

PR #6098 完整报告

verl-project/verl

[fsdp] feat: Qwen3.5 Adds Docker Files Based on CANN8.5.2 A2

合并时间: 2026-04-21 21:36

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6098>

执行摘要

- 一句话: 新增基于 CANN 8.5.2 A2 的 Qwen3.5 Docker 镜像构建文件, 支持多架构 CI/CD 流水线。
- 推荐动作: 该 PR 主要涉及基础设施, 对核心算法或训练逻辑无直接影响, 适合负责部署或 NPU 平台的工程师精读。重点关注 Dockerfile 中的安全警告(证书验证)和环境变量持久化问题, 这些是镜像可靠性的关键。设计上展示了多架构镜像构建的 CI/CD 模式, 值得借鉴。

功能与动机

PR 标题和提交信息表明, 需要为 Qwen3.5 模型在 Ascend NPU 平台上提供基于 CANN 8.5.2 A2 的 Docker 镜像构建能力。结合近期历史 PR (如 #6066、#5967) 中对 NPU 平台支持、脚本参数和补丁的持续优化, 此 PR 旨在扩展基础设施支持, 为特定硬件 (Ascend 910b1) 和模型 (Qwen3.5) 提供标准化的部署环境。

实现拆解

1. 新增 Dockerfile 定义镜像构建步骤: 在 `docker/ascend/Dockerfile.ascend_8.5.2_a2_qwen3-5` 中, 基于官方 CANN 8.5.1 基础镜像, 升级到 CANN 8.5.2, 安装系统依赖, 并克隆 `vllm`、`vllm-ascend` 等仓库, 最后安装 `verl` 项目。
2. 新增 CI/CD 工作流实现自动化构建: 在 `.github/workflows/docker-build-ascend-a2-qwen3_5.yml` 中, 定义 GitHub Actions 工作流, 支持手动触发, 使用矩阵策略并行构建 `amd64` 和 `arm64` 架构镜像, 通过 Buildx 推送镜像层, 并合并为多架构镜像后打上标签 `verl-8.5.2-910b-ubuntu22.04-py3.11-qwen3-5`。
3. 配置构建参数与安全措施: 工作流中配置了 Quay.io 仓库推送、缓存优化和磁盘清理步骤, 但 Dockerfile 中使用了 `--no-check-certificate` 等存在安全风险的命令 (已在 review 中提出)。

关键文件:

- `docker/ascend/Dockerfile.ascend_8.5.2_a2_qwen3-5` (模块 Docker 配置; 类别 `infra`; 类型 `infrastructure`): 定义了基于 CANN 8.5.2 A2 的 Qwen3.5 Docker 镜像构建步骤, 是核心基础设施文件。
- `.github/workflows/docker-build-ascend-a2-qwen3_5.yml` (模块 CI 流水线; 类别 `infra`; 类型 `infrastructure`): 定义了自动化构建、推送和合并多架构 Docker 镜像的 CI/CD 工作

流。

关键符号：未识别

关键源码片段

docker/ascend/Dockerfile.ascend_8.5.2_a2_qwen3-5

定义了基于 CANN 8.5.2 A2 的 Qwen3.5 Docker 镜像构建步骤，是核心基础设施文件。

```
# 基于官方 CANN 8.5.1 基础镜像
```

```
FROM swr.cn-south-1.myhuaweicloud.com/ascendhub/cann:8.5.1-910b-ubuntu22.04-py3.11
```

```
ARG SOC_VERSION="ascend910b1"
```

```
# 安装系统依赖，包括 gcc、cmake、wget 等
```

```
RUN apt-get update -y && \
```

```
    apt-get install -y --no-install-recommends gcc g++ cmake libnuma-dev wget git curl jq vim  
    build-essential && \
```

```
    apt-get clean && \
```

```
    rm -rf /var/lib/apt/lists/* /tmp/* /var/tmp/* && \
```

```
    pip install --upgrade pip packaging setuptools==80.10.2 && \
```

```
    pip cache purge
```

```
# 升级 CANN 到 8.5.2 版本，根据架构下载并安装
```

```
RUN set -e && \
```

```
    cd /tmp && \
```

```
    ARCH=$(uname -m) && \
```

```
    if [ "$ARCH" = "aarch64" ]; then \
```

```
        # 注意：使用--no-check-certificate存在安全风险，应移除
```

```
        wget -v --no-check-certificate https://ascend-repo.obs.cn-east-2.myhuaweicloud.com/  
        CANN/CANN%208.5.2/Ascend-cann-toolkit_8.5.2_linux-aarch64.run && \
```

```
        bash Ascend-cann-toolkit_8.5.2_linux-aarch64.run --quiet --install && \
```

```
        rm -f Ascend-cann-toolkit_8.5.2_linux-aarch64.run && \
```

```
        # 类似步骤安装其他组件...
```

```
    elif [ "$ARCH" = "x86_64" ]; then \
```

```
        wget -v --no-check-certificate https://ascend-repo.obs.cn-east-2.myhuaweicloud.com/  
        CANN/CANN%208.5.2/Ascend-cann-toolkit_8.5.2_linux-x86_64.run && \
```

```
        bash Ascend-cann-toolkit_8.5.2_linux-x86_64.run --quiet --install && \
```

```
        rm -f Ascend-cann-toolkit_8.5.2_linux-x86_64.run && \
```

```
        # 类似步骤安装其他组件...
```

```
    fi && \
```

```
    source /usr/local/Ascend/cann/set_env.sh && \ # 环境变量未持久化，可能导致后续步骤缺失
```

```
    rm -rf /tmp/*
```

```
# 克隆 vllm 和 vllm-ascend 仓库，为后续安装准备
```

```
RUN ARCH=$(uname -m) && \
```

```
    echo "[LOG INFO] Detected architecture: $ARCH" && \
```

```
    if [ "$ARCH" = "x86_64" ]; then \
```

```
        pip config set global.extra-index-url "https://download.pytorch.org/whl/cpu/"; \
```

```
fi && \
git clone --depth 1 --branch v0.18.0 https://github.com/vllm-project/vllm.git && \
git clone https://github.com/vllm-project/vllm-ascend.git && \
cd vllm-ascend && git checkout 0.1.0 && cd ..

# 设置全局构建并行度限制，可能影响性能
ENV MAX_JOBS=1 \
  MAKEFLAGS="-j1" \
  CMAKE_BUILD_PARALLEL_LEVEL=1

# 安装 verl 项目
RUN pip install -e .
```

