

PR #36443 完整报告

sgl-project/sglang

[CPU] Fix rotary_embedding_cpu fake for in-place layouts

合并时间: 2026-08-27 09:48

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36443>

执行摘要

- 一句话: 修复 CPU 图捕获中 rotary 嵌入的布局错误
- 推荐动作: 此 PR 是一个小而重要的 bugfix, 值得精读以理解 fake 内核与真实内核的重要性。维护者应关注是否有测试覆盖此场景, 但本次未能添加测试, 因此建议后续补充 CPU 图捕获的回归测试, 以覆盖不同布局的 rotary embedding。

功能与动机

在 CPU 上启用 `--enable-torch-compile` 时, 对于任何带有融合 QKV 投影的模型, 图捕获失败, 错误信息为: `assert_size_stride(buf6, (32, 4096), (4096, 1), 'torch.ops.sgl_kernel.rotary_embedding_cpu.default')`, 期望 stride 为 4096 但实际为 6144。原因是 fake 内核未正确模拟真实内核的原地行为, 导致编译器基于错误的元数据生成不兼容的布局。该回归由 PR #22498 ([CPU] Add support for Gemma4 on Xeon) 引入, 它用一个无条件的 fake 替换了原先布局相关的 fake。

实现拆解

1. 定位问题: 在 `python/sglang/srt/model_executor/cpu_graph_runner.py` 中找到 `rotary_embedding_cpu` 的 fake 内核实现, 发现其无条件返回 `torch.empty_like(query)`, `torch.empty_like(key)`。
2. 分析真实内核逻辑: 在 `python/sglang/kernels/aot/csrc/cpu/rope.cpp` 中, 真实内核根据 `input_dim != 3` 决定是原地旋转还是分配新张量, 但 fake 未反映这一点。
3. 修改 fake 内核: 在 `@register_cpu_compile_fake("rotary_embedding_cpu")` 函数中, 增加 `if query.ndim == 3`: 判断, 当输入为 3D 时返回 `torch.empty_like`, 否则返回原 `query` 和 `key`, 从而与真实内核保持一致。
4. 影响范围: 此变更仅影响 CPU 图捕获路径的元数据推断, 不会改变实际内核计算逻辑, 对模型输出和性能无影响。

关键文件:

- `python/sglang/srt/model_executor/cpu_graph_runner.py` (模块图捕获; 类别 source; 类型 data-contract): 这是唯一修改的文件, 包含 `rotary_embedding_cpu` 的 fake 内核, 此次修复的核心逻辑在此。

关键符号: `rotary_embedding_cpu fake`

关键源码片段

`python/sglang/srt/model_executor/cpu_graph_runner.py`

这是唯一修改的文件，包含 `rotary_embedding_cpu` 的 fake 内核，此次修复的核心逻辑在此。

```
# 关键源码片段: cpu_graph_runner.py 中 rotary_embedding_cpu 的 fake 内核
@register_cpu_compile_fake("rotary_embedding_cpu")
def _(positions, query, key, head_size, cos_sin_cache, is_neox):
    # 真实内核在 C++ 中根据 input_dim != 3 决定是否原地旋转
    # 因此这里也需要匹配: 3D 时分配新张量, 其他维度返回原张量
    if query.ndim == 3:
        return torch.empty_like(query), torch.empty_like(key)
    return query, key
```

评论区精华

review 中只有维护者 mingfeima 的批准，没有实质讨论。PR body 中作者 htzo 已详细解释了问题根因和修复方式，并注明是正确性变更。

- 暂无高价值评论线程

风险与影响

- 风险：该修复仅涉及 fake 内核的返回值，用于元数据推断，不影响运行时计算。但需要确保 `query.ndim` 的判断与真实内核的 `input_dim != 3` 完全一致，因为真实内核根据 `input_dim`（即 `query` 的维度数）判断。本次修复与真实内核逻辑一致，风险较低。潜在的兼容性风险是如果未来真实内核的布局处理变化，这里可能再次出现不一致，但当前无此风险。
- 影响：影响范围：CPU 上使用 `--enable-torch-compile` 且带有融合 QKV 投影的模型（如 Gemma4）的图捕获功能。修复后，图捕获将不再因 `stride` 断言失败而中断，从而允许编译优化正常进行。对不使用 `torch.compile` 的用户无影响。
- 风险标记：缺少测试覆盖，CPU 特定路径

关联脉络

- PR #22498 [CPU] Add support for Gemma4 on Xeon: 引入该回归的 PR，破坏了原先布局相关的 fake 实现。
- PR #36413 fix mmlu download issues: CI 失败（mmlu 下载问题）在该 PR 中修复，维护者 mingfeima 在 issue 评论中提及。