = embeddings_1 @ embeddings_2.T
print(similarity)

# for s2p(short query to long passage) retrieval task, suggest to use encode_queries() which will automatically add the instruction to each query
# corpus in retrieval task can still use encode() or encode_corpus(), since they don't need instruction
queries = ['query_1', 'query_2']
passages = ["样例文档-1", "样例文档-2"]
q_embeddings = model.encode_queries(queries)
p_embeddings = model.encode(passages)
scores = q_embeddings @ p_embeddings.T
```

3. 执行推理--bge-large-zh-1.5_flagembedding.py

```
python3 bge-large-zh-1.5_flagembedding.py
```

```
[root@r740 inference_scripts]# python3 bge-large-zh-1.5_flagembedding.py
2024-09-02 08:10:21.080005: I tensorflow/core/util/port.cc:110] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-09-02 08:10:22.685908: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: SSE3 SSE4.1 SSE4.2 AVX AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
WARNING:tensorflow:Deprecation warnings have been disabled. Set TF_ENABLE_DEPRECATION_WARNINGS=1 to re-enable them.
----------using 2*GPUs----------
[[0.855  0.8516]
 [0.874  0.8555]]
```

5.2 sentence_transformers方式

1. 推理脚本--bge-large-zh-1.5_transformers.py

```
from sentence_transformers import SentenceTransformer
queries = ['query_1', 'query_2']
passages = ["样例文档-1", "样例文档-2"]
instruction = "为这个句子生成表示以用于检索相关文章："

```

```
model = SentenceTransformer('/home/r740/lsc/models/bge-large-zh-v1.5')
q_embeddings = model.encode([instruction+q for q in queries], normalize_embeddings=True)
p_embeddings = model.encode(passages, normalize_embeddings=True)
scores = q_embeddings @ p_embeddings.T
print(scores)
```

2. 执行推理--bge-large-zh-1.5_transformers.py

```
python3 bge-large-zh-1.5_transformers.py
```

```
[root@r740 inference_scripts]# python3 bge-large-zh-1.5_transformers.py
2024-09-02 08:21:20.788641: I tensorflow/core/util/port.cc:110] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-09-02 08:21:20.843561: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: SSE3 SSE4.1 SSE4.2 AVX AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
WARNING:tensorflow:Deprecation warnings have been disabled. Set TF_ENABLE_DEPRECATION_WARNINGS=1 to re-enable them.
[[0.3372504  0.20507807]
 [0.22591081 0.38495794]]
```

6. bge-reranker-v2-m3模型推理

6.1 FlagEmbedding方式

1. 环境准备--安装FlagEmbedding

```
pip3 install -U FlagEmbedding
```

2. 推理脚本--bge-reranker-v2-m3_flagembedding.py

```
from FlagEmbedding import FlagReranker
reranker = FlagReranker('/home/r740/lsc/models/bge-reranker-v2-m3', use_fp16=True) # Setting use_fp16 to True speeds up computation with a slight performance degradation

score = reranker.compute_score(['query', 'passage'])
print(score) # -5.65234375

# You can map the scores into 0-1 by set "normalize=True", which will apply sigmoid function to the score
score = reranker.compute_score(['query', 'passage'], normalize=True)
```

```
print(score) # 0.003497010252573502

scores = reranker.compute_score([['what is panda?', 'hi'], ['what is panda?', 'The giant panda (Ailuropoda melanoleuca), sometimes called a panda bear or simply panda, is a bear species endemic to China.']])
print(scores) # [-8.1875, 5.26171875]

# You can map the scores into 0-1 by set "normalize=True", which will apply sigmoid function to the score
scores = reranker.compute_score([['what is panda?', 'hi'], ['what is panda?', 'The giant panda (Ailuropoda melanoleuca), sometimes called a panda bear or simply panda, is a bear species endemic to China.']], normalize=True)
print(scores) # [0.0002780309638775155, 0.9948403768236574]
```

3. 执行推理--bge-reranker-v2-m3_flagembedding.py

```
python3 bge-reranker-v2-m3_flagembedding.py
```

```
[root@r740 inference_scripts]# python3 bge-reranker-v2-m3_flagembedding.py
2024-09-02 08:26:33.624788: I tensorflow/core/util/port.cc:110] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off errors from different computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2024-09-02 08:26:33.679763: I tensorflow/core/platform/cpu_feature_guard.cc:182] This TensorFlow binary is optimized to use available CPU instructions in performance-critical operations.
To enable the following instructions: SSE3 SSE4.1 SSE4.2 AVX AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
WARNING:tensorflow:Deprecation warnings have been disabled. Set TF_ENABLE_DEPRECATION_WARNINGS=1 to re-enable them.
----------using 2*GPUs----------
[-5.65234375]
[0.003497010252573502]
[-8.171875, 5.26171875]
[0.0002824080799315707, 0.9948403768236574]
```

6.2 sentence_transformers方式

1. 推理脚本--bge-reranker-v2-m3_transformers.py

```
import torch
from transformers import AutoModelForSequenceClassification, AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained('/home/r740/lsc/models/bge-reranker-v2-m3')
model = AutoModelForSequenceClassification.from_pretrained('/home/r740/lsc/models/bge-reranker-v2-m3')
model.eval()

```

```
pairs = [['what is panda?', 'hi'], ['what is panda?', 'The giant panda (Ailuropoda melanoleuca), sometimes called a panda bear or simply panda, is a bear species endemic to China.']]
with torch.no_grad():
    inputs = tokenizer(pairs, padding=True, truncation=True, return_tensors='pt', max_length=512)
    scores = model(**inputs, return_dict=True).logits.view(-1, ).float()
    print(scores)
```

2. 执行推理--bge-reranker-v2-m3_transformers.py

```
python3 bge-reranker-v2-m3_transformers.py
```

```
[root@r740 inference_scripts]# vim bge-reranker-v2-m3_transformers.py
[root@r740 inference_scripts]# python3 bge-reranker-v2-m3_transformers.py
tensor([-8.1838,  5.2650])
```