

PR #49912 完整报告

vllm-project/vllm

[CI] Initialize DeepEP FP8 test weights

合并时间: 2026-07-28 05:29

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49912>

执行摘要

- 一句话: 修复 DeepEP FP8 测试权重依赖随机内容
- 推荐动作: 无需精读, 可直接合并。这是一个 CI 稳定性修复, 设计简单, 值得类似测试编写者借鉴 (避免 `torch.empty` 依赖全量内容)。

功能与动机

DeepEP FP8 测试的权重通过 `torch.empty` 初始化, 导致测试依赖于分配器未写入的随机内容; 同时大数值可能超出 W8A8 的固定容忍度, 造成测试不稳定。PR 引用关联 Issue #46758 确保修改兼容后续优化。

实现拆解

修改 `tests/kernels/moe/test_deepep_moe.py` 中的 `make_weights` 函数:

1. FP8 分支中将 `torch.empty((e, 2*n, k), device="cuda", dtype=torch.float16)` 替换为 `torch.randn((e, 2*n, k), device=current_platform.device_type, dtype=torch.float16).div_(100)`, 确保权重数值有限且小。
2. 同理处理第二个权重 `w2`。
3. 设备类型改为 `current_platform.device_type` 以适配 ROCm 平台, 不再硬编码 `"cuda"`。

关键文件:

- `tests/kernels/moe/test_deepep_moe.py` (模块测试; 类别 `test`; 类型 `test-coverage`; 符号 `make_weights`): 所有变更均在此文件中, 修复了 FP8 测试权重的初始化方式, 消除测试随机失败源。

关键符号: `make_weights`

关键源码片段

`tests/kernels/moe/test_deepep_moe.py`

所有变更均在此文件中, 修复了 FP8 测试权重的初始化方式, 消除测试随机失败源。

```
# ... (before actual changes)
# per-out-channel weight quantization
assert dtype == current_platform.fp8_dtype()
# Keep FP8 inputs finite and bounded. torch.empty made this test depend on
```

```
# allocator contents, while larger values exceed its fixed W8A8 tolerance.
w1 = torch.randn(
    (e, 2 * n, k), device=current_platform.device_type, dtype=torch.float16
).div_(100) # 标准差为 0.01, 量化后不会超出容忍范围
w2 = torch.randn(
    (e, k, n), device=current_platform.device_type, dtype=torch.float16
).div_(100)
# ... (rest of quantization and return)
```

评论区精华

无 reviewer 讨论或争议。tjtanaa 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅修改测试数据生成方式，不涉及生产代码逻辑。新随机值范围有限（均值为 0，标准差为 0.01），量化后不易溢出。需要确保 `current_platform.device_type` 在 ROCm 环境中正确返回 "hip" 等值。
- 影响：影响范围仅限于 `test_deepep_moe.py` 中的 FP8 测试用例。预期提高 DeepEP FP8 测试的稳定性和可重复性，消除因分配器内容不可控导致的随机失败。对其他测试或无 DeepEP 的环境无影响。
- 风险标记：测试数据生成改动

关联脉络

- PR #46758 [ROCm][CI TG] refactor and fix deepep_moe test group: 本 PR 引用的基础，修复了相同的测试文件，本 PR 在其基础上进一步改进权重初始化。