

PR #51074 完整报告

vllm-project/vllm

[CI] Prune PyTorch Fullgraph Test

合并时间: 2026-08-06 03:27

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51074>

执行摘要

- 一句话: 精简 PyTorch Fullgraph 测试矩阵与 CI 步骤
- 推荐动作: 该 PR 属于常规测试维护, 值得快速浏览 diff 了解测试裁剪决策; 不建议精读。关注点在于合并后的超时风险与覆盖缩减是否可接受。

功能与动机

PR 标题与提交信息表明意图是裁剪 (Prune) Fullgraph 测试。原配置中 smoke test 与 fullgraph 分离导致重复运行, 且测试矩阵包含大量低价值组合 (GPTQ、w8a8、旧模型)。body 未提供详细说明, 但从 diff 看是为了降低 CI 开销并集中维护。

实现拆解

1. 精简正确性测试矩阵 (tests/compile/fullgraph/test_basic_correctness.py): 删除 GPTQ 量化用例与 BAAI/bge-base-en-v1.5 的 ROCm 跳过用例; 将 MoE 模型从 ibm/PowerMoE-3b (tp=2) 替换为 ibm-granite/granite-3.0-1b-a400m-instruct (tp=1); 将原先按三种编译模式分别与 NONE 配对比较的 inductor 循环重构为一次性比较全部四种模式, compare_all_settings 调用次数从 3 次降为 1 次。
2. 精简全图测试模型列表 (tests/compile/fullgraph/test_full_graph.py): 从 models_list 删除 Llama-3.2-1B-Instruct、w8a8/w8w8 量化模型和 GPTQ 模型, 仅保留 opt-125m、FP8 dynamic 模型及 gptq_marlin/AWQ (按平台支持); FP8 模型由 neuralmagic 的仓库替换为 RedHatAI 的同一模型镜像; 由于不再覆盖 int8 模型, 移除了 Blackwell 上的 w8a8/w8w8 跳过分支, custom_ops 融合测试的关键字从 ['FP8-dynamic', 'quantized.w8a8'] 缩减为 ['FP8-dynamic']。
3. 合并 CI 步骤 (.buildkite/test_areas/pytorch.yaml): 将“PyTorch Fullgraph Smoke Test”重命名为“PyTorch Fullgraph Test”, key 改为 pytorch-fullgraph-test, 命令由排除 test_full_graph.py 改为仅排除 test_full_cudagraph.py, 使 test_full_graph.py 纳入统一扫描; 删除独立于 h200_18gb 的“PyTorch Fullgraph”步骤, 全部用例集中在 h200_35gb (超时 90 分钟) 运行。
4. 配套说明: 无新增测试或文档; 依赖原有 tests/compile/utils.py 中的 compare_all_settings 与 create_new_process_for_each_test 机制。

关键文件:

- tests/compile/fullgraph/test_basic_correctness.py (模块 正确性测试; 类别 test; 类型 test-coverage; 符号 test_compile_correctness, TestSetting) : 核心测试矩阵调整, 删除了量化与 embedding 用例, 重构编译模式比较逻辑, 直接影响正确性测试的覆盖范围与运行效率。
- tests/compile/fullgraph/test_full_graph.py (模块 全图测试; 类别 test; 类型 test-coverage; 符号 models_list, test_full_graph, test_custom_compile_config) : 模型列表精简与 FP8 模型替换, 移除 Blackwell skip 逻辑, 缩小了 fullgraph 测试的模型覆盖。
- .buildkite/test_areas/pytorch.yaml (模块 测试配置; 类别 config; 类型 configuration) : 合并 smoke test 与 fullgraph 步骤, 删除独立 fullgraph step, 统一测试入口。

关键符号: test_compile_correctness, models_list, test_full_graph, test_custom_compile_config

关键源码片段

tests/compile/fullgraph/test_basic_correctness.py

核心测试矩阵调整, 删除了量化与 embedding 用例, 重构编译模式比较逻辑, 直接影响正确性测试的覆盖范围与运行效率。

```
# tests/compile/fullgraph/test_basic_correctness.py (head 版本核心逻辑)
final_args = [
    *model_args,
    '-pp', str(pp_size),
    '-tp', str(tp_size),
    '-cc.cudagraph_mode=none',
    f'--attention-backend={attn_backend}',
]

all_args: list[list[str]] = []
all_envs: list[dict[str, str] | None] = []

# Test all compilation modes with inductor backend
for mode in [
    CompilationMode.NONE,
    CompilationMode.STOCK_TORCH_COMPILE,
    CompilationMode.DYNAMO_TRACE_ONCE,
    CompilationMode.VLLM_COMPILE,
]:
    all_args.append(final_args + [f'-cc.mode={mode.name}', '-cc.backend=inductor'])
    all_envs.append({})
# inductor 会更改输出, 因此比较时使用 generate_close 方法
compare_all_settings(
    model,
    all_args,
    all_envs,
    method=method if method != 'generate' else 'generate_close',
    force_v1_runner=True,
)
```

```
all_envs.clear()
all_args.clear()

# Test all compilation modes with eager backend
for mode in [
    CompilationMode.NONE,
    CompilationMode.STOCK_TORCH_COMPILE,
    CompilationMode.DYNAMO_TRACE_ONCE,
    CompilationMode.VLLM_COMPILE,
]:
    all_args.append(final_args + [f'-cc.mode={mode.name}', '-cc.backend=eager'])
    all_envs.append({})
compare_all_settings(model, all_args, all_envs, method=method, force_v1_runner=True)
```

评论区精华

PR 无实质审查讨论；claude[bot] 因来自 fork 自动跳过审查，仅提示维护者可触发人工审查。作者 mgoin 触发了 Buildkite CI (#82527) 作为验证。

- 自动化审查被禁用 (other): 无后续实质讨论，作者直接触发 CI 并合并。

风险与影响

- 风险：主要风险：一是测试覆盖缩减可能遗漏量化路径（GPTQ、w8a8）与 Blackwell int8 支持问题；二是合并后的 fullgraph step 在 h200_35gb 上仍设 90 分钟超时，若 test_full_graph.py 用例变多可能超时；三是模型源替换（RedHatAI/Llama-3.2-1B-Instruct-FP8-dynamic）可用性与 license 依赖外部服务。
- 影响：对用户无直接功能影响；对 CI 系统减少约一个 GPU 测试步骤和运行时长；对开发团队，fullgraph 测试的维护入口从两个 step 变为一个，降低了配置复杂度，但测试覆盖面有所收窄。
- 风险标记：测试覆盖缩减，合并 CI 步骤超时风险，外部模型依赖替换

关联脉络

- PR #51177 [CI Bug] Fix Chunked prefill is required for mamba cache mode 'align'.: 同为 CI 测试修复，体现测试矩阵维护的持续工作。
- PR #51153 [Bugfix] Enable chunked prefill for qwen3.5-0.8B ppl test: 调整测试配置修复 CI 失败，与本次精简测试目标一致。