PR #40629 完整报告

vllm-project/vllm

[Bugfix][CI] Fix wrong residual shape in TestFusedAddRMSNorm.example_inputs that causes flaky test

合并时间：2026-04-25 04:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40629

执行摘要

- 一句话：修复 RMSNorm 测试中残差张量形状不匹配问题
- 推荐动作：值得精读，这是一个典型的内存越界导致 flaky 测试的案例。PR 作者提供了详细的复现脚本和根因分析，对理解 CUDA 内存分配和测试编写有借鉴意义。建议阅读 PR body 中的复现脚本以加深理解。

功能与动机

关联 issue #40622 报告了 tests/compile/passes/test_functionalization.py::TestFusedAddRMSNorm 测试的不稳定性。根本原因是 example_inputs 生成的残差张量形状为 (128, 16)，但 fused_add_rms_norm 内核期望读取 (128, 32) 元素（因为 RMSNorm 的 hidden dim 是 intermediate_size=32）。内核越界读取时，若两次运行中越界区域恰好相似，测试可能通过，导致间歇性失败。

实现拆解

1. 修改残差张量形状：在 tests/compile/passes/test_functionalization.py 中，将 example_inputs 方法中的 residual 张量从 torch.randn((batch_size * seq_len, hidden_size)) 改为 torch.randn((batch_size * seq_len, self.intermediate_size))，使其与 RMSNorm 期望的维度一致。
2. 移除 hidden_size 参数并改用实例属性：同时将 hidden_states 的生成从 torch.randn((batch_size * seq_len, hidden_size)) 改为 torch.randn((batch_size * seq_len, self.hidden_size))，并移除方法签名中的 hidden_size 参数。这使得方法能自动适应模型实例的实际配置，避免未来因参数不一致导致的形状错误。
3. 相关文件：仅修改了 tests/compile/passes/test_functionalization.py，变更量 +3/-3。

关键文件：

- tests/compile/passes/test_functionalization.py（模块 测试；类别 test；类型 test-coverage；符号 example_inputs）：包含核心修复：修改 example_inputs 方法，使残差张量形状与模型 intermediate_size 对齐，并移除遮蔽参数的 hidden_size 参数。

关键符号：example_inputs

关键源码片段

tests/compile/passes/test_functionalization.py

包含核心修复：修改 example_inputs 方法，使残差张量形状与模型 intermediate_size 对齐，并移除遮蔽参数的 hidden_size 参数。

```
# 修复前：residual 形状为 (batch*seq_len, hidden_size=16)，但内核期望 32 列
# 修复后：residual 形状与 self.intermediate_size 对齐

def example_inputs(self, batch_size=8, seq_len=16):
    # hidden_states 使用实例属性 self.hidden_size，而非硬编码参数
    hidden_states = torch.randn((batch_size * seq_len, self.hidden_size))
    # residual 使用 self.intermediate_size，确保与 RMSNorm 期望维度一致
    residual = torch.randn((batch_size * seq_len, self.intermediate_size))
    return (hidden_states, residual)
```

评论区精华

gemini-code-assist[bot] 在 review 中指出 hidden_size 参数会遮蔽 self.hidden_size，可能导致在非默认 hidden_size 模型配置下出现形状不匹配，建议使用实例属性。zhangj1an 接受了该建议并在第二次提交中进行了修复。没有其他重大争议。

- 移除隐藏的 hidden_size 参数以使用实例属性 (design): 作者采纳建议，在第二次提交中移除了 hidden_size 参数，改为使用 self.hidden_size 和 self.intermediate_size。

风险与影响

- 风险：风险极低：变更仅涉及测试辅助方法，不改变任何生产代码。修改已确保输入张量形状与模型层兼容，消除了越界读取的风险。未引入新的依赖或性能影响。
- 影响：直接影响是修复了 TestFusedAddRMSNorm 的 flaky 测试，提高 CI 稳定性。对系统其他部分无影响。团队可减少因该测试反复失败而排查的时间成本。
- 风险标记：暂无

关联脉络

- PR #40622 [CI Failure]: PyTorch Compilation Passes Unit Tests: 该 issue 报告了具体的 flaky 测试，是本 PR 直接修复的问题。