PR #43695 完整报告

vllm-project/vllm

Fix test_aot_compile for torch 2.12

合并时间：2026-05-27 11:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43695

执行摘要

- 一句话：修复 PyTorch 2.12 下 AOT 编译测试失败
- 推荐动作：此 PR 是适配 PyTorch 2.12 的必要修复，变更简单直接，值得合并。建议后续统一检查其他使用 VLLM_USE_MEGA_AOT_ARTIFACT 的测试点，确保 torch 版本升级后的兼容性。

功能与动机

PyTorch 2.12 引入回归：VLLM_USE_MEGA_AOT_ARTIFACT=1 需要 VLLM_USE_STANDALONE_COMPILE=1，但测试中 *env2* 仅设置 VLLM_USE_STANDALONE_COMPILE=0，未设置 VLLM_USE_MEGA_AOT_ARTIFACT，导致断言失败。关联 Issue #184550 详细描述了该错误。

实现拆解

1. 修改 tests/compile/test_aot_compile.py 中的 test_standalone_compile_correctness 函数。
2. 将 env2 字典从单键 {"VLLM_USE_STANDALONE_COMPILE": "0"} 扩展为双键 {"VLLM_USE_STANDALONE_COMPILE": "0", "VLLM_USE_MEGA_AOT_ARTIFACT": "0"}。
3. 这样，当 PyTorch 2.12 默认启用 VLLM_USE_MEGA_AOT_ARTIFACT 时，测试会显式关闭它，避免因缺少 VLLM_USE_STANDALONE_COMPILE 而触发断言。
4. 无其他文件修改，变更极小且聚焦。

关键文件：

- tests/compile/test_aot_compile.py（模块 AOT 编译；类别 test；类型 test-coverage）：唯一修改文件，修复了 PyTorch 2.12 下 AOT 编译测试的失败。

关键符号：test_standalone_compile_correctness

关键源码片段

tests/compile/test_aot_compile.py

唯一修改文件，修复了 PyTorch 2.12 下 AOT 编译测试的失败。

```
# tests/compile/test_aot_compile.py 中 test_standalone_compile_correctness 的变更
# 原 env2 字典：env2={"VLLM_USE_STANDALONE_COMPILE": "0"}
```

```
# 新 env2 字典:
compare_two_settings(
    "facebook/opt-125m",
    common_args,
    common_args,
    env1={"VLLM_USE_STANDALONE_COMPILE": "1"},
    env2={
        "VLLM_USE_STANDALONE_COMPILE": "0",
        "VLLM_USE_MEGA_AOT_ARTIFACT": "0", # 显式关闭 mega artifact，兼容 torch 2.12
        默认开启的行为
    },
)
```

评论区精华

讨论中，Harry-Chen 询问谁实际设置了 VLLM_USE_MEGA_AOT_ARTIFACT；zou3519 回应不确定，但如果 PyTorch 2.12 默认开启，可以使用关闭选项。最终补丁选择了显式关闭该变量。

- VLLM_USE_MEGA_AOT_ARTIFACT 默认值问题 (question): PR 通过显式设置 VLLM_USE_MEGA_AOT_ARTIFACT=0 来解决 torch 2.12 的默认行为变化。

风险与影响

- 风险：风险极低。变更仅涉及测试配置，不改变产品代码。测试覆盖了 OPT-125m 模型的 standalone compile 与非 standalone compile 的正确性对比，添加的环境变量显式关闭了可能因 torch 版本变化而引入的默认行为，确保测试稳定性。
- 影响：直接影响：修复 tests/compile/test_aot_compile.py::test_standalone_compile_correctness 在 PyTorch 2.12 下的失败。间接影响：其他依赖于相同环境变量默认值的测试可能也会受益于类似的显式设置，但本 PR 未对其余测试修改。不影响用户或系统。
- 风险标记：依赖 PyTorch 版本行为

关联脉络

- PR #42848 涉及 AOT 编译的早期 PR: 关联 Issue #184550 提到该测试最初在 PR #42848 的提交中失败，本 PR 是对后续 torch 2.12 兼容的修复。