PR #6021 完整报告

verl-project/verl

[ci] chore: add sglang new version docker for NPU

合并时间：2026-04-16 14:15

原文链接：http://prhub.com.cn/verl-project/verl/pull/6021

执行摘要

- 一句话：新增 Ascend NPU 的 SGLang v0.5.10 Dockerfile 及配套 CI 和文档更新。
- 推荐动作：对于基础设施维护者和 NPU 用户，此 PR 值得浏览以了解最新 NPU 环境配置。建议关注 Dockerfile 中的安全优化建议，未来可考虑采纳以提高镜像安全性和效率。

功能与动机

PR body 中说明：'Add new dockerfile with CANN 8.5.0 base image, update sglang version to v0.5.10, sgl-kernel-npu-2026.02.01, Mindspeed version to 2.3.0_core_r0.12.1.'，旨在为 NPU 硬件提供最新的 SGLang 支持，确保训练环境与最新库版本兼容。

实现拆解

1. 新增 Dockerfile：在 docker/ascend/ 目录下添加 Dockerfile.ascend.sglang_8.5.0_a2 和 _a3，基于 CANN 8.5.0 基础镜像，定义安装系统依赖、克隆 sglang v0.5.10 和 MindSpeed 仓库、安装相关 NPU 库（如 torch_npu、sgl-kernel-npu）及 verl 项目的步骤。这样改是为了提供标准化、可复制的 NPU SGLang 环境。
2. 更新 CI 工作流：修改 .github/workflows/docker-build-ascend-sglang-a2.yml 和 -a3.yml，将路径引用从旧版 Dockerfile（如 8.3.rc1）更新到新版 8.5.0，确保 CI 在代码变更时自动构建和推送新镜像。这保证了基础设施的持续集成。
3. 同步文档：在 docs/ascend_tutorial/quick_start/dockerfile_build_guidance.rst 中添加新 Dockerfile 的链接，帮助用户快速查找和使用最新版本。这是配套文档更新，提升用户体验。影响：为 Ascend NPU 用户提供最新 SGLang 训练环境，CI 自动化构建减少手动部署负担。

关键文件：

- docker/ascend/Dockerfile.ascend.sglang_8.5.0_a2（模块 Docker 镜像；类别 infra；类型 infrastructure）：新增 A2 架构的 SGLang Dockerfile，定义了从基础镜像到完整环境的构建流程，是 NPU 部署的核心文件。
- docker/ascend/Dockerfile.ascend.sglang_8.5.0_a3（模块 Docker 镜像；类别 infra；类型 infrastructure）：新增 A3 架构的 SGLang Dockerfile，结构与 A2 类似但针对不同硬件变体，确保跨架构支持。

- .github/workflows/docker-build-ascend-sglang-a2.yml（模块 CI 流水线；类别 infra；类型 infrastructure）：更新 CI 工作流以引用新 Dockerfile 路径，确保代码变更时自动触发镜像构建和发布。
- .github/workflows/docker-build-ascend-sglang-a3.yml（模块 CI 流水线；类别 infra；类型 infrastructure）：类似 A2 工作流更新，确保 A3 架构的 CI 自动化。
- docs/ascend_tutorial/quick_start/dockerfile_build_guidance.rst（模块 文档教程；类别 docs；类型 documentation）：更新文档添加新 Dockerfile 链接，帮助用户快速访问和使用最新 NPU SGLang 环境。

