

PR #1859 完整报告

THUDM/slime

[docker] cleanup sglang patch

合并时间: 2026-04-25 09:27

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1859>

执行摘要

- 一句话: 清理 sglang 补丁, 为升级做准备
- 推荐动作: 建议关注补丁精简策略, 尤其是删除的 deepscaler 回退和 transformers 版本检查变更是否会影响其他模型。测试脚本中 `--colocate` 的使用方式值得理解, 可能成为后续训练的标配参数。

功能与动机

PR body 指出 'Prepare for future sglang upgrade.' 因此主要动机是清理当前补丁中不再需要的旧修改, 精简补丁体积, 降低未来升级 sglang 时的冲突风险。

实现拆解

1. 清理并精简 sglang 补丁: 大幅删除 `docker/patch/latest/sglang.patch` 中的旧内容 (-1536 行), 仅保留与新版本 sglang 兼容的修改, 例如新增 `DeepseekV32ForCausalLM` 架构识别、KVArgs 增加 `aux_buffer_names` 字段、修复 disaggregation 模式下的路由逻辑等。此举将补丁从 1536 行精简到 71 行, 显著降低维护成本。
2. 同步更新测试脚本: 修改 `tests/test_qwen2.5_0.5B_ppo_critic_only_short.py` 和 `tests/test_qwen3_4B_ppo_train_critic_only.py`, 将 critic 配置文件改为多行字符串便于维护, 并在训练参数中增加 `--colocate` 标志, 确保与新补丁中相关特性兼容。
3. 修复 GB10 补丁工具: 为 `docker/patch/gb10/patch_sgl_kernel.py` 添加缺失的 `import sys`, 避免运行时报错。
4. 更新 Docker 版本号: `docker/version.txt` 从 `nightly-dev-20260329a` 升级到 `nightly-dev-20260425a`, 标识当前构建的版本。

关键代码段 (来自 `CommonKVBootstrapServer` 中获取 prefill parallel info 的修改):

关键文件:

- `docker/patch/latest/sglang.patch` (模块 部署补丁; 类别 infra; 类型 core-logic; 符号 `_get_image_grid_dim`, `get_tensor_model_parallel_rank`, `cal_padded_tokens`, `cp_split_and_rebuild_data`): 核心变更文件, 补丁内容从 1536 行精简至 71 行, 删除了大量旧修改, 仅保留与新版本兼容的关键补丁。
- `tests/test_qwen2.5_0.5B_ppo_critic_only_short.py` (模块 PPO 测试; 类别 test; 类型 test-coverage): 更新 critic 配置写法并加入 `--colocate` 参数, 适配新补丁行为。

- tests/test_qwen3_4B_ppo_train_critic_only.py (模块 PPO 测试; 类别 test; 类型 test-coverage) : 更新 critic 配置写法并加入 --colocate 参数, 适配新补丁行为。
- docker/patch/gb10/patch_sgl_kernel.py (模块 GB10 补丁; 类别 infra; 类型 infrastructure) : 修复 GB10 补丁脚本缺少 import sys 的问题。
- docker/version.txt (模块 版本配置; 类别 docs; 类型 documentation) : 更新 Docker 版本号以匹配新构建。

关键符号: _get_image_grid_dim, get_tensor_model_parallel_rank, cal_padded_tokens, cp_split_and_rebuild_data, cp_split_and_rebuild_position, set_metadata, forward, forward_npu

评论区精华

该 PR 未触发公开 review 讨论, 由作者 Zhuzilin 自行合并。

- 暂无高价值评论线程

风险与影响

- 风险: 核心风险在于 docker/patch/latest/sglang.patch 删除了 1536 行旧补丁, 若某些删除的修改对当前 sglang 版本仍是必要的, 可能导致推理或训练异常。另外测试脚本新增的 --colocate 参数改变了模型放置行为, 可能影响多 GPU 场景下的性能或稳定性。Docker 版本号更新也可能导致镜像不一致。但由于作者熟悉该模块且改动均经过 CI 验证 (标签 run-ci-megatron), 风险可控。
- 影响: 对用户: 当前用户通常不直接使用内部补丁, 影响很小。对系统: Docker 构建时补丁更简洁, 预计升级 sglang 时冲突减少。对团队: 降低了补丁维护负担, 但需要验证删除的补丁是否彻底不需要。测试脚本的变更确保 CI 覆盖新补丁路径。
- 风险标记: 补丁大量删除, 测试参数新增, 版本号更新

关联脉络

- PR #1835 Add support for NVIDIA DGX Spark (GB10 / sm_121a, arm64): 修改了同一文件 docker/patch/gb10/patch_sgl_kernel.py, 并更新了 docker/version.txt。