PR #26025 完整报告

sgl-project/sglang

[fix] Fallback DeepGEMM activation for unsupported shapes

合并时间：2026-05-22 17:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26025

执行摘要

- 一句话：修复 DeepGEMM JIT activation 在非对齐 shape 下的崩溃
- 推荐动作：值得合并，修复明确，风险低。建议回归测试 fallback 路径下的性能和正确性。

功能与动机

修复用户使用 Qwen3-30B-A3B-FP8 等模型时因 N 或 G 不是 4 的倍数导致 DeepGEMM JIT EP activation 内核断言失败（assert N % 4 == 0 and G % 4 == 0）而崩溃的问题。PR body 提供了可复现的命令行。

实现拆解

1. 读取环境变量：在 python/sglang/srt/layers/moe/moe_runner/deep_gemm.py 的 _varlen_deep_gemm_silu_mul_quant 函数中，将 envs.SGLANG_OPT_USE_JIT_EP_ACTIVATION.get() 的返回值赋给局部变量 use_jit_ep_activation。
2. 检查 shape 约束：添加条件判断 if N % 4 != 0 or G % 4 != 0: use_jit_ep_activation = False，当 N 或 G 不是 4 的倍数时，禁用 JIT activation。
3. 条件分流：后续逻辑根据 use_jit_ep_activation 的值选择 JIT 路径或原有 fallback 路径。原有 fallback 路径中的 assert 保持不变。

关键文件：

- python/sglang/srt/layers/moe/moe_runner/deep_gemm.py（模块 MoE 运行器；类别 source；类型 core-logic）：核心变更文件，修复了 _varlen_deep_gemm_silu_mul_quant 函数中 JIT activation 在非对齐 shape 下崩溃的问题，添加 fallback 逻辑。

关键符号：未识别

关键源码片段

python/sglang/srt/layers/moe/moe_runner/deep_gemm.py

核心变更文件，修复了 _varlen_deep_gemm_silu_mul_quant 函数中 JIT activation 在非对齐 shape 下崩溃的问题，添加 fallback 逻辑。

```
# python/sglang/srt/layers/moe/moe_runner/deep_gemm.py
# 第 866-891 行（变更后）
```

```
# 当 DeepGEMM JIT EP activation 不支持当前 N/G 形状时（非 4 的倍数），
# 回退到原有的非 JIT 量化路径。

use_jit_ep_activation = envs.SGLANG_OPT_USE_JIT_EP_ACTIVATION.get()
# 如果 N 或 G 不是 4 的倍数，JIT 内核无法处理，禁用 JIT activation
if N % 4 != 0 or G % 4 != 0:
    use_jit_ep_activation = False

if use_jit_ep_activation:
    packed_ue8m0 = deep_gemm_wrapper.DEEPGEMM_SCALE_UE8M0
    down_input_scale = torch.empty(
        (E, G // 4, N) if packed_ue8m0 else (E, N, G),
        device=hidden_states_device,
        dtype=torch.int32 if packed_ue8m0 else torch.float32,
    )
    silu_and_mul_masked_post_quant(
        gateup_output,
        down_input,
        down_input_scale,
        group_size,
        masked_m,
        scale_ue8m0=packed_ue8m0,
        topk=topk,
        transposed=packed_ue8m0,
        swiglu_limit=swiglu_limit,
        swizzle=swizzle,
    )
    if packed_ue8m0:
        down_input_scale = down_input_scale.transpose(-1, -2)
else:
    # 原有 fallback 路径：使用 sglang_per_token_group_quant_8bit 等函数
    assert swiglu_limit is None, (
        "swiglu_limit (DeepSeek V4) requires SGLANG_OPT_USE_JIT_EP_ACTIVATION=True"
    )
    assert not swizzle, (
        "SGLANG_OPT_FIX_MEGA_MOE_MEMORY requires SGLANG_OPT_USE_JIT_EP_
        ACTIVATION=True"
    )
    down_input_scale = torch.empty(
        (E, N, G),
        device=hidden_states_device,
        dtype=torch.float32,
    )
    # ... 后续使用 sglang_per_token_group_quant_8bit 或类似调用
```

评论区精华

审核者 ch-wan 直接批准（LGTM），无其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。仅当 JIT activation 不支持的 shape 时才会触发 fallback，原有 fallback 路径已存在且经过测试。但建议确认 fallback 路径在非对齐 shape 下是否性能显著下降，以及检查 swiglu_limit 和 swizzle 参数的兼容性断言是否合理。
- 影响：影响范围小：仅影响使用 DeepGEMM JIT EP activation 且 N 或 G 不是 4 的倍数的模型（如部分 FP8 MoE 模型）。修复了启动崩溃，用户无需手动设置环境变量。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #25805 Fix SWA double-free in disagg decode with MTP speculation: 同为 DeepGEMM 相关 bugfix，涉及 MoE runner 的稳定性修复。