PR #7308 完整报告

PaddlePaddle/FastDeploy

[TI-consistent] support quant use pow2scale

合并时间：2026-04-13 15:01

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7308

PR 分析报告：支持 FP8 量化使用 pow2scale 模式

执行摘要

本 PR 新增环境变量 FD_FP8_QUANT_WITH_POW2SCALE，支持在 FP8 量化中使用 pow2scale 模式，以对齐训练推理一致性（TI-consistent）。修改涉及配置、MoE 后端和量化层文件，增强系统可配置性，但需注意 review 中提到的逻辑不一致风险。

功能与动机

为满足训练推理一致性对齐需求，本 PR 在 FastDeploy 中新增环境变量控制，允许用户在 FP8 量化时选择使用 pow2scale 模式。动机源于训练和推理在量化精度上的对齐需求，通过环境变量 FD_FP8_QUANT_WITH_POW2SCALE 实现灵活控制。

实现拆解

实现分为两个关键部分：

1. 环境变量添加：在 fastdeploy/envs.py 中新增变量： python " FD_FP8_QUANT_WITH_POW2SCALE": lambda: bool(int(os.getenv("FD_FP8_QUANT_WITH_POW2SCALE", "0"))) 默认值为 0，确保向后兼容。同时调整了其他训练一致性相关变量（如 FD_USE_PHI_FP8_QUANT）的顺序，以逻辑分组。
2. 量化逻辑修改：在多个文件中更新 using_pow2_scale 参数：
   - fused_moe_deepgemm_backend.py 和 block_wise_fp8.py 中，参数改为 self.quant_config.deepgemm_scale_ue8m0 or fastdeploy.envs.FD_FP8_QUANT_WITH_POW2SCALE。
   - fused_moe_triton_backend.py 中，参数直接使用 fastdeploy.envs.FD_FP8_QUANT_WITH_POW2SCALE，但 review 指出此处需对齐。

关键影响函数包括 apply_ep_prefill、python_op_fused_moe_kernel_paddle 等，覆盖 MoE 激活和权重量化。

评论区精华

review 讨论中，fastdeploy-bot 指出了核心技术问题：

> 🤔 Bug此处 using_pow2_scale 直接使用 FD_FP8_QUANT_WITH_POW2SCALE，但没有考虑 quant_config.deepgemm_scale_ue8m0 的配置，与

> fused_moe_deepgemm_backend.py 中的行为不一致。

此问题涉及量化正确性，建议修改以整合现有配置。同时，讨论了 PR 规范问题，如标题标签应使用官方 [Quantization]，而非 [TI-consistent]。

风险与影响

风险：

- 逻辑不一致：fused_moe_triton_backend.py 中的 using_pow2_scale 逻辑可能导致量化模式错误，影响模型输出精度。
- 配置兼容性：新增环境变量默认值为 0，但需确保所有调用点正确处理，避免引入回归。
- 量化精度：pow2scale 模式可能改变量化行为，需测试验证对性能的影响。

影响：

- 用户可通过环境变量控制量化模式，提升训练推理一致性配置灵活性。
- 系统在 MoE 层和量化层应用此变量，可能影响运行时行为，范围限于相关模块。
- 团队需关注 review 未解决问题，以维护代码质量。

关联脉络

本 PR 是 FastDeploy 量化功能演进的一部分，与近期 PR 紧密相关：

- #7281：支持 CLI 配置量化参数，与本 PR 共同扩展量化配置选项。
- #7337：优化 MoE 层并涉及训练推理一致性（RL），与本 PR 在 MoE 量化改进上形成互补。
- #7269：通过环境变量控制 RMSNorm 用于训练对齐，与本 PR 使用环境变量实现一致性目标思路相似。

整体上，这些 PR 反映了系统在训练推理对齐和量化优化方面的持续投入，本 PR 是这一趋势的具体体现。