PR #7337 完整报告

PaddlePaddle/FastDeploy

[RL]moe bf16 ep support paddle batch_gemm

合并时间：2026-04-11 21:51

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7337

执行摘要

本 PR 为 MoE（Mixture of Experts）层在 BF16 精度下的 EP（Expert Parallelism）prefill 阶段添加了 Paddle batched_gemm 支持，旨在对齐训练实现，提升推理一致性。变更涉及核心计算路径重构，但存在外部依赖未验证和测试覆盖不足的风险，需团队关注后续验证。

功能与动机

动机：根据 PR body，主要目标是“将 MoE BF16 精度下的 EP prefill group_gemm 和 swiglu 对齐训练”。这意味着推理阶段的实现需要与训练保持一致，以确保模型输出的正确性。作者进一步说明：“由于 batched_gemm 算子第三个输入需要 list，会破坏 decode 进 CUDAGraph，暂不适配”，因此变更仅针对 prefill 阶段。

实现拆解

实现围绕两个关键文件展开：

1. fastdeploy/model_executor/layers/moe/fused_moe_cutlass_backend.py：修改 apply_ep_prefill 函数，核心变更如下：
   - 移除原有的 compute_ffn 调用。
   - 使用 paddle.incubate.nn.functional.batched_gemm 进行两次矩阵乘法：python out = batched_gemm(permute_input, up_gate_proj_weight, recv_num_tokens_per_expert_list) if FD_MOE_PROB_IN_ADVANCE: out = paddlefleet_ops.fused_swiglu_scale(out, dst_weights) else: out = swiglu(out) ffn_out = batched_gemm(out, down_proj_weight, recv_num_tokens_per_expert_list)
   - 调整 moe_unpermute 中的 using_weighted_combine 参数逻辑，使用环境变量控制。
2. tests/layers/test_fused_moe_cutlass_backend.py：调整测试中 MoE 层权重的形状，从 [专家数, 输出维度, 输入维度] 改为 [专家数, 输入维度, 输出维度]，以匹配 batched_gemm 的输入要求，并添加 align 辅助函数。

评论区精华

review 讨论中，fastdeploy-bot 提出了多个关键问题，其中最值得关注的有：

> down_proj_bias 缺失处理：新代码使用 batched_gemm 后，未处理 layer.with_bias=True 时的 down_proj_bias，可能导致计算结果不正确。

> 函数名不一致：使用的 paddlefleet_ops.fused_swiglu_scale 函数与其他 backend 中的 paddlefleet_ops.fuse_weighted_swiglu_fp8_quant 不一致，且可能不存在，存在运行时 AttributeError 风险。
>
> 测试覆盖不足：变更核心计算逻辑但缺少测试覆盖，Codecov 报告显示 patch 覆盖率仅 66.67%。

讨论未显示这些问题是否被解决，但 PR 最终被合并，可能作者在后续提交中进行了修复或确认。

风险与影响

风险：

1. 正确性风险：如果 layer.with_bias=True，缺失 down_proj_bias 处理可能导致输出偏差。
2. 运行时风险：paddlefleet_ops.fused_swiglu_scale 函数可能不存在，引发运行时错误。
3. 测试风险：新代码路径测试覆盖不足，增加回归风险。

影响：

- 用户：需设置环境变量（如 FD_USE_PHI_MOE_PERMUTE=1）启用新路径，可能提升训练 - 推理一致性。
- 系统：仅影响 MoE EP prefill 计算，decode 阶段因 CUDAGraph 限制未适配，性能影响待评估。
- 团队：延续了近期 [RL] 标签的推理优化趋势，需关注外部依赖管理。

关联脉络

本 PR 是 FastDeploy 仓库中 MoE 模块持续优化的一部分：

- PR 7340：优化 MoE 层属性访问，使用缓存的 self.hidden_size，同涉及 MoE 优化。
- PR 7269：为 GLM4 MoE 模型添加 RMSNorm 支持，同标签 [RL]，显示推理优化是重点方向。
- PR 7259：为 NVFP4 MoE 添加 TBO 支持，同涉及 MoE 功能扩展。 这些 PR 共同显示团队在 MoE 性能和功能上持续投入，本 PR 通过对齐训练实现，进一步提升了推理阶段的可靠性。