

PR #2670 完整报告

radixark/miles

fix(docker): update torch_memory_saver for CUDA VMM granularity

合并时间: 2026-08-20 13:57

原文链接: <http://prhub.com.cn/radixark/miles/pull/2670>

执行摘要

- 一句话: 更新 torch_memory_saver 依赖, 修复 CUDA VMM 分配粒度问题
- 推荐动作: 这是一个低风险的依赖升级 PR, 值得关注的是它解除了重要的训练阻塞。建议阅读以了解团队维护第三方依赖 pin 的实践, 以及如何处理此类上游修复的集成节奏。

功能与动机

PR 背景明确: GB200 OPD workflows 需要使用 torch_memory_saver 的 CUDA VMM 分配粒度修复, 该修复由作者本人 (kaixih) 在上游 PR fzyzcjy/torch_memory_saver#82 中贡献, 但当前镜像中 pin 的版本过旧, 未包含此修复, 导致相关训练流程 (如两节点 GB200 HybridEP 训练, 关联 Issue #1726) 被阻塞。作者在 Issue 评论中询问更新 TMS pin 的惯例, 并请求确认。

实现拆解

1. 修改 docker/Dockerfile 中 torch_memory_saver 的安装 pin, 从提交 74d68c5e4bedf2b6774f2c92ed0f81b7c8d91ed0 更新为 f05a8754daf68238d54e4cf31cb3ba866684bbaf (即上游 PR #82 的合并提交)。
2. 同步修改 docker/Dockerfile.rocm 中相同的 pin, 保持两个镜像依赖一致。
3. 无其他源码或测试改动, 仅为基础设施依赖升级。

关键文件:

- docker/Dockerfile (模块 部署脚本; 类别 infra; 类型 infrastructure) : CUDA 镜像中 torch_memory_saver 的安装 pin 被更新到上游修复提交, 是本次变更的核心文件。
- docker/Dockerfile.rocm (模块 部署脚本; 类别 infra; 类型 infrastructure) : ROCm 镜像中相同的 pin 更新, 保持依赖一致性。

关键符号: 未识别

评论区精华

Issue 评论中 kaixih 询问更新 TMS pin 的惯例, 并指出该版本包含其添加的修复, 用于解阻 GB200 OPD 实验。@maocheng23 和 @yueming-yuan 被提及, 无公开回复记录。Review 中 maocheng23 直接批准 (APPROVED), 无其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。升级 torch_memory_saver 版本可能带来新行为，但上游修复已通过测试，且该库主要用于 RoCE 等场景的内存管理。主要风险是 pin 的提交在大型仓库中通常稳定，但若上游 PR 被 revert 或存在未发现的回归，可能影响相关训练。另外，两个镜像的 pin 同步更新，若 CUDA 与 ROCm 路径行为不一致，需关注。但本 PR 仅改动依赖版本，回归风险有限。
- 影响：影响范围限于 Docker 镜像构建，对使用最新镜像的 CUDA 与 ROCm 用户产生依赖变化。对开发团队而言，解除了 GB200 OPD 工作流的阻塞，使得相关训练实验能够继续。镜像层面需重新构建，但无需代码迁移。影响程度中等。
- 风险标记：依赖更新，镜像构建影响

关联脉络

- PR #1726 [OPD] Stabilize two-node GB200 HybridEP training: 该 PR 正是本次依赖升级所解阻的工作流，二者直接关联。
- PR #2660 [AMD] Point the ROCm image at the v0.5.16 wheels release with TE 2.17.0: 同为 Docker 镜像依赖更新，且涉及 ROCm 构建。