1. 在做注意力头部测试时发现，

query\_states shape: [8, 16, 1, 24]

key\_states shape: [8, 16, 1, 24]

value\_states shape: [8, 16, 1, 48]

dtype为float16时，torch.nn.functional.scaled\_dot\_product\_attention会报错：

Head dimension of query/key must greater or equal to head dimension in query。

dtype 为float32时则不会报错。

1. 宿主机环境：

CPU：Intel(R) Xeon(R) Gold 5318Y \*2

内存：256GB（32GB\*8）

GPU：N260\*2

OS：Ubuntu 22.04.4 LTS

MACA：3.0.0.8

vBIOS：1.26.1.0

docker：27.5.1

1. 容器环境：

cr.metax-tech.com/public-ai-release/maca/vllm:maca.ai3.0.0.5-torch2.6-py310-ubuntu22.04-amd64

cr.metax-tech.com/public-ai-release/maca/sglang:maca.ai2.33.1.7-torch2.6-py310-ubuntu22.04-amd64

1. 简单测试内容介绍

后面在做测试发现在sglang的容器内会出现上述报错，而在vllm容器中不会出现上述报错。使用附件中的test1（上传限制需要修改为.py文件运行。）可稳定复现。

后续在代码中添加下面全局开关后：

torch.backends.cuda.enable\_flash\_sdp(False)

torch.backends.cuda.enable\_mem\_efficient\_sdp(False)

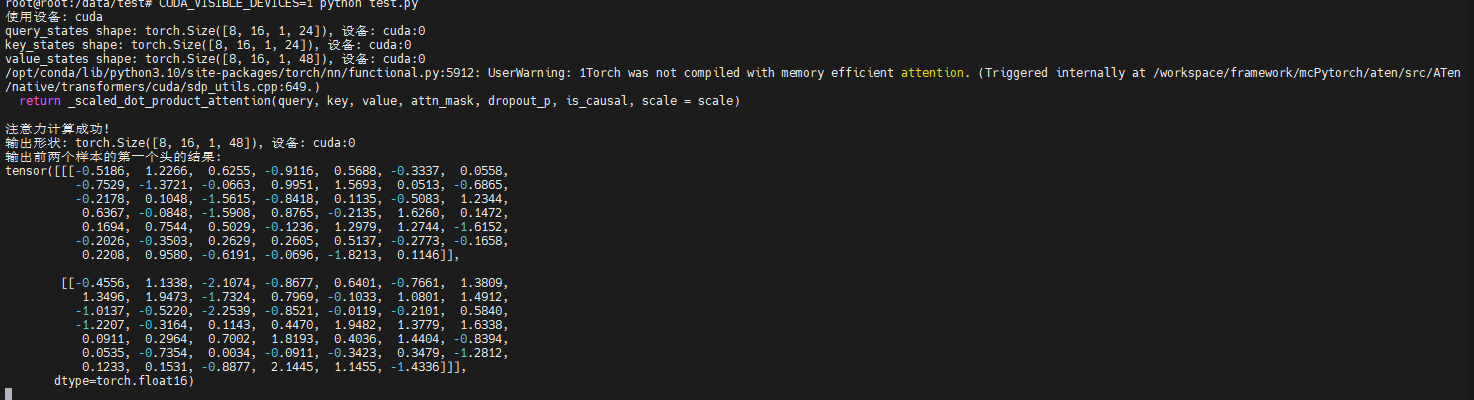
torch.backends.cuda.enable\_math\_sdp(True)

在sglang和vllm容器中就不会在出现上述报错。使用附件中的test2复现。

在后续测试中发现即使设置了上面的全局开关，在使用 flash\_attn\_func 时还是会出现上述错误。附件test3可稳定复现。

1. 回显：

成功运行的回显



报错的回显

