

PR #2178 完整报告

THUDM/slime

[docker] Update dependencies

合并时间: 2026-07-05 21:33

原文链接: <http://prhub.com.cn/THUDM/slime/pull/2178>

执行摘要

- 一句话: Docker 依赖批量升级关键库
- 推荐动作: 建议阅读该 PR 的 Dockerfile 变更, 了解依赖版本对齐策略。可考虑在后续 PR 中补充集成测试以验证兼容性。

功能与动机

为了保持与 TransformerEngine 2.16 及相应 FlashAttention 版本的兼容性, Docker 镜像需要更新依赖, 确保在 SGLang 基础镜像中移除有冲突的 FlashAttention-4 后端。

实现拆解

1. Dockerfile 依赖调整:
 - 添加显式卸载 FA4 的命令以避免后端选择冲突。
 - 升级 FlashAttention 可从 2.7.4.post1 升级至 2.8.3。
 - 更新 FA3 源码构建: 使用特定 commit 并添加 FLASH_ATTENTION_FORCE_BUILD=TRUE 环境变量确保强制构建。
 - 升级 TransformerEngine: CUDA 12 环境下使用 2.16.1 wheel, CUDA 13 环境下从 release_v2.16 分支源码构建。
2. 版本标签更新: docker/version.txt 从 nightly-dev-20260703b 更新为 nightly-dev-20260705a 以反映新的构建版本。

关键文件:

- docker/Dockerfile (模块 Docker; 类别 infra; 类型 infrastructure): 核心变更文件, 修改了 FlashAttention 和 TransformerEngine 的安装版本和构建选项, 影响整个训练环境的依赖栈。
- docker/version.txt (模块 Docker; 类别 docs; 类型 documentation): 更新 Docker 镜像版本标签, 与依赖变更对应, 方便追溯镜像内容。

关键符号: 未识别

关键源码片段

[docker/Dockerfile](#)

核心变更文件，修改了 FlashAttention 和 TransformerEngine 的安装版本和构建选项，影响整个训练环境的依赖栈。

```
# ...
# The SGLang base image may include FA4. The validated TE 2.16 CP stack uses
# FA2 + FA3, so remove FA4 explicitly to avoid backend selection conflicts.
RUN pip uninstall -y flash-attn-4 flash_attn_4 || true

# The compilation is slow, thus should be put at top
RUN MAX_JOBS=64 pip -v install flash-attn==2.8.3 --no-build-isolation

# The compilation is slow, thus should be put at top. This FA3 commit uses the
# window_size_left/window_size_right API required by TransformerEngine 2.16 CP.
RUN git clone https://github.com/Dao-AILab/flash-attention.git && \
  cd flash-attention/ && git checkout 002cce0a1068f8c07dfccb5a1d232b9a3276947c && git
  submodule update --init && \
  cd hopper/ && \
  FLASH_ATTENTION_FORCE_BUILD=TRUE MAX_JOBS=96 pip -v install . --no-build-isolation
  && \
  cd /root/ && rm -rf flash-attention/

# ... TransformerEngine 升级:
RUN if [ "${ENABLE_CUDA_13}" = "1" ]; then \
  pip install nvidia-mathdx==26.6.0 && \
  pip -v install --no-build-isolation git+https://github.com/NVIDIA/TransformerEngine.
  git@release_v2.16; \
else \
  pip -v install --no-build-isolation "transformer_engine[pytorch]==2.16.1"; \
fi
# ...
```

评论区精华

该 PR 没有讨论记录。

- 暂无高价值评论线程

风险与影响

- 风险：依赖版本升级可能导致与其他组件（如 SGLang 补丁）的兼容性问题，尤其在 FlashAttention 后端切换时。需确保 CI 测试覆盖典型训练流程。TransformerEngine 2.16 的引入可能影响 CUDA 13 环境下的编译稳定性。FA3 强制构建可能增加镜像构建时间。
- 影响：直接影响所有使用该 Docker 镜像的训练任务，涉及 FlashAttention 和 TransformerEngine 的版本变更，可能改变训练速度与内存占用。对于使用旧依赖的外部代码可能需要适配。低风险，但建议在 CI 中验证关键训练用例。
- 风险标记：依赖升级可能不兼容，构建时间增加

关联脉络

- PR #2172 [docker] Update training side dependencies: 同样是更新 Docker 训练依赖, 涉及版本对齐和库升级, 具有类似动机。