

PR #1884 完整报告

THUDM/slime

[docker] update megatron-bridge and add qwen3.6 tests

合并时间: 2026-04-30 17:57

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1884>

执行摘要

- 一句话: 更新 Megatron-Bridge 并新增 Qwen3.6 测试
- 推荐动作: 建议关注 Bridge 版本切换后的集成测试结果, 并验证 Qwen3.6 模型的训练收敛性。该 PR 的测试文件可作为模板, 供开发者快速编写新模型的 PD Mooncake 测试。

功能与动机

为适配新发布的 Qwen3.6-35B-A3B 模型, 需要添加对应的 CI 测试用例; 同时, Megatron-Bridge 的维护仓库已迁移, 因此更新 Docker 安装源以保证后续兼容性。

实现拆解

1. 更新 Docker 依赖 (docker/Dockerfile) : 将 torch_memory_saver 的 git commit 从 dc6876905830430b5054325fa4211ff302169c6b 改为 d64a639; 将 Megatron-Bridge 的安装源从 https://github.com/fzyzcjy/Megatron-Bridge.git@dev_rl 改为 <https://github.com/radixark/Megatron-Bridge.git@bridge> 并使用 --no-deps 标志。
2. 新增 Qwen3.6 测试 (tests/test_qwen3.6_35B_A3B_pd_mooncake.py) : 将原 test_qwen3_30B_A3B_pd_mooncake.py 文件重命名为新模型名, 并调整 MODEL_NAME、MODEL_TYPE、TORCH_DIST_CKPT 等常量。将 tensor-model-parallel-size 从 4 改为 2 ; 在 sglang 参数字段中新增 --sglang-disaggregation-transfer-backend mooncake、--sglang-speculative-algorithm EAGLE 及相关推测解码参数 (num-steps、eagle-topk、num-draft-tokens、mamba-scheduler-strategy) 。
3. 增强 GLM4.7 测试 (tests/test_glm4.7_30B_A3B_pd_mooncake.py) : 在 sglang 参数字段中追加与 Qwen3.6 测试相同的 EAGLE 推测解码参数, 保持 CI 测试一致性。
4. 更新 CI 配置 (.github/workflows/pr-test.yml 及模板 .pr-test.yml.j2) : 将测试矩阵中原 test_qwen3_30B_A3B_pd_mooncake.py 替换为 test_qwen3.6_35B_A3B_pd_mooncake.py。
5. 同步文档: 更新中英文 CI 文档、版本号及 README 中的测试列表, 反映 CI 测试变更。
6. 微调 Rollout 引擎 (slime/ray/rollout.py) : 删除 start_engines 方法中一处多余的空行, 无逻辑影响。

关键文件:

- tests/test_qwen3.6_35B_A3B_pd_mooncake.py (模块 PD 测试; 类别 test; 类型 rename-or-move; 符号 MODEL_NAME, MODEL_TYPE, NUM_GPUS,

TORCH_DIST_CKPT)：核心变更文件，重命名并修改为新模型 Qwen3.6 的 PDMooncake 测试，引入 EAGLE 推测解码等新参数，是本次 PR 的主要目的。

- docker/Dockerfile (模块 Docker 构建；类别 infra；类型 infrastructure；符号 torch_memory_saver, Megatron-Bridge)：核心基础设施变更，更新了 Megatron-Bridge 和 torch_memory_saver 的安装来源与版本。
- tests/test_glm4.7_30B_A3B_pd_mooncake.py (模块 PD 测试；类别 test；类型 test-coverage；符号 sglang_args, execute)：为 GLM4.7 测试添加 EAGLE 推测解码参数，保持一致性和增强测试覆盖。
- .github/workflows/pr-test.yml (模块 CI 配置；类别 infra；类型 infrastructure)：CI 配置更新，将 Qwen3-30B 测试替换为 Qwen3.6 测试，确保新模型在 CI 中运行。
- slime/ray/rollout.py (模块 Rollout 引擎；类别 source；类型 core-logic)：仅删除一处空行，无逻辑实质变更，但属于核心文件。

关键符号：prepare, execute

关键源码片段

tests/test_glm4.7_30B_A3B_pd_mooncake.py

为 GLM4.7 测试添加 EAGLE 推测解码参数，保持一致性和增强测试覆盖。

```
# 在 sglang_args 中新增 EAGLE 推测解码四连参数
sglang_args = (
    "--rollout-num-gpus 8 "
    "--rollout-num-gpus-per-engine 4 "
    "--sglang-enable-dp-attention "
    "--sglang-dp-size 4 "
    "--sglang-enable-dp-lm-head "
    "--sglang-ep-size 4 "
    "--sglang-moe-dense-tp-size 1 "
    "--sglang-mem-fraction-static 0.45 "
    "--sglang-cuda-graph-max-bs 8 "
    "--sglang-max-running-requests 16 "
    "--sglang-disaggregation-transfer-backend mooncake "
    # 以下四行是本次新增的 EAGLE 推测解码参数
    "--sglang-speculative-algorithm EAGLE "
    "--sglang-speculative-num-steps 3 "
    "--sglang-speculative-eagle-topk 1 "
    "--sglang-speculative-num-draft-tokens 4 "
    "--sglang-watchdog-timeout 1200 "
    "--sglang-router-request-timeout-secs 1200 "
    "--sglang-enable-metrics "
    f"--sglang-config {sglang_config} "
)
```

评论区精华

本 PR 无 review 讨论，由作者直接合入。

- 暂无高价值评论线程

风险与影响

- 风险:

1. 依赖源变更风险: Megatron-Bridge 从 fzyzcjy 仓库切换到 radixark 仓库, 版本不一致可能导致集成问题, 需在真实训练场景中验证。
2. 新增测试不稳定风险: Qwen3.6 为新模型, 并行配置 (TP=2, EP=8, CP=2, DP=4) 可能与预期不符, EAGLE 推测解码参数可能影响 rollout 正确性与性能。
3. CI 覆盖缺失风险: 原 Qwen3-30B 测试被替换, 旧模型在 CI 中不再有专门的 PD Mooncake 测试, 若后续回归可能无法及时发现。
4. 文档不同步风险: 文档中修改了测试列表但未涉及其他配置说明, 可能对开发者造成困惑。
 - 影响: 影响范围: Docker 镜像构建 (Bridge 与 memory saver 版本)、CI 测试矩阵 (新增 Qwen3.6 测试替换 QWen3-30B)、GLM4.7 测试配置、中英文文档及版本号。影响程度: 中等。对用户无直接影响, 但后续 CI 将验证新模型, Bridge 更新可能影响训练稳定性。

- 风险标记: 依赖源变更, 新增测试不稳定, 并行参数调整

关联脉络

- PR #1867 [docker] upgrade megatron to 1dcf0dafa: 同为 Docker 依赖升级 (Megatron), 与本 PR 的 Bridge 更新属于同一演进线。
- PR #1874 [docker] upgrade slang to v0.5.10.post1: 同为 Docker 镜像更新与测试文件调整, 背景一致。