PR #25733 完整报告

sgl-project/sglang

[Bug] Fix V4-Pro NaN on Blackwell by converting fp8_einsum input scale to ue8m0

合并时间：2026-05-19 14:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25733

执行摘要

- 一句话：修复 Blackwell GPU 上 DeepSeek-V4-Pro NaN 问题
- 推荐动作：此 PR 为关键 bugfix，建议尽快合入。值得关注的是作者对问题的深入诊断（外部 gist 分析），体现了底层数值问题的调试方法。

功能与动机

DeepSeek-V4-Pro 在 Blackwell GPU 上自回归解码时输出乱码（garbage text）。作者在 PR body 和关联的详细分析 gist 中定位到根因：deep_gemm 的 transpose_and_pack_fp32_into_ue8m0 CUDA kernel 打包存在 bug—没有屏蔽 fp32 指数的尾数位，导致非 2 的幂 scale 值被损坏，进而使 fp8_einsum 在 batch=1（单 token 解码）时产生 NaN。

实现拆解

在 python/sglang/srt/models/deepseek_v4.py 的 forward 方法中，_FP8_WO_A_GEMM 分支内，sglang_per_token_group_quant_fp8 得到量化激活 o_fp8 和 scale o_s 后，在调用 deep_gemm.fp8_einsum 之前插入一行 o_s = deep_gemm.ceil_to_ue8m0(o_s)。该函数将 scale 值向上取整为最接近的 2 的幂，确保尾数位为零，从而绕过 deep_gemm 内部 packing kernel 的 bug。scale 张量形状很小（如 (2, 32)），因此计算开销可忽略。

关键文件：

- python/sglang/srt/models/deepseek_v4.py（模块 模型推理；类别 source；类型 data-contract）：核心修复文件，在模型 forward 的 FP8 量化分支中添加 scale 规范化调用，直接修复 NaN 问题。

关键符号：DeepseekV4Attention.forward

关键源码片段

python/sglang/srt/models/deepseek_v4.py

核心修复文件，在模型 forward 的 FP8 量化分支中添加 scale 规范化调用，直接修复 NaN 问题。

```
# python/sglang/srt/models/deepseek_v4.py ( 修改后 )
if _FP8_WO_A_GEMM:
```

```
import deep_gemm
T, G, D = o.shape
R = self.o_lora_rank
o_fp8, o_s = sglang_per_token_group_quant_fp8(
    o.reshape(T * G, D).contiguous(),
    group_size=128,
)
# 关键修复：将 scale 向上取整为 ue8m0 格式，
# 绕过 deep_gemm 内部 pack kernel 尾数位泄露 bug
o_s = deep_gemm.ceil_to_ue8m0(o_s)
output = torch.empty(T, G, R, device=o.device, dtype=torch.bfloat16)
deep_gemm.fp8_einsum(
    "bhr,hdr->bhd",
    (o_fp8.view(T, G, D), o_s.view(T, G, -1)),
    (self.wo_a.weight.view(G, R, D), self.wo_a.weight_scale_inv.data),
    output,
    recipe=(1, 1, 128),
)
o = output
```

评论区精华

Reviewer ch-wan 直接批准，无实质讨论。评论中只有 bot 自动化 rerun 测试的记录。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅添加一行对 scale 进行 ceil_to_ue8m0 的操作，该函数返回结果与 o_s 形状相同，不影响后续逻辑形状。如果 deep_gemm 未来修复了 packing kernel 的 bug，此 ceil_to_ue8m0 调用可安全移除或保留为非幂值提供额外保护。但需注意该修复依赖 deep_gemm 的 ceil_to_ue8m0 函数，若该函数被删除或重命名会导致报错。
- 影响：直接影响 DeepSeek-V4-Pro 在 Blackwell GPU（B300/B200）上的推理正确性，修复了 NaN 问题。对其他模型和 GPU 架构无影响，因为 _FP8_WO_A_GEMM 分支仅在特定配置下启用。测试验证通过。
- 风险标记：依赖底层库实现，修复兼容性

关联脉络

- 暂无明显关联 PR