PR #27191 完整报告

sgl-project/sglang

Fix DeepSeek V4 DP reduce scatter when use attention DP + MoE TP

合并时间：2026-06-07 09:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27191

执行摘要

- 一句话：修复 DeepSeek V4 DP 注意力 + TP MoE 下 reduce-scatter 问题
- 推荐动作：该 PR 值得精读，以理解 DeepSeek V4 独特的手写 _use_tp_moe_gather 路径以及 DP 注意力与 TP MoE 交互时的数据流问题。对于关注 DeepSeek V4 模型推理或大规模并行训练的工程师有参考价值。建议合并或已合并。

功能与动机

DeepSeek V4 具有手写的 _use_tp_moe_gather 路径。当使用 DP 注意力 + TP MoE 时，MLP 输出是 TP-partial 的。若 should_use_dp_reduce_scatterv() 为 true，直接对部分张量做 scatter 会导致本地隐藏状态在后续 HC post block 之前未经规约。通用通信器已经对该模式使用 reduce_scatterv，此 PR 让 DeepSeek V4 自定义路径与之对齐。触发条件：DeepSeek V4 运行 DP 注意力 + attention_dp_size > 1 + 无 MoE A2A 后端 + 专家并行大小等于注意力 DP 大小。

实现拆解

1. 导入新增：在 python/sglang/srt/models/deepseek_v4.py 中，从 sglang.srt.layers.moe 额外导入 should_use_dp_reduce_scatterv，并从 sglang.srt.layers.dp_attention 导入 get_dp_global_num_tokens。
2. 控制流改造：在 forward 方法的 _use_tp_moe_gather 分支中，将原先单一的 dp_scatter(hidden_states, global_hidden_states, forward_batch) 替换为条件判断：如果 should_use_dp_reduce_scatterv() 为真，则调用 get_tp_group().reduce_scatterv(global_hidden_states, output=hidden_states, sizes=get_dp_global_num_tokens())，否则沿用原有的 dp_scatter。

关键文件：

- python/sglang/srt/models/deepseek_v4.py（模块 DeepSeek V4 模型；类别 source；类型 data-contract）：所有变更均在此文件，包括导入新增和控制流修改。是 DeepSeek V4 模型的核心实现。

关键符号：未识别

关键源码片段

python/sglang/srt/models/deepseek_v4.py

所有变更均在此文件，包括导入新增和控制流修改。是 DeepSeek V4 模型的核心实现。

```
# python/sglang/srt/models/deepseek_v4.py (head)
# 导入新增（第 71 行附近）:
from sglang.srt.layers.moe import get_moe_a2a_backend, should_use_dp_reduce_scatterv
# （第 62 行附近）:
from sglang.srt.layers.dp_attention import (
    ...
    get_dp_global_num_tokens, # 新增导入
    ...
)

# forward 方法中 _use_tp_moe_gather 分支（第 1429-1441 行）:
elif _use_tp_moe_gather:
    # 准备本地和全局隐藏状态缓冲区
    hidden_states, global_hidden_states = (
        get_local_dp_buffer(get_tp_group()),
        hidden_states,
    )
    # 根据配置选择是使用 reduce_scatterv 还是传统的 dp_scatter
    if should_use_dp_reduce_scatterv():
        # 使用 reduce_scatterv 进行规约和散布，与通用通信器行为一致
        get_tp_group().reduce_scatterv(
            global_hidden_states,
            output=hidden_states,
            sizes=get_dp_global_num_tokens(), # 需要正确的 token 数量
        )
    else:
        # 原有路径：只做 scatter
        dp_scatter(hidden_states, global_hidden_states, forward_batch)
```

评论区精华

该 PR 没有 review 评论，只有一个审批（ch-wan APPROVED）。Issue 评论均为 CI rerun 指令，无技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更局限于 DeepSeek V4 模型的一条特定控制流分支（_use_tp_moe_gather），且该分支仅在特定配置组合（DP 注意力 + TP MoE + attention_dp_size > 1 等）下才会触发。若 should_use_dp_reduce_scatterv() 为 false，行为退化为原有逻辑，无回归。由于没有直接对应的单元测试，建议作者或团队补充相关测试覆盖。
- 影响：影响范围小，仅影响 DeepSeek V4 模型在特定并行配置下的正确性。受影响用户：使用 DeepSeek V4 且开启 DP 注意力（例如 --sglang-enable-dp-attention）同时 attention_dp_size、tp_size、ep_size 均大于 1 的场景。修复后确保隐藏状态正确规约，

避免下游 HC post block 得到错误输入。

- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- 暂无明显关联 PR