PR #43963 完整报告

vllm-project/vllm

[XPU] Enable rms_norm/act quant fusions

合并时间：2026-06-03 00:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43963

执行摘要

- 一句话：XPU 启用 norm/act 量化融合
- 推荐动作：该 PR 值得合并，但建议作者补充测试用例验证 XPU 上融合 pass 的正确性和性能。

功能与动机

在 XPU 平台上启用 norm 和 act 量化融合，以提升推理效率。之前在 vllm/platforms/xpu.py 的 check_and_update_config 方法中，fuse_norm_quant 和 fuse_act_quant 被列入了禁用列表，导致这两个优化在 XPU 上无法生效。

实现拆解

1. 移除禁用配置：在 vllm/platforms/xpu.