

PR #27075 完整报告

sgl-project/sglang

Add DeepGEMM prerelease wheel tests

合并时间: 2026-06-12 06:14

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27075>

执行摘要

- 一句话: 为 DeepGEMM wheel 添加 CI 预发布测试
- 推荐动作: 对于关注 CI/CD 和 DeepGEMM 发布流程的开发者值得精读, 尤其是如何通过环境变量确保依赖一致性, 以及将测试作为发布门禁的做法。建议后续补充 FP8 combine 和 MXF4 的测试覆盖。

功能与动机

避免类似 Issue #26541 (GLM-5-FP8 CUDA graph IMA) 的回归问题, 以及 DeepGEMM 仓库 PR #342 中暴露的测试缺口。在 wheel 构建后立即运行 DeepGEMM 测试套件, 可以在发布前捕获潜在的错误。

实现拆解

1. 在 workflow 顶层定义 TORCH_VER 环境变量 (值 2.11.0), 确保运行时 PyTorch 版本与 Docker 构建镜像一致, 并强制重装匹配的 torch。
2. 新增 test-cu129 和 test-cu130 作业, 依赖同 CUDA 版本的构建矩阵, 在 8-GPU H200 (sm90) 和 8-GPU B200 (sm100) 上安装 wheel 并执行 DeepGEMM 测试脚本 run_tests.sh。
3. 修改 release-cu129 和 release-cu130 作业, 将其 needs 列表加入对应的测试作业, 将 PyPI 上传后移到测试通过之后, 形成“构建 → 测试 → 发布”的门禁流程。

关键文件:

- .github/workflows/release-whl-deepgemma.yml (模块 CI/ 部署; 类别 infra; 类型 infrastructure): 此文件是唯一的变更文件, 新增了测试作业并重组了发布流程, 是 PR 的核心载体。

关键符号: 未识别

评论区精华

Review 中 Fridge003 提出了几点建议:

- 1) 将测试仓库分支固定为 dev 以避免硬编码;
- 2) 考虑将依赖管理移至 DeepGEMM 仓库侧;
- 3) 减少测试 GPU 数量以节省资源;
- 4) 增加 sm103 架构的测试。b8zhong 接受了分支固定建议, 并解释了依赖版本对齐的必要性, 但对减少 GPU 和增加 sm103 测试未做实际修改。最

终 Fridge003 给予 APPROVE。

- 将测试仓库分支固定为 dev (other): 已采用, 代码中改为 `ref: ${{ inputs.branch || 'dev' }}`。
- 依赖管理放在 DeepGEMM 侧 (design): 保留在 CI 侧, 添加了 TORCH_VER 环境变量和强制重装。
- 减少测试 GPU 数量 (performance): 未实施, 仍使用 8 卡 runner。

风险与影响

- 风险: 主要风险包括: CI 测试占用较多 GPU 资源 (8 卡 H200/B200), 可能导致排队; 测试依赖外部仓库分支 (dev), 如果被破坏可能导致 CI 失败; 测试覆盖仍不完全 (FP8 combine / MXF4 路径无测试), 无法保证所有功能正确。此外, 如果运行时 torch 版本与 wheel 构建版本不匹配, 可能出现链接错误, 已通过强制 reinstall 缓解。
- 影响: 对用户无直接影响但提升了发布质量; 对系统 CI 流程增加了 2 个测试作业 (每个大约 2 小时), 延长了发布总时间; 对团队增加了维护测试环境和 runner 的工作量。
- 风险标记: 缺少 FP8 combine 和 MXF4 测试覆盖, CI 资源消耗大, 依赖外部分支 (dev)

关联脉络

- PR #27922 fix(deepgemm): align PP-parallel warmup bs to CP padding: 都与 DeepGEMM 相关, 前者修复编译对齐问题, 本 PR 增加测试覆盖, 共同保障 DeepGEMM 在 SGLang 中的可靠性。