

PR #7035 完整报告

verl-project/verl

[ci] test: update ascend ci docker tag

合并时间: 2026-07-15 10:02

原文链接: <http://prhub.com.cn/verl-project/verl/pull/7035>

执行摘要

- 一句话: 更新 Ascend CI Docker 镜像标签并移除旧 Qwen3.5 构建工作流
- 推荐动作: ### 建议 1. 立即修复 shell 脚本: 将 `enforce_eager` 改回 `True`, 并移除 `cuda_graph_mode` 配置, 以保持与 Ascend NPU 的兼容性。 2. 验证 CI 新镜像: 检查所有更新标签后的 CI 工作流是否正常运行, 尤其是 `nightly` 和 `e2e` 测试。 3. 关注讨论中的未解决意见: 虽然 PR 已合并, 但留下了一个已知问题, 建议在后续 PR 中解决。

功能与动机

根据 PR 标题和描述, 目的是更新 Ascend CI 的 Docker 镜像标签, 确保 CI 使用最新构建的镜像。这可能是因为之前的镜像标签格式不够规范或需要统一命名, 以及 Qwen3.5 的专用构建工作流已被新的通用工作流取代。

实现拆解

实现拆解

1. 删除旧 Qwen3.5 专用构建工作流: 删除 `.github/workflows/docker-build-ascend-a2-qwen3_5.yml` 和 `.github/workflows/docker-build-ascend-a3-qwen3_5.yml`, 共移除约 246 行。这两个工作流用于构建 Qwen3.5 专用 Docker 镜像, 可能已被合并到通用工作流中。
2. 更新 CI 工作流的镜像标签: 修改 `nightly_ascend.yml`, `model_ascend.yml`, `e2e_ascend.yml` 等 8 个 CI 工作流文件, 将容器镜像标签从类似 `verl-9.0.0-910b-ubuntu22.04-py3.11-latest` 的格式统一替换为 `latest-cann9.0.0-torch_npu2.9.0post2-910b-ubuntu22.04-py3.11-vllm` 的格式, 明确包含 CANN 和 Torch 版本等详细信息。
3. 修改 GRPO 训练示例配置: 修改 `verl/experimental/one_step_off_policy/shell/grpo_qwen3_8b_gsm8k_fsdp2_8_8_npu.sh`, 将 `actor_rollout_ref.rollout.enforce_eager` 从 `True` 改为 `False`, 并添加了 `cuda_graph_mode` 配置。但此修改在 Ascend NPU 上可能不兼容, 因为 CUDA graphs 不受支持。

关键文件:

- `.github/workflows/docker-build-ascend-a2-qwen3_5.yml` (模块 CI 工作流; 类别 `infra`; 类型 `deletion`): 删除此文件, 移除了 Qwen3.5 专用 Docker 构建工作流, 是本次变更中最大的结构调整。

- `.github/workflows/docker-build-ascend-a3-qwen3_5.yml` (模块 CI 工作流; 类别 infra; 类型 deletion) : 与上一个文件对称, 删除 A3 版本的 Qwen3.5 专用构建工作流。
- `.github/workflows/nightly_ascend.yml` (模块 CI 工作流; 类别 infra; 类型 infrastructure) : 包含最全面的镜像标签更新, 覆盖多个 job, 是理解变更核心的入口。
- `verl/experimental/one_step_off_policy/shell/grpo_qwen3_8b_gsm8k_fsdp2_8_8_npu.sh` (模块 实验脚本; 类别 other; 类型 configuration) : 该脚本的配置修改可能引入 Ascend NPU 兼容性问题, 是讨论焦点。
- `.github/workflows/model_ascend.yml` (模块 CI 工作流; 类别 infra; 类型 infrastructure) : 同样更新了镜像标签, 且是模型测试 CI 的一部分。

关键符号: 未识别

关键源码片段

`.github/workflows/nightly_ascend.yml`

包含最全面的镜像标签更新, 覆盖多个 job, 是理解变更核心的入口。

```
# 更新 CI 容器镜像标签 (部分示例)
# 旧标签格式: verl-9.0.0-910b-ubuntu22.04-py3.11-latest
# 新标签格式: latest-cann9.0.0-torch_npu2.9.0post2-910b-ubuntu22.04-py3.11-vllm
jobs:
  build:
    container:
      # 变更前
      # image: swr.cn-southwest-2.myhuaweicloud.com/modelfoundry/ascend-ci/verl/verl:verl-9.0.0-910b-ubuntu22.04-py3.11-latest
      # 变更后
      image: swr.cn-southwest-2.myhuaweicloud.com/modelfoundry/ascend-ci/verl/verl:latest-cann9.0.0-torch_npu2.9.0post2-910b-ubuntu22.04-py3.11-vllm
# 更多类似更新见实际文件
```

评论区精华

讨论焦点

- `gemini-code-assist[bot]`指出在 Ascend NPU 上 CUDA graphs 不受支持, 设置 `enforce_eager=False` 并启用 `cudagraph_mode` 会导致 vLLM 初始化失败。建议改回 `enforce_eager=True` 并移除 `cudagraph_mode`。然而, 该评论未被采纳, PR 在无进一步讨论的情况下被批准合并。这可能导致 NPU 训练环境出现运行时错误。
- Shell 脚本 `enforce_eager` 设置与 NPU 不兼容 (correctness): 该问题未解决, PR 仍被批准合并。

风险与影响

- 风险: ### 风险分析
- shell 脚本配置不兼容: `grpo_qwen3_8b_gsm8k_fsdp2_8_8_npu.sh` 中 `enforce_eager=False` 和 `cudagraph_mode` 的设置与 Ascend NPU 不兼容, 直接使用将导

致 vLLM 崩溃。这是未解决的高风险问题。

- 镜像标签变更：新的 Docker 镜像标签指向新构建的镜像，可能包含不同的依赖版本，存在回归风险。需要 CI 验证通过。
- 删除旧工作流：如果还有外部依赖引用旧工作流，可能会受影响。但考虑到是 CI 内部文件，影响可控。
- 影响：### 影响分析
- CI 系统：所有 Ascend NPU 相关的 CI 工作流都将使用新的 Docker 镜像，可能带来依赖或行为变化。
- 开发者：使用 `grpo_qwen3_8b_gsm8k_fsdp2_8_8_npu.sh` 脚本运行 GRPO 训练的开发者将遇到 NPU 兼容性问题。
- 团队：需要尽快修复 shell 脚本配置，并确保新镜像在 CI 中通过测试。
- 风险标记：shell 脚本 NPU 配置不兼容，CI 镜像更新依赖验证

关联脉络

- PR #7028 [env] fix: Sync Qwen3.5 Dockerfile from 0.8.0 Branch to Main and Pin PyArrow Version: 同样涉及 Qwen3.5 Docker 镜像的同步和修复，与本文删除 Qwen3.5 专用构建工作流有间接关联。
- PR #7008 [doc] refactor: update ascend docker name: 更新了 Ascend Docker 镜像命名规范，本 PR 进一步应用了新的命名格式。
- PR #7039 [env] chore: add megatron-bridge in sglang docker env: 同为 Ascend Docker 环境变更，但侧重不同镜像。