

PR #49394 完整报告

vllm-project/vllm

[XPU] Enable QK Norm + RoPE fusion pass on XPU

合并时间: 2026-07-27 11:22

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49394>

执行摘要

- 一句话: XPU 启用 QK Norm + RoPE 融合编译 pass
- 推荐动作: 值得精读并合并。设计清晰且安全: 将条件从 `is_cuda_alike` 扩展到 `is_cuda_alike or is_xpu`, 同时将两个 pass 一起移入, 保持了平台间的对称性。基准测试数据扎实, 证明了性能收益和精度无损。

功能与动机

XPU 平台此前缺少对 QK Norm + RoPE 融合 pass 的支持, 导致该优化被强制禁用。PR body 中的基准测试表明启用后吞吐量提升 11.7%, TTFT 降低约 10%, 且准确率与 eager 模式相当, 因此需要将融合 pass 扩展到 XPU。

实现拆解

1. 调整编译 pass 注册条件: 在 `vllm/compilation/passes/pass_manager.py` 中, 将 `QKNormRoPEFusionPass` 的导入从 `if current_platform.is_cuda_alike():` 块移动到 `if current_platform.is_cuda_alike() or current_platform.is_xpu():` 块, 同时将 `SplitCoalescingPass` 也一同移入, 使得这些 pass 在 XPU 上可用。
2. 移除 XPU 禁用列表条目: 在 `vllm/platforms/xpu.py` 的 `check_and_update_config` 方法中, 从 `fusion_passes_to_disable` 字典中删除了 `"enable_qk_norm_rope_fusion": "QK Norm + RoPE fusion"` 条目, 从而允许用户启用该 pass。
3. 新增 CI 测试任务: 在 `.buildkite/intel_jobs/misc_intel.yaml` 中添加名为 `Fusion Unit Tests` 的 CI 步骤, 自动运行 `tests/compile/passes/test_qk_norm_rope_fusion.py` 测试, 确保融合 pass 在 XPU 上正确性。

关键文件:

- `vllm/compilation/passes/pass_manager.py` (模块 编译 Pass; 类别 source; 类型 dependency-wiring): 核心变更加载逻辑: 将 `QKNormRoPEFusionPass` 和 `SplitCoalescingPass` 的导入从 `is_cuda_alike` 块扩展为 `is_cuda_alike or is_xpu`, 使 XPU 可注册这些 pass。
- `vllm/platforms/xpu.py` (模块 平台配置; 类别 source; 类型 core-logic): 移除禁用列表中的 `enable_qk_norm_rope_fusion` 条目, 允许 XPU 用户启用该 pass。
- `.buildkite/intel_jobs/misc_intel.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 新增 CI 步骤 `Fusion Unit Tests`, 用于自动运行融合 pass 的单元测试, 确保变更质量。

关键符号：未识别

关键源码片段

vllm/compilation/passes/pass_manager.py

核心变更加载逻辑：将 QKNormRoPEFusionPass 和 SplitCoalescingPass 的导入从 `is_cuda_alike` 块扩展为 `is_cuda_alike or is_xpu`，使 XPU 可注册这些 pass。

vllm/compilation/passes/pass_manager.py (部分)

原代码：仅 is_cuda_alike 时导入

新代码：is_cuda_alike 或 is_xpu 时都导入

if current_platform.is_cuda_alike() or current_platform.is_xpu():

QK Norm + RoPE fusion pass —— 核心变更：从下面的 is_cuda_alike 块移到这里

from .fusion.qk_norm_rope_fusion import QKNormRoPEFusionPass

Sequence parallelism pass —— 保持原有位置不变

from .fusion.sequence_parallelism import SequenceParallelismPass

Split coalescing pass —— 同样从 is_cuda_alike 块移到这里

from .utility.split_coalescing import SplitCoalescingPass

下面的只剩下 NVIDIA 独有的 pass，XPU 不会走这里

if current_platform.is_cuda_alike():

from .fusion.act_quant_fusion import ActivationQuantFusionPass

... 其他 NVIDIA 专有 pass 保持不变

vllm/platforms/xpu.py

移除禁用列表中的 `enable_qk_norm_rope_fusion` 条目，允许 XPU 用户启用该 pass。

vllm/platforms/xpu.py (check_and_update_config 方法片段)

该字典列出 XPU 上尚不支持的融合 pass，会被强制禁用

fusion_passes_to_disable = {

"fuse_gemm_comms": "Async TP",

"fuse_allreduce_rms": "AllReduce + RMSNorm fusion",

"fuse_attn_quant": "Attention + quant fusion",

"fuse_act_padding": "Activation + padding fusion",

"fuse_rope_kvcache": "RoPE + KV cache fusion",

"fuse_rope_kvcache_cat_mla": "RoPE + KV cache + MLA fusion",

删除以下行 —— enable_qk_norm_rope_fusion 现在已支持

"enable_qk_norm_rope_fusion": "QK Norm + RoPE fusion",

}

评论区精华

无实质技术讨论。审核人 jikunshang 仅评论了 [cc @Yejing-Lai](#)，随后批准；claude[bot] 自动回复因来自 fork 无法执行审查。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及条件导入和配置列表调整，核心融合 pass 逻辑本身未改动。CI 中新增的单元测试提供了基本覆盖，但缺少集成 / 端到端测试验证实际模型推理准确率。XPU 上其他融合 pass（如 ActivationQuantFusionPass、RMSNormQuantFusionPass）的注册保持不变。
- 影响：影响范围较小，仅针对 Intel XPU 平台。启用后用户可通过 `--compilation-config '{"pass_config":{"enable_qk_norm_rope_fusion":true}}'` 获得性能提升（吞吐量 +11.7%，TTFT-10%）。对 NVIDIA/AMD 平台无影响，因导入条件仍然包含它们。新增 CI 任务会增加少许 XPU 测试资源消耗。
- 风险标记：缺少集成测试覆盖，仅依赖单元测试

关联脉络

- PR #46877 [Core][Distributed] Add process-checkpoint lifecycle hooks for communicators (starting with Flashinfer): 同属 compilation/ 模块，涉及 pass 管理架构变更，但无直接功能关联。
- PR #49134 [Bugfix] Reject contradictory custom-op directives: 也涉及编译配置变更，但影响不同平台。