

PR #32294 完整报告

sgl-project/sglang

[UT][NPU] add NPU attention unit tests for ascend_backend and ascend_dsv4_backend

合并时间: 2026-07-26 19:59

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32294>

执行摘要

- 一句话: 为 NPU Ascend attention 后端添加 CPU 单元测试
- 推荐动作: 建议直接合并。此 PR 填补了 NPU 后端的单元测试空白, 且采用了的 mock/stub 模式值得学习。测试结构清晰, 覆盖全面, CI 配置合理。

功能与动机

NPU attention backend 模块 (ascend_backend.py、ascend_dsv4_backend.py) 缺乏对纯逻辑辅助函数和元数据设置函数的单元测试覆盖。现有测试只覆盖了部分函数且未遵循源文件命名惯例。此 PR 旨在通过 CPU 可运行的单测填补这一空白, 确保重构时不会意外破坏这些核心逻辑。

实现拆解

1. 创建 test_npu_ascend_backend.py: 在顶层通过 sys.modules.setdefault 模拟 torch_npu、sgl_kernel_npu 等 NPU 专有模块, 然后从源文件导入 _expand_dsa_sparse_indices、_reshape_kv_for_fia_nz、ForwardMetadata 等符号。测试类覆盖了二维 / 三维输入下的下标展开行为、KV 重排输出形状与数据一致性、数据类默认值和字段设置。
2. 创建 test_npu_ascend_dsv4_backend.py: 以相同手段模拟 NPU 依赖, 额外构造 deepseek_v2 模块 stub 以提供 _is_hip 常量, 并实现 faithful 的 _per_step_draft_out_cache_loc 函数作为 eagle_utils 的替代。测试涵盖 _walsh_hadamard_matrix 的形状、正交性、缓存行为、dtype 固定逻辑, 以及 _apply_hadamard、_get_kv_indices 等函数的边缘情况。
3. 更新 CI 配置: 在 .github/workflows/pr-test-npu.yml 中新增 stage-a-unit-test-npu 阶段, 使用 q10b 容器, 通过 python3 run_suite.py --hw npu --suite stage-a-unit-test-npu 执行, 限时 15 分钟。该阶段依赖 check-changes 输出, 仅当 main_package 内容变更时触发。
4. 注册测试套件: 在 test/run_suite.py 中将 stage-a-unit-test-npu 加入 BASE_SUITES 的 NPU 列表, 并调整 run_a_suite 中的文件遍历逻辑排除 run_tests.py (避免与已有分散测试冲突)。

关键文件:

- test/registered/unit/npu/attention/test_npu_ascend_backend.py (模块 Ascend 后端; 类别 test; 类型 test-coverage; 符号 TestExpandDsaSparseIndices, TestReshapeKvForFiaNz, _expand_dsa_sparse_indices, _reshape_kv_for_fia_nz) : 核心测试文件之一, 覆盖 _expand_dsa_sparse_indices、_reshape_kv_for_fia_nz 和 ForwardMetadata 的全面测试, 共 767 行。
- test/registered/unit/npu/attention/test_npu_ascend_dsv4_backend.py (模块 DSv4 后端; 类别 test; 类型 test-coverage; 符号 TestWalshHadamardMatrix, _walsh_hadamard_matrix, _apply_hadamard, _get_kv_indices) : 核心测试文件之二, 覆盖 _walsh_hadamard_matrix、_apply_hadamard 等函数, 以及模拟 speculative 依赖的巧妙设计。
- .github/workflows/pr-test-npu.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : CI 配置文件, 新增 stage-a-unit-test-npu 阶段以自动运行 NPU 单元测试。
- test/run_suite.py (模块 测试套件; 类别 test; 类型 test-coverage) : 测试调度脚本, 注册新的 stage-a-unit-test-npu 套件并修复文件遍历排除规则。

关键符号: _expand_dsa_sparse_indices, _reshape_kv_for_fia_nz, _walsh_hadamard_matrix, _apply_hadamard, _get_kv_indices, _overlap_transform

评论区精华

此 PR 未产生实质性讨论。sglang-npu-bot 在 CI 通过后自动批准并合并, 表明符合 NPU 团队的测试规范要求。

- 无讨论 (other): 无需进一步讨论。

风险与影响

- 风险: 低风险。变更均限于测试文件与 CI 配置, 不触及任何推理或模型代码。主要潜在问题:
 - 测试依赖大量 mock, 可能无法覆盖 NPU 真实硬件上的行为 (如算子精度差异), 但被测函数均为纯张量形状变换与数学运算, CPU 结果应可等价。
 - 新增 CI 阶段 stage-a-unit-test-npu 会增加约 15 分钟的整体 CI 时间, 但仅影响 NPU 相关 PR, 且与已有阶段并行执行。
 - 已在 pr-test-npu.yml 中为单元测试设置了 15 分钟超时, 防止挂起。
 - 影响: 用户影响: 无。开发者影响: 对 NPU 后端辅助函数的任何修改将立即被这些单元测试捕获, 降低回归风险。测试文件中的 mock 模式可作为后续 NPU 单元测试的参考模板。CI 影响: 新增一个快速单元测试阶段, 增强对 NPU 模块的持续验证。
- 风险标记: 低风险

关联脉络

- PR #32427 add fill_draft_extend_prepare_buffers_native for NPU: 同属 NPU 后端优化, 修改了部分 speculative 相关模块, 本 PR 的测试覆盖了相似依赖路径。

- PR #31189 [NPU]Add Ascend transfer version compatibility.: 近期 NPU 后端 bugfix, 本 PR 补充的单元测试可在未来防止类似回归。