

PR #49388 完整报告

vllm-project/vllm

[CI] stabilize GDN prefill CuTeDSL test

合并时间: 2026-07-23 00:03

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49388>

执行摘要

- 一句话: 稳定 GDN prefill CuTeDSL 测试
- 推荐动作: 值得合并, 作为最小且有效的 CI 稳定性修复。建议后续考虑引入更多 seed 或参数化变异来补偿随机性覆盖率。

功能与动机

该 PR 源自 #49143, 旨在解决 `test_gdn_prefill_cutedsl.py` 测试因随机 RNG 输入导致结果不稳定、偶尔超容差失败的问题。PR description 指出 'different RNG inputs may shift the tolerance a bit. In the original test, I intentionally put the tolerance quite tight'。

实现拆解

1. 添加确定性 RNG: 在测试函数 `test_gdn_chunk_cutedsl_correctness` 开头创建两个固定种子的 `torch.Generator` 实例: `rng_cpu (seed=1234)` 用于 CPU 上的 `seq_lens` 生成, `rng (seed=2345)` 用于 GPU 上的所有随机张量生成。
2. 将生成器传递给所有随机操作: 将 `generator=rng` 参数添加到所有 `torch.randn`、`torch.rand`、`torch.empty.uniform_` 调用中, 确保每次运行产生相同的随机数据序列。
3. 保持其他逻辑不变: 测试的核心逻辑、参数化范围和参考实现对比逻辑不作更改。

关键文件:

- `tests/kernels/mamba/test_gdn_prefill_cutedsl.py` (模块测试; 类别 test; 类型 test-coverage): 唯一变更文件, 核心修改为使 GDN prefill CuTeDSL 测试的随机输入确定化, 避免 CI 不稳定。

关键符号: `test_gdn_chunk_cutedsl_correctness`

关键源码片段

`tests/kernels/mamba/test_gdn_prefill_cutedsl.py`

唯一变更文件, 核心修改为使 GDN prefill CuTeDSL 测试的随机输入确定化, 避免 CI 不稳定。

```
# 测试函数开始处创建确定性 RNG
rng_cpu = torch.Generator("cpu").manual_seed(1234) # 固定 CPU 种子
rng = torch.Generator("cuda").manual_seed(2345) # 固定 CUDA 种子
```

```
seq_lens = torch.randint(1, 130, (num_seqs,), dtype=torch.int32, generator=rng_cpu)
cu_seqlens = torch.zeros(num_seqs + 1, device="cuda", dtype=torch.int32)
cu_seqlens[1:] = seq_lens.to(device="cuda").cumsum(0)
total_tokens = int(cu_seqlens[-1].item())
# ... 所有 torch.randn/generator 调用均传入 generator=rng, 确保每次运行数据一致
q = torch.randn(1, total_tokens, num_k_heads, head_k_dim, device="cuda", dtype=dtype,
generator=rng)
k = torch.randn_like(q, generator=rng)
v = torch.randn(1, total_tokens, num_v_heads, head_v_dim, device="cuda", dtype=dtype,
generator=rng)
# ... 更多随机张量及统一 _ 调用均使用 generator=rng
```

评论区精华

无人工 reviewers 讨论。claude[bot] 自动评论指出 PR 来自 fork，自动审查已禁用。
WoosukKwon 已批准该 PR。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅影响单个测试文件，且仅涉及随机数种子设置和 generator 传递，不影响任何生产代码。但需注意，如果测试关注的数值稳定性依赖于随机性覆盖多种场景，固定 seed 可能减少潜在 bug 的暴露机会。
- 影响：正向影响显著：消除 CI 中该测试的非确定性失败，提高 CI 稳定性。对用户无直接影响。
- 风险标记：测试随机性覆盖降低

关联脉络

- PR #49143 original PR from which this was extracted: 该 PR 来自 #49143 的一部分，用于隔离和简化 CI 稳定性修复。