

PR #6532 完整报告

verl-project/verl

[doc] refactor: optimize ascend doc

合并时间: 2026-05-29 15:51

原文链接: <http://prhub.com.cn/verl-project/verl/pull/6532>

执行摘要

- 一句话: 优化 Ascend NPU 文档与安装脚本, 升级依赖版本
- 推荐动作: 推荐阅读 `install_guidance.rst` 的重构思路以及一键安装脚本的集成方式, 可作为以后多后端文档的模板。关注 SGLang 路线从源码编译到 release whl 的转变, 这可能提升安装稳定性。

功能与动机

消除 Ascend 文档中陈旧信息, 统一教程结构与格式; 升级关键软件包 (vLLM 0.18.0、SGLang 0.5.10、torch 2.8.0) 以跟进社区进展; 提供更清晰的安装流程, 降低用户上手门槛。

实现拆解

1. 重构安装指南 `docs/ascend_tutorial/get_start/install_guidance.rst`, 将“后端拆分说明”替换为“框架后端支持说明”, 精简后端列表并增加目录导航和 Docker 镜像使用指引。
2. 新增 vLLM 路线一键安装脚本 `scripts/install_vllm_mcore_npu.sh`, 整合 pip 包安装与 Megatron/MindSpeed 部署流程; 同步修改 `scripts/install_sglang_mcore_npu.sh`, 升级 SGLang 至 v0.5.10、torch 至 2.8.0, 改用 release whl 安装 sgl-kernel-npu, 移除源码编译与冗余注释。
3. 更新性能调优、模型开发、FAQ 等子文档, 修复 Markdown 链接语法在 `.rst` 文件中的错误; 统一内部文档相对路径引用。
4. 根据 review 反馈修正安装脚本 `install_sglang_mcore_npu.sh` 中的 `cd ..` 残留和 trailing `&&` 语法问题。

关键文件:

- `scripts/install_vllm_mcore_npu.sh` (模块 安装脚本; 类别 infra; 类型 core-logic) : 新增的 vLLM 路线一键安装脚本, 整合 vLLM-Ascend 0.18.0、Megatron/MindSpeed 安装流程, 是此 PR 的核心产出之一。
- `scripts/install_sglang_mcore_npu.sh` (模块 安装脚本; 类别 infra; 类型 core-logic) : 更新 SGLang 安装脚本, 升级框架版本并改用 release whl 安装 sgl-kernel-npu, 大幅简化安装流程。Review 中指出的脚本错误也在此文件。
- `docs/ascend_tutorial/get_start/install_guidance.rst` (模块 Ascend 教程; 类别 docs; 类型 documentation) : 重构后的 Ascend 安装指南, 更新框架版本、补充 Docker 镜像指导, 格式调整最大, 是文档部分的核心变更。

关键符号：未识别

关键源码片段

scripts/install_vllm_mcore_npu.sh

新增的 vLLM 路线一键安装脚本，整合 vLLM-Ascend 0.18.0、Megatron/MindSpeed 安装流程，是此 PR 的核心产出之一。

```
#!/bin/bash
set -ex

USE_MEGATRON=${USE_MEGATRON:-1}

echo "1. install basic packages"
pip uninstall -y triton triton-ascend
# 安装与 vLLM-Ascend 0.18.0 对应的 torch 组件
pip install torchvision==0.24.0
pip install torchaudio==2.9.0
pip install triton-ascend==3.2.1 --extra-index-url https://triton-ascend.osinfra.cn/pypi/simple/ --
trusted-host triton-ascend.osinfra.cn
pip install "transformers==5.3.0"
pip install setuptools-scm

echo "2. install vllm & vllm-ascend"
git clone --depth 1 --branch v0.18.0 https://github.com/vllm-project/vllm.git
cd vllm
VLLM_TARGET_DEVICE=empty pip install -v -e .
cd ..
git clone -b releases/v0.18.0 https://github.com/vllm-project/vllm-ascend.git
cd vllm-ascend
git submodule update --init --recursive
pip install -v -e . --no-build-isolation --extra-index-url https://triton-ascend.osinfra.cn/pypi/simple/
--trusted-host triton-ascend.osinfra.cn
cd ..

if [ $USE_MEGATRON -eq 1 ]; then
    echo "3. install Megatron & MindSpeed"
    git clone https://gitcode.com/Ascend/MindSpeed.git
    cd MindSpeed && git checkout core_r0.16.0 && cd ..
    git clone --depth 1 --branch core_r0.16.0 https://github.com/NVIDIA/Megatron-LM.git
    pip install -e Megatron-LM
    pip install -e MindSpeed
    pip install mbridge
fi

echo "4. install verl"
cd verl
pip install -r requirements-npu.txt --extra-index-url https://triton-ascend.osinfra.cn/pypi/simple/ --
trusted-host triton-ascend.osinfra.cn
```

```

pip install -v -e .
cd recipe
git checkout main
cd ..

echo "5. May need to check other necessary packages"
pip install transformers==5.3.0 xgrammar==0.1.33
echo "Successfully installed all packages"

