

PR #2172 完整报告

THUDM/slime

[docker] Update training side dependencies

合并时间: 2026-07-03 15:07

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2172>

执行摘要

- 一句话: 更新 Docker 训练依赖, 升级 FlashLinearAttention 版本
- 推荐动作: 建议快速合并, 是常规的依赖维护更新。

功能与动机

保持训练镜像中依赖的最新兼容性, 修复可能存在的 PyTorch 版本兼容问题 (参考 comment 中关于 #168167 的说明)。

实现拆解

1. 升级 flash-linear-attention: 在 docker/Dockerfile 中将 flash-linear-attention 版本从 0.4.1 升级到 0.4.2。
2. 移除 nvidia-cudnn-cu12 显式安装: 删除 `pip install nvidia-cudnn-cu12==9.16.0.29` 行, 因为该依赖已不再是必须或由基础镜像提供 (关联 PyTorch issue #168167 已解决)。
3. 更新版本号: 在 docker/version.txt 中将版本从 nightly-dev-20260703a 更新为 nightly-dev-20260703b。
4. 提交历史包含 2 次 bugfix 调整, 最终版本确定。

关键文件:

- docker/Dockerfile (模块 部署脚本; 类别 infra; 类型 infrastructure): 核心变更文件, 更新了依赖版本并移除了不必要的安装步骤。
- docker/version.txt (模块 部署脚本; 类别 docs; 类型 documentation): 同步更新 Docker 镜像版本号。

关键符号: 未识别

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅限于依赖版本升级和移除一个可选依赖，不涉及核心逻辑。但 flash-linear-attention 升级可能引入微小的行为变化，需确保与当前模型训练兼容。
- 影响：影响范围仅限于 Docker 训练镜像的用户。升级 flash-linear-attention 可能带来性能或功能改进，移除 nvidia-cudnn-cu12 可减少镜像大小并避免潜在冲突。
- 风险标记：依赖版本变更

关联脉络

- PR #2173 [docker] Update SGLang patch for PD R3 routed experts: 同为 Docker 依赖更新 PR，涉及类似的基础设施维护。