

PR #1975 完整报告

THUDM/slime

[release] bump to v0.3.0

合并时间: 2026-05-31 07:40

原文链接: <http://prhub.com.cn/THUDM/slime/pull/1975>

执行摘要

- 一句话: 版本升级至 v0.3.0, 重写构建脚本并更新补丁
- 推荐动作: 建议阅读 build_conda.sh 的重写逻辑和新增补丁的内容, 理解 CUDA 环境兼容性处理方法。对于涉及分布式训练的开发者, 重点关注 megatron.patch 中的 checkpoint 加载与 DDP 修改。该 PR 值得精读, 尤其是基础设施和部署相关的部分。

功能与动机

版本升级至 v0.3.0, 旨在发布新版本并同步更新底层依赖 (sglang、megatron) 的环境适配与补丁。同时重写构建脚本以简化环境安装、适配 CUDA 12.9。

实现拆解

1. 版本号更新: 修改 setup.py 中的 version 字段从 "0.2.4" 为 "0.3.0"。
2. 构建脚本重写: 重写 build_conda.sh, 调整环境变量引用为版本号 (SGLANG_VERSION、MEGATRON_COMMIT、PATCH_VERSION), 修复 nodefaults 问题, 优化 CUDA 12.9 安装, 添加 Rust 依赖, 解决 cu13 与 cu12 的包冲突 (force-reinstall torch 和 sglang-kernel, 然后修复 nvidia 共享库)。
3. 新增补丁文件: 在 docker/patch/v0.5.12.post1/ 下新增 sglang.patch 和 megatron.patch, 包含 checkpoint 加载 (weights_only=False, 允许 shape 不匹配)、分布式数据并行 (disable_grad_buffers_cpu_backup、disable_param_sync)、Flash Attention 缩放因子修复等。
4. CI 配置调整: 修改 .github/workflows/conda-ci.yml 增加构建与测试流程; 微调 pr-test.yml 和 pr-test.yml.j2。
5. 测试与断言调整: 修改 tests/test_qwen3_0.6B_parallel_check.py, 简化并行组合循环 (使用 parallel_sizes = [1,2,4]); 在 slime/backends/megatron_utils/data.py 中放宽 CI 断言的 log_prob 和 entropy 范围。

关键文件:

- docker/patch/v0.5.12.post1/megatron.patch (模块 补丁文件; 类别 infra; 类型 infrastructure; 符号 _validate_global_shapes, TEGroupedLinear, _encode_extra_state, _postprocess): 新增 1007 行补丁, 对 megatron 核心库进行多处修复, 包括 checkpoint 加载、分布式数据并行、Flash Attention 等, 是本次版本升级的关键基础设施变更。

- `build_conda.sh` (模块 构建脚本; 类别 `other`; 类型 `core-logic`) : 核心构建脚本, 被大幅重写以支持 CUDA 12.9、修复 conda 配置冲突、整合环境变量同步, 是版本升级的主要工程改动。
- `docker/patch/v0.5.12.post1/sglang.patch` (模块 补丁文件; 类别 `infra`; 类型 `infrastructure`) : 新增 2850 行 `sglang` 补丁, 适配 `v0.5.12.post1` 版本, 是 `sglang` 集成的重要更新。
- `slime/backends/megatron_utils/data.py` (模块 后端核心; 类别 `source`; 类型 `core-logic`; 符号 `log_rollout_data`) : CI 中断言范围被放宽 (`log_probs` 和 `entropy`) , 这是为了适应实际训练中的浮动, 但可能掩盖异常。
- `tests/test_qwen3_0.6B_parallel_check.py` (模块 并行测试; 类别 `test`; 类型 `test-coverage`) : 测试并行组合逻辑被简化, 以支持更灵活的 CI 执行, 减少硬件限制。
- `setup.py` (模块 版本配置; 类别 `source`; 类型 `core-logic`) : 版本号从 0.2.4 升级到 0.3.0 , 是版本发布的核心标识。
- `.github/workflows/conda-ci.yml` (模块 CI 配置; 类别 `infra`; 类型 `infrastructure`) : 新增 conda 构建的 CI 配置, 确保新构建脚本在 PR 中通过测试。

关键符号: `_validate_global_shapes`, `TEGroupedLinear`, `_encode_extra_state`, `_postprocess`, `_maintain_float32_expert_bias`, `_checkpointed_forward`, `checkpoint_handler`, `log_rollout_data`

关键源码片段

`slime/backends/megatron_utils/data.py`

CI 中断言范围被放宽 (`log_probs` 和 `entropy`) , 这是为了适应实际训练中的浮动, 但可能掩盖异常。

```
if "rollout/log_probs" in reduced_log_dict:
    # 放宽下限从 -0.5 到 -1, 以应对实际训练中的数值波动
    assert -1 < reduced_log_dict["rollout/log_probs"] < 0
if "rollout/entropy" in reduced_log_dict:
    # 放宽上限从 0.5 到 1, 以覆盖更多的正常情况
    assert 0 < reduced_log_dict["rollout/entropy"] < 1
```

评论区精华

该 PR 没有公开的 review 评论, 合并者为提交者本人, 属于快速发布流程。

- 暂无高价值评论线程

风险与影响

- 风险: 新增的补丁文件 (特别是 `megatron.patch` 中修改了 `checkpoint` 加载和分布式数据并行流程) 可能影响分布式训练的稳定性 and 兼容性。`build_conda.sh` 重写后改变了环境安装方式, 可能对已有部署环境的手动升级步骤有影响。CI 断言的放宽 (`log_probs` 从 `-0.5~0` 改为 `-1~0`, `entropy` 从 `0~0.5` 改为 `0~1`) 降低了检测精度, 可能掩盖潜在数值异常。

- 影响：对用户：升级至 v0.3.0 需要重新构建 conda 环境或使用更新后的 docker 镜像。对系统：构建流程发生重大变化，CI 增加了 conda 构建测试；补丁文件的变更影响底层训练引擎。对团队：需要维护新的补丁版本和构建脚本，确保与官方同步。
- 风险标记：补丁兼容性风险，构建流程重写，断言范围放宽，CUDA 版本冲突处理

关联脉络

- PR #1983 [docker] fix GLM4.7 Flash in sglang v0.5.12: 与本 PR 同属 sglang 补丁更新，涉及相同补丁目录和版本适配。
- PR #1991 [ci] Add e2e test for delta weight update: 同属 v0.3.0 发布周期内的 CI 增强，与本 PR 的 CI 配置修改互补。
- PR #1990 [ci] fix kl check on R3: 与本 PR 均修改了 `slime/backends/megatron_utils/data.py` 中的 CI 检查逻辑。