```

scripts/install_sglang_mcore_npu.sh

更新 SGLang 安装脚本，升级框架版本并改用 release whl 安装 sgl-kernel-npu，大幅简化安装流程。Review 中指出的脚本错误也在此文件。

```

#!/bin/bash
set -e
USE_MEGATRON=${USE_MEGATRON:-1}

echo "1. install SGLang from source"
git clone -b v0.5.10 https://github.com/sgl-project/sglang.git
cd sglang
# 设置 NPU 安装配置
mv python/pyproject.toml python/pyproject.toml.backup
mv python/pyproject_npu.toml python/pyproject.toml
pip install -e python[srt_npu]
cd ..

echo "2. install torch & torch_npu & other basic packages"
pip install torch==2.8.0 torch_npu==2.8.0.post2 torchvision==0.23.0 pyyaml
pip install pybind11 click==8.2.1 mbridge "numpy<2.0.0" cachetools

echo "3. install sgl-kernel-npu from release whl"
ARCH=$(uname -m)
wget --no-check-certificate https://github.com/sgl-project/sgl-kernel-npu/releases/download/2026.02.01/sgl-kernel-npu-2026.02.01-torch2.8.0-py311-cann8.5.0-a3-${ARCH}.zip
unzip sgl-kernel-npu*.zip
pip install torch_memory_saver*.whl
pip install sgl_kernel_npu*.whl
pip install deep_ep*.whl
# 创建符号链接使 deep_ep 可导入
cd "$(pip show deep-ep | grep -E '^Location:' | awk '{print $2}')" && ln -s deep_ep/deep_ep_cpp*.so . && cd -
python -c "import deep_ep; print(deep_ep.__path__)"
cd ..

if [ $USE_MEGATRON -eq 1 ]; then
    echo "4. install Megatron & MindSpeed"
    git clone https://gitcode.com/Ascend/MindSpeed.git
    cd MindSpeed && git checkout core_r0.16.0 && cd ..
    git clone --depth 1 --branch core_r0.16.0 https://github.com/NVIDIA/Megatron-LM.git

```

```
pip install -e Megatron-LM
pip install -e MindSpeed
pip install mbridge
fi

echo "5. install verl"
# ... (后续步骤根据 review 修复, 移除了残留 cd .. 和 trailing &&)
```

评论区精华

Review 中 [gemini-code-assist\[bot\]](#) 指出了三个高优先级问题:

- `scripts/install_sglang_mcore_npu.sh` 第 28 行的 `cd ..` 是重构后残留的导航错误, 会破坏后续目录依赖的命令。
- 第 46 行的 `trailing &&` 在严格 POSIX shell 下可能导致语法错误。
- `docs/ascend_tutorial/get_start/install_guidance.rst` 中 CANN 社区 超链接缺少下划线 `_` 或 `__`, 影响 Sphinx 渲染。
- `docs/ascend_tutorial/dev_guide/performance/ascend_profiling_*.rst` 中使用了 Markdown 链接语法而非 reStructuredText 语法。

审核者 [wucong25](#) 最终批准。

- 移除残留 `cd ..` 和 `trailing &&` (correctness): 建议移除 `cd ..` 并清理命令链, 采用更清晰的书写方式。
- `.rst` 文件中使用了 Markdown 链接语法 (documentation): 应使用 `text <url>_` 格式。
- CANN 社区超链接缺少下划线 (documentation): 建议使用双下划线 `__` 匿名链接。

风险与影响

- 风险: 安装脚本中的路径和语法问题可能在新环境下导致安装失败 (已在 review 中修复); 文档中的 `.rst` 格式问题在修复前可能导致 Sphinx 构建警告。此外, 依赖版本升级 (如 torch 2.8.0、SGLang 0.5.10) 可能与其他组件存在兼容性风险, 需在实际环境中验证。
- 影响: 对所有使用 Ascend NPU 的用户: 安装指南结构化重写后更易阅读, 一键安装脚本降低了部署复杂度; SGLang 安装流程简化, 不再需要源码编译 `sgl-kernel-npu`。团队需确保新脚本在各 NPU 型号 (A2/A3) 上通过 CI 验证。
- 风险标记: 脚本路径残留风险, `.rst` 文档格式错误, 依赖版本兼容性

关联脉络

- PR #6291 [vllm, ascend] upgrade vllm-ascend to 0.18.0: 本 PR 将 Ascend 文档和安装脚本中的 vLLM-Ascend 升级至 0.18.0, 与 PR#6291 的版本更新一致。
- PR #6435 [docker] chore: upgrade sglang to 0.5.12: 本 PR 将 SGLang 升级至 0.5.10 (相近版本), 并简化了 `sgl-kernel-npu` 安装方式。
- PR #6374 [megatron] feat: ascend bump into megatron 016: 本 PR 安装脚本中 Megatron-LM 和 MindSpeed 版本升级至 `core_r0.16.0`, 与 PR#6374 一致。