评论区精华

review 中 gemini-code-assist[bot] 提出了三个关键问题：

1. 安全风险：Dockerfile 中使用 `wget --no-check-certificate` 绕过 SSL/TLS 验证，存在中间人攻击风险，建议更新 `ca-certificates` 并移除该标志。
 2. 性能影响：全局设置 `MAX_JOBS=1` 等环境变量会显著降低所有后续构建步骤的速度，建议仅针对特定命令（如 `vllm-ascend` 安装）设置。
 3. 环境变量持久性：在 `RUN` 指令中使用 `source` 或 `export` 设置的环境变量不会持久化到后续步骤或容器运行时，可能导致 Ascend 环境设置缺失，建议使用 `ENV` 指令或写入 `/etc/bash.bashrc`。这些评论均未在 PR 中直接回复或解决，但 PR 已被 `wucong25` 批准合并，表明团队可能接受这些风险或计划后续处理。
- Dockerfile 安全风险：使用 `--no-check-certificate (security)`：建议更新 `ca-certificates` 并移除该标志，但 PR 中未直接回应。
 - 构建性能：全局并行度限制 (`performance`)：建议仅针对特定命令设置，但 PR 中未修改。
 - 环境变量持久性：`source` 命令不持久 (`correctness`)：建议使用 `ENV` 指令或写入 `bashrc`，但 PR 中未调整。

风险与影响

- 风险：
 1. 安全风险：Dockerfile 中 `wget --no-check-certificate` 绕过证书验证，可能引入恶意软件或中间人攻击，影响镜像完整性。
 2. 性能风险：全局构建并行度限制 (`MAX_JOBS=1`) 会显著延长镜像构建时间，尤其在 CI/CD 流水线中可能成为瓶颈。
 3. 环境配置风险：Ascend 环境变量（如 `LD_LIBRARY_PATH`）未持久化，可能导致容器运行时依赖库缺失，引发运行时错误。
 4. 兼容性风险：镜像基于特定 CANN 版本（8.5.2）和硬件（Ascend 910b1），可能不适用于其他 NPU 型号或 CANN 版本。
- 影响：

1. 对用户的影响：为在 Ascend NPU 上运行 Qwen3.5 模型的用户提供了标准化的 Docker 镜像，简化了环境部署，但需注意安全警告。
2. 对系统的影响：新增了 CI/CD 流水线，增加了仓库的自动化构建能力，但未修改核心训练或推理逻辑，不影响现有功能。
3. 对团队的影响：扩展了基础设施支持范围，与近期 NPU 相关 PR（如 #6066、#5967）形成协同，但引入的安全和性能问题需后续关注。 - 风险标记：安全风险：证书验证绕过，性能风险：全局构建并行度限制，环境配置风险：变量未持久化

关联脉络

- PR #6066 [recipe, cfg] fix: update NPU script parameters for Qwen3 GSPO and DAPO math recipes: 同样涉及 NPU 平台和 Qwen3 模型，调整训练脚本参数，与此 PR 的 Docker 镜像支持形成互补。
- PR #5967 [fully_async] fix: Add Mindspeed Patch for Async Training on Ascend NPUs: 涉及 Ascend NPU 上的异步训练补丁，与此 PR 共同扩展 NPU 基础设施支持。