

PR #32246 完整报告

sgl-project/sglang

Fix nvfp4 online scale with pcg

合并时间: 2026-07-24 05:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32246>

执行摘要

- 一句话: 修复 NVFP4 在线量化时 scale 为 Python 标量的问题
- 推荐动作: 可以快速合并。但建议作者或团队后续补充针对 per_token_activation 路径的单元测试, 避免回归。

功能与动机

修复 NVFP4 在线量化时 scale 参数类型 / 设备不匹配导致的潜在错误。原始代码中 `1.0 / (e4m3_max * 6.0)` 返回 Python float, 而 `nvfp4_quantize` 期望一个 tensor, 可能引发隐式转换问题或设备不匹配。

实现拆解

1. 在 `flashinfer_trtllm.py` 的 `fused_experts_none_to_flashinfer_trtllm_fp4` 函数中, 将原先直接传入 Python float 的 `1.0 / (e4m3_max * 6.0)` 替换为通过 `torch.full` 显式创建的 float32 张量, 并指定与 `hidden_states` 相同的 device。
2. 该变更仅影响 `per_token_activation=True` 的 NVFP4 在线量化分支, 不改变其他量化路径。

关键文件:

- `python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py` (模块 MoE 计算; 类别 source; 类型 core-logic; 符号 `fused_experts_none_to_flashinfer_trtllm_fp4`): 修复 NVFP4 在线量化时 `global_scale_inv` 参数从 Python float 改为 torch tensor

关键符号: `fused_experts_none_to_flashinfer_trtllm_fp4`

关键源码片段

`python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py`

修复 NVFP4 在线量化时 `global_scale_inv` 参数从 Python float 改为 torch tensor

```
# 修改前: 传入 Python float, 可能导致 device 不匹配
hs_fp4_bytes, hs_sf_bytes, per_token_scale = nvfp4_quantize(
    hidden_states,
    1.0 / (e4m3_max * 6.0), # Python float, 无 device 信息
    sfLayout=SfLayout.layout_linear,
    per_token_activation=True,
    backend="cute-dsl",
```

)

修改后：创建显式 tensor，确保 device 与 hidden_states 一致

```
global_scale_inv = torch.full(
    (1,),
    1.0 / (e4m3_max * 6.0),
    dtype=torch.float32,
    device=hidden_states.device,
)
hs_fp4_bytes, hs_sf_bytes, per_token_scale = nvfp4_quantize(
    hidden_states,
    global_scale_inv,
    sfLayout=SfLayout.layout_linear,
    per_token_activation=True,
    backend="cute-dsl",
)
```

评论区精华

无实质性 review 讨论，变更获得 approve。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅将 Python 标量替换为同值的 tensor，且设备与 hidden_states 对齐，不会改变数值结果。但缺少针对此场景的专项单元测试。
- 影响：影响范围小，仅影响启用了 per_token_activation 的 NVFP4 在线量化路径。修复可能消除某些 CUDA kernel 因设备不匹配导致的崩溃。
- 风险标记：缺少专项测试

关联脉络

- 暂无明显关联 PR