关键符号：未识别

关键源码片段

docker/ascend/Dockerfile.ascend.sglang_8.5.0_a2

新增 A2 架构的 SGLang Dockerfile，定义了从基础镜像到完整环境的构建流程，是 NPU 部署的核心文件。

```
# 基于CANN 8.5.0的Ascend NPU基础镜像，针对A2架构
FROM swr.cn-south-1.myhuaweicloud.com/ascendhub/cann:8.5.0-910b-ubuntu22.04-py3.11

# 设置pip镜像源以加速中国区下载
ARG PIP_INDEX_URL="https://mirrors.aliyun.com/pypi/simple"
ARG PTA_BASE_VERSION="torch_npu-2.8.0.post2-cp311-cp311-manylinux_2_28"
ARG PTA_URL="https://gitcode.com/Ascend/pytorch/releases/download/v7.3.0-pytorch2.8.0"

# 安装系统依赖并配置pip环境
RUN apt-get update -y && \
    apt-get install -y --no-install-recommends gcc g++ cmake libnuma-dev wget git curl jq vim
    build-essential net-tools iputils-ping unzip ca-certificates && \
    apt-get clean && \
    rm -rf /var/lib/apt/lists/* /tmp/* /var/tmp/* && \
    pip config set global.index-url ${PIP_INDEX_URL} && \
    pip config set install.trusted-host mirrors.aliyun.com && \
    pip install --upgrade pip setuptools==80.10.2 packaging && \
    pip cache purge

# 克隆sglang和MindSpeed仓库（注意：review建议使用--depth 1优化镜像大小）
RUN ARCH=$(uname -m) && \
    echo "[LOG INFO] Detected architecture: $ARCH" && \
    # 为x86_64平台设置额外的pip索引以支持CPU版本
    if [ "$ARCH" = "x86_64" ]; then \
        pip config set global.extra-index-url "https://download.pytorch.org/whl/cpu/"; \
    fi && \
    # 克隆libs，指定版本标签
    git clone https://github.com/sgl-project/sglang.git && cd sglang && git checkout v0.5.10 &&
    cd .. && \
    git clone https://gitcode.com/Ascend/MindSpeed.git && \
    cd MindSpeed && git checkout 2.3.0_core_r0.12.1 && cd ..
```

评论区精华

gemini-code-assist[bot] 在 review 中提出优化和安全建议：

- 优化建议：指出克隆 sglang 和 MindSpeed 仓库时应使用 --depth 1 进行浅克隆，以减少镜像大小和构建时间。评论原话：'Cloning large repositories... without --depth 1 downloads the entire git history, which significantly increases the image size and build time.'
- 安全建议：批评使用 --no-check-certificate 标志绕过 SSL 验证，有安全风险。评论原话：'Using --no-check-certificate with wget is a security risk as it bypasses SSL/TLS verification.' 这些建议在 PR 合并前未显示是否被采纳，可能作为未来改进点。
- Dockerfile 优化：使用浅克隆减少镜像大小 (performance): 建议未在 PR 中明确采纳，可能作为未来优化点。
- 安全风险：移除 --no-check-certificate 标志 (security): 建议未在 PR 中明确采纳，安全风险仍需关注。

风险与影响

- 风险：安全风险：Dockerfile 中使用 --no-check-certificate 标志下载 sgl-kernel-npu，可能易受中间人攻击，影响镜像完整性。性能风险：未使用浅克隆可能导致 Docker 镜像过大，增加存储和传输开销。兼容性风险：新版本库（如 SGLang v0.5.10、MindSpeed 2.3.0_core_r0.12.1）可能与现有 verl 代码或训练脚本存在不兼容问题，需测试验证。
- 影响：对用户：使用 Ascend NPU 的研究者可以获得最新 SGLang 环境，简化部署流程并支持新特性。对系统：CI 工作流更新确保新镜像自动构建和发布，提升部署自动化水平。对团队：增强 NPU 基础设施支持，与近期 NPU 相关 PR（如 Docker 和 CI 更新）保持一致，推动硬件生态完善。
- 风险标记：安全风险 :SSL 验证绕过，镜像大小优化不足

关联脉络

- PR #5991 [fsdp] feat: qwen3.5 add npu docker file: 同为 NPU Docker 基础设施更新，涉及类似 Dockerfile 和 CI 工作流变更。
- PR #5935 [ci] chore: Add veomni npu ci test: 涉及 NPU CI 测试工作流增强，与本 PR 的 CI 更新有技术关联。