PR #7078 完整报告

PaddlePaddle/FastDeploy

[Iluvatar] Support wi4a16 group_gemm

合并时间：2026-03-30 19:03

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7078

执行摘要

本 PR 为 FastDeploy 的 Iluvatar GPU 后端新增了 wi4a16（权重 int4 激活 float16）的groupgemm支持，通过添加CUDA内核、更新量化逻辑和文档，扩展了模型的量化部署选项。变更影响特定硬件后端的推理流程，建议关注量化设计和兼容性限制，以平衡性能与正确性。

功能与动机

动机源自 PR body 中所述的“新增 feature：支持 w4a16”和“ci 迁移到新仓库 paddle-iluvatar，并修复 ci 报错问题”。具体来说，旨在为 Iluvatar GPU 提供更高效的量化推理能力，wi4a16 格式可降低内存占用并可能提升速度，同时调整 CI 以适配新仓库环境，确保自动化测试的稳定性。引用 PR body：“- 新增 feature：支持 w4a16 - ci 迁移到新仓库 paddle-iluvatar，并修复 ci 报错问题”。

实现拆解

实现分为多个层次，按模块梳理关键改动：

- CUDA 内核层：新增 wi4a16_group_gemm.cu 实现 group gemm 操作，支持 int4 权重和 float16 激活；新增 wi4a16_weight_quantize.cu 实现权重量化，采用每组对称 int4 量化，scale 计算为 maxlwl/7。cuda // 示例代码来自 wi4a16_group_gemm.cu std::vector<paddle::Tensor> WI4A16GroupGemm(const paddle::Tensor& x, const paddle::Tensor& weight, const paddle::Tensor& weight_scale, const paddle::Tensor& weight_zeros, const paddle::Tensor& prefix_sum, const int32_t group_size) { // 核心计算逻辑... }
- Python 整合层：修改 moe_ops.py，引入 wi4a16_group_gemm 并调整 _pre_process_expert_ffn 函数以支持新量化类型；修改 fuse_moe_cutlass_iluvatar_backend.py，在 process_prequanted_weights 中添加零张量处理和量化配置。
- 编译与配置：更新 setup_ops.py，将新文件加入编译列表，并统一 CUDA 编译标志以优化构建。
- 文档与测试：更新中英文安装文档，在“支持的量化策略”部分添加 wi4a16 说明；新增测试脚本 run_ernie_vl_28B_wint4.py 验证功能；调整 CI 脚本 run_ci_iluvatar.sh 使用新 paddle 包并添加测试项。
- 兼容性处理：在 iluvatar_model_runner.py 中添加断言，限制 wi4a16 不支持 CUDA 图，避免运行时错误。

评论区精华

review 中仅由维护者 EmmonsCurse 提交空批准，无技术讨论。Issue 评论中维护者要求跳过 CI 构建，例如："/skip-ci build_gpu /skip-ci build_xpu"，这表明 PR 已通过基础审查，但缺乏对设计、性能或正确性的深度交锋，可能依赖后续测试验证。

风险与影响

技术风险具体说明：

- 新 CUDA 内核 wi4a16_group_gemm.cu 和 wi4a16_weight_quantize.cu 涉及复杂量化逻辑，如共享内存规约和数值钳位，可能存在数值误差或边界条件处理不当，影响模型输出精度。
- 性能风险：wi4a16 量化虽提升效率，但实现优化不足（如内存访问模式）可能导致性能不如预期，需基准测试验证。
- 兼容性风险：iluvatar_model_runner.py 中明确断言 wi4a16 不支持 CUDA 图，限制高性能推理场景；同时，量化配置依赖于特定 group_size（128），可能不适用于所有模型架构。
- 测试覆盖风险：新增测试仅针对 ERNIE-VL 28B 模型，未覆盖其他模型或边缘情况（如不同输入形状、专家数量），可能遗漏回归问题。

影响范围：

- 用户：可使用 --quantization wint4 选项进行更轻量级推理，扩展部署灵活性；文档更新帮助用户正确配置。
- 系统：扩展了 MOE 系统的量化支持，增强系统模块化；新增代码增加维护复杂度。
- 团队：CI 流程更新可能影响后续开发和测试；团队需熟悉新量化逻辑以确保正确集成。

关联脉络

由于未提供历史 PR 分析，无法识别具体关联 PR。但基于代码变更，推测本 PR 可能与早期支持 w8a16 的 Iluvatar PR 有功能演进关系，例如修改了相同文件 w8a16_group_gemm.cu 和 w8a16_group_gemv.cu（函数重命名和参数调整），显示量化功能的逐步扩展。未来可能继续优化或添加更多硬件特定支持。