PR #6986 完整报告

PaddlePaddle/FastDeploy

[Optimization] merge matmul and add

合并时间：2026-04-03 18:02

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/6986

执行摘要

此 PR 通过将线性层中的 matmul 和 add 操作合并为 paddle.nn.functional.linear，优化了未量化线性方法的性能。带 bias 时显著加速，小 shape 不带 bias 时略有性能下降，实现了性能与开销的权衡。变更影响核心模型执行路径，建议团队关注性能监控。

功能与动机

动机是性能优化。根据 PR body 描述："带 bias 情况基本上有加速，不带 bias 情况小 shape 下性能有下降（主要是 python 层 if 等调度开销）"。目标是减少算子调用开销，提升推理效率。

实现拆解

- 核心模块：fastdeploy/model_executor/layers/linear.py 中的 UnquantizedLinearMethod.apply 方法。
  - 修改前：使用 paddle.matmul 和 paddle.add。
  - 修改后：带 bias 时使用 paddle.nn.functional.linear，并添加 assert 验证 bias 形状；不带 bias 时保留 paddle.matmul。
- 测试模块：tests/e2e/utils/rollout_routing_replay_test_utils.py 更新基准路径，确保测试准确性。

评论区精华

- zhangbo9674建议："建议直接换成 paddle.nn.functional.linear"，引导了设计决策。
- qingqing01建议："建议单测不增加环境变量"，强调了测试简洁性，被采纳移除环境变量。
- AI 审核总结："代码变更逻辑正确，性能优化合理"，确认了不带 bias 时保留 matmul 的权衡。

风险与影响

- 技术风险：核心路径变更可能引入回归 bug，需通过测试覆盖验证；小 shape 不带 bias 场景性能下降，需监控实际影响。
- 影响范围：影响所有使用未量化线性层的模型，可能提升带 bias 操作的推理速度，但对团队要求更新测试基准。

关联脉络

与此前 PR 7039（优化 AllReduce 合并）类似，都属于模型执行层的性能优化，反映了团队在核心算子优化上的持续努力。近期 PR 如 7139 支持 GLM4.7 Flash，也涉及模型层优化，但本 PR 更专注于基础线性算子的改进。