

PR #32126 完整报告

sgl-project/sglang

Seed the GDN CuteDSL correctness test inputs to fix flakiness

合并时间: 2026-07-23 12:30

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32126>

执行摘要

- 一句话: 修复 GDN CuteDSL 测试因随机种子未设置导致的偶发失败
- 推荐动作: 建议合并。这是一个小但有效的测试稳定性修复, 通过简单明确的种子机制解决了 CI 中的间歇性失败。值得注意的实践是: 当测试断言阈值接近误差分布尾部时, 应主动固定随机种子以保证可重复性。

功能与动机

`test_gdn_chunk_cutedsl_correctness` 测试在 CI 中偶发失败, 失败案例 `[state_dtype0-1]` 的错误均值为 $6.31e-4$, 超过了 $6e-4$ 的阈值, 而四次通过运行均使用同一内核代码。根本原因是测试输入来自未设置种子的 `torch.randn/randint`, 导致极端随机输入触发断言失败。

实现拆解

1. 固定随机种子: 在 `test_gdn_chunk_cutedsl_correctness` 函数开头添加 `torch.manual_seed(num_seqs)`, 使用每用例的 `num_seqs` 作为种子, 确保同一参数组合下输入数据确定, 消除随机性造成的波动。
2. 添加解释注释: 在种子设置前添加注释, 说明阈值 $6e-4$ 接近观测到的误差分布尾部, 无种子输入会使测试变得不稳定。
3. 对齐同类测试: 同文件中的另一个测试 `test_gdn_chunk_cutedsl_pool_mode_matches_de_nse` 已使用了种子 `seed=11`, 本修改使 `correctness` 测试与其保持一致。

关键文件:

- `test/registered/attention/test_gdn_prefill_cutedsl.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 唯一修改的文件, 在测试函数开头添加了 `torch.manual_seed(num_seqs)` 以固定随机种子, 消除因随机输入导致的偶发测试失败。

关键符号: `test_gdn_chunk_cutedsl_correctness`

关键源码片段

`test/registered/attention/test_gdn_prefill_cutedsl.py`

唯一修改的文件, 在测试函数开头添加了 `torch.manual_seed(num_seqs)` 以固定随机种子, 消除因随机输入导致的偶发测试失败。

```
def test_gdn_chunk_cutedsl_correctness(num_seqs: int, state_dtype: torch.dtype):
```

```
# Fixed per-case seed: the mean-error assertions sit close to the observed
# error distribution (state_error.mean() threshold 6e-4 vs ~6.3e-4 seen on
# unlucky draws in CI), so unseeded inputs make the test flaky.
torch.manual_seed(num_seqs) # 使用 num_seqs 作为种子，确保同一参数组合下输入确定
# 后续的 randn/randint 调用将基于固定种子生成，消除随机性导致的断言失败
seq_lens = torch.randint(1, 130, (num_seqs,), dtype=torch.int32)
# ... 其余代码不变
```

评论区精华

无，PR 无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：无技术风险。仅修改测试文件，添加了确定性种子，不影响内核逻辑、性能或任何生产代码。测试失败源已消除，但种子固定后可能覆盖的随机输入空间变小，不过原本的阈值已接近误差分布尾部，固定种子后仍能有效检测回归。
- 影响：影响范围极小，仅影响 GDN CuteDSL 正确性测试的稳定性。CI 中该测试不再偶发失败，减少误报，提升开发效率。对其他用户无影响。
- 风险标记：测试稳定性修复

关联脉络

- PR #30169 Unify GDN chunk_cutedsl and chunk_triton code paths: PR body 提到内核代码自 #30169 后未变化，本 PR 的测试稳定性问题是独立于该 PR 的遗留问题。
- PR #30822 GDN cutedsl: fix A_log usage: PR body 引用该 PR 的 CI 日志作为偶发失败的证据。