PR #20967 完整报告

sgl-project/sglang

【BugFix】fix the bug of minimax_m2.5 model that causes repeated outputs when using tp16

合并时间：2026-04-10 22:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20967

执行摘要

该 PR 修复了 MiniMax M2.5 模型在 TP=16 配置下因 KV 头复制导致的 RMSNorm 权重分片错误，解决了重复输出问题。通过重构 MiniMaxM2RMSNormTP 类，使其感知头复制并正确分片权重，同时修复了前向传播中的方差归约逻辑。这是一个针对特定模型和 TP 配置的关键 bugfix，影响范围有限但修复了高端硬件下的模型可用性问题。

功能与动机

MiniMax M2.5 模型使用 8 个 KV 头，当 TP=16 时，TP 大小超过 KV 头数，多个 TP rank 必须共享同一个 KV 头。原有的 MiniMaxM2RMSNormTP 计算权重分片大小时使用 hidden_size / attn_tp_size，对于 K norm 得到 8d / 16 = 0.5d——一个非整数大小，导致错误的权重分片和归一化，最终产生重复或乱码输出。根本原因是 MiniMaxM2RMSNormTP 未感知到头级结构，假设权重维度总能被 tp_size 整除，这在 tp_size > num_kv_heads 时失效。

实现拆解

所有修改集中在 python/sglang/srt/models/minimax_m2.py 文件的 MiniMaxM2RMSNormTP 类中：

修改点	关键代码逻辑	目的
初始化	新增 num_heads 参数，根据 QKVParallelLinear 模式计算 num_heads、num_head_replicas 和 head_dim	确保权重大小始终为整数，支持头复制场景
权重加载器	从 @staticmethod 改为实例方法，使用 attn_tp_rank // num_head_replicas 计算分片索引	使复制 rank 正确加载相同权重分片
防御性断言	在 __init__ 中添加整除性检查，在 weight_loader 中添加边界检查	提前捕获配置错误，防止静默失败
前向传播	修复方差归约逻辑，确保正确 all-reduce 和除法	解决 reviewer 指出的计算错误

评论区精华

reviewer JustinTong0323 指出了三个关键问题，并推动了 PR 的改进：

> 初始化整除性验证："Both branches use integer division (//) without asserting exactness. If tp_size % num_heads != 0 (or vice versa), the remainder is silently dropped, causing incorrect shard calculations downstream."
>
> 方差归约逻辑错误："All-reduce sums all 16 ranks' variances (from 8 different heads × 2 replicas). Dividing by 2 gives sum_of_8_head_variances, but the correct per-layer variance requires dividing by 8..."
>
> weight_loader 边界检查："If the divisibility invariant from __init__ is violated, this will silently load incorrect weights — the model runs but produces wrong outputs with no error raised."

PR 作者采纳了这些建议，添加了 assert 和边界检查，并修复了方差归约逻辑，体现了良好的代码审查文化。

风险与影响

技术风险：

1. 回归风险：修改了核心的 RMSNorm 实现，如果头复制逻辑有误，可能导致其他 TP 配置下的模型输出错误。
2. 兼容性风险：MiniMaxM2RMSNormTP 的 __init__ 签名改变，破坏了向后兼容性，但仅影响 MiniMax M2.5 模型内部使用。

影响评估：

- 对用户：修复了 TP=16 时 MiniMax M2.5 模型的重复输出问题，提升了模型在高端硬件配置下的可用性。
- 对系统：仅影响 MiniMax M2.5 模型的 RMSNorm 实现，不涉及其他模型或子系统。
- 对团队：提供了头复制场景下权重分片的参考实现，可作为其他类似 TP 问题的解决模板。

关联脉络

从近期历史 PR 看，该 PR 与以下 PR 有相似之处：

- PR #22312：修复 GDN 内核以支持非连续张量输入，解决 Qwen3.5-27B 准确性回归问题，同为 bugfix 且涉及模型准确性。
- PR #22423：修复 Flux.2 模型 TI2I 准确性，通过对齐编码器、VAE 和图像预处理行为，同为准确性修复。

这些 PR 共同反映了团队对模型准确性的持续关注，特别是在复杂配置（如 TP>1、多模态）下的边缘 case 处理。该 PR 的解决方案——借鉴 QKVParallelLinear 的成熟模式——展示了代码复用的价值，为未来类似问题提供了参考。