PR #7039 完整报告

PaddlePaddle/FastDeploy

[Optimization] merge_allreduce

合并时间：2026-04-02 19:52

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7039

执行摘要

- 一句话：优化 GLM4-MoE 模型在纯 TP 并行模式下的 AllReduce 通信，合并两个 FFN 分支的归约操作。
- 推荐动作：该 PR 值得精读，特别是对于关注分布式优化和 MoE 模型性能的工程师。核心设计决策在于如何根据并行模式动态调整归约策略，这种条件化通信优化模式值得借鉴。建议重点关注 merge_ffn_tp 的判断逻辑和 reduce_results 的参数传递一致性。

功能与动机

根据 PR 描述中的 "Modifications" 部分，优化动机是 " 将普通专家和共享专家在计算 ffn 后各自的 allreduce 合并为一个 allreduce"。代码注释进一步说明：在纯 TP 模式下（tp>1, ep=1），两个分支都返回部分和，因此推迟 AllReduce 到合并之后可以节省一次集体通信。

实现拆解

实现主要分为三个部分：1) 在 glm4_moe.py 的 __init__中新增 merge_ffn_tp 标志，判断是否为纯 TP 模式（use_tp 且非 use_ep）。2) 修改 FusedMoE 和 SharedExperts 初始化参数，根据 merge_ffn_tp 设置 reduce_results（为 True 时内部执行归约，为 False 时返回部分和）。3) 在 forward 方法中，如果 merge_ffn_tp 为 True，则在合并两个专家输出后执行一次 tensor_model_parallel_all_reduce。此外，更新了两个测试文件中的基线路径和预期输出。

关键文件：

- fastdeploy/model_executor/models/glm4_moe.py（模块 model_executor/models）：核心优化逻辑所在，实现了 AllReduce 合并的条件判断和前向计算修改。
- tests/e2e/4cards_cases/test_GLM_45_AIR_mtp_tp4.py（模块 tests/e2e）：更新了端到端测试的预期输出，反映优化后模型输出的变化。
- tests/e2e/utils/rollout_routing_replay_test_utils.py（模块 tests/e2e/utils）：更新了测试基线路径，确保测试环境一致性。

关键符号：init, forward

评论区精华

review 讨论非常简短，仅有一次命名建议：zhoutianzi666 建议将变量名改为 self.merge_ffn_tp 以更易理解，作者 fxyfxy777 立即采纳并修改。没有出现设计争议或未解决

的疑虑。

- 变量命名优化 (style): 作者 fxyfxy777 采纳建议并修改

风险与影响

- 风险：主要风险包括：1) 条件逻辑风险：merge_ffn_tp 的判断条件（self.use_tp and not self.use_ep）需要确保在所有并行配置下正确，错误判断可能导致归约缺失或重复。2) 通信模式一致性：需确保 FusedMoE 和 SharedExperts 的 reduce_results 参数与 merge_ffn_tp 逻辑完全匹配，否则可能破坏分布式计算语义。3) 测试覆盖不足：Codecov 报告显示 patch 覆盖率仅 20%，有 4 行代码缺少测试覆盖，可能影响变更的可靠性验证。
- 影响：对系统性能有积极影响：在纯 TP 并行场景下减少一次 AllReduce 通信，可提升 MoE 模型推理吞吐量，尤其在大规模分布式训练 / 推理中效果显著。对用户透明，不影响 API 接口。对团队开发影响较小，但需要确保相关测试充分覆盖变更逻辑。
- 风险标记：条件逻辑风险，测试覆盖不足

关联脉络

- PR #7073 [OP] support deepgeem for sm103: 同属 Optimization 标签的 PR，涉及模型执行层的性能优化。
- PR #7001 [Feature] Support mtp overlap schedule: 同属性能优化相关 PR，关注调度和通信优化。
- PR #6993 [XPU] Refactor pre process: 同属模型执行层优化，涉及前处理逻辑重构。