PR #7071 完整报告

PaddlePaddle/FastDeploy

[XPU] support glm-4.5-air (fix neox+partial_rotary_factor)

合并时间：2026-04-14 11:31

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7071

执行摘要

此 PR 为 XPU 平台添加 GLM-4.5-air 模型支持，通过优化 MoE 算子和 RoPE 实现，确保模型正确运行并提升性能。变更涉及多个核心组件，包括自定义算子、模型层和 worker 代码，但存在 breaking change 和测试覆盖不足等风险，建议技术团队仔细审查。

功能与动机

根据 PR body，动机是“为 XPU 平台添加 GLM-4.5-air 模型支持，优化 MoE 算子和 RoPE 实现。”作者在评论中补充说明，修复了从 --quantization 参数传 json 形式配置不生效的问题，以及 torch 模型开 ep+tp 报错的问题，以解决特定场景下的运行时错误。这扩展了 FastDeploy 在 XPU 后端的模型库，提升用户推理能力。

实现拆解

- MoE 算子层：修改 custom_ops/xpu_ops/src/ops/fused_noaux_tc.cc，将输出顺序从 {gating_logits, topk_ids, topk_weights} 调整为 {gating_logits, topk_weights, topk_ids}，数据类型从 INT64 改为 INT32，并在 fastdeploy/model_executor/layers/backends/xpu/moe/fused_moe.py 中适配解包顺序。
- 注意力层：更新 custom_ops/xpu_ops/src/ops/block_attn.cc，添加 rope_head_dim 参数传递，优化 encoder 模式下的 qkv 指针偏移，代码示例如下：cpp param.head_num, param.kv_head_num, param.head_dim, rope_head_dim, // 新增参数
- RoPE 层：在 fastdeploy/model_executor/layers/rotary_embedding.py 中，使用 current_platform.is_xpu() 检测平台，优化 GlmRotaryEmbedding 实现，并修复变量名错误（ictemb→emb）。
- 模型层：修正 fastdeploy/model_executor/models/glm4_moe.py 中 MergedReplicatedLinear 的参数名从 output_size 到 output_sizes。
- 辅助层：fastdeploy/model_executor/layers/linear.py 添加 output_dim 的 None 检查；fastdeploy/worker/xpu_model_runner.py 传递 partial_rotary_factor 参数以支持 RoPE 配置。

评论区精华

- 变量名错误：fastdeploy-bot 指出：“变量名拼写错误：ictemb 应为 emb。”作者已修复，确保代码正确性。

- Breaking Change：fastdeploy-bot 强调：“算子输出顺序变更，未同步更新所有调用方。”测试文件 test_fused_noaux_tc.py 未适配，存在未解决风险。
- 版本稳定性：fastdeploy-bot 建议：“Using latest version may cause CI instability. Consider using a specific version number.”但未采纳，可能引入依赖问题。
- 平台检测一致性：fastdeploy-bot 建议：“建议将 get_rope 函数中的平台检测也更新为 current_platform.is_xpu()。”显示代码风格优化空间。

风险与影响

- 技术风险：
  - Breaking Change：fused_noaux_tc 算子变更可能导致下游依赖错误，需紧急检查所有调用方。
  - 测试覆盖不足：Codecov 报告补丁覆盖率仅 22.22%，测试文件未完全适配，增加回归风险。
  - 版本不稳定：使用 latest 版本可能引发 CI 失败或部署不兼容。
  - 平台检测不一致：部分代码仍使用旧检测方法，可能影响跨平台兼容性。
- 影响范围：
  - 用户：XPU 平台用户获得 GLM-4.5-air 模型支持，扩展推理能力。
  - 系统：MoE 和 RoPE 优化可能提升性能，但需管理 breaking change 带来的维护成本。
  - 团队：变更涉及 OP、Models 等多个模块，需加强协作以确保代码质量。

关联脉络

- 与近期 PR #7029（XPU 算子重构）、#7361（MoE 参数支持）和 #7313（RoPE 优化）相关，共同推进 XPU 平台和模型模块的演进。
- 历史 PR 显示 FastDeploy 在 XPU、MoE 和 Optimization 领域持续投入，此 PR 是这一趋势的一部分，旨在增强模型兼容性和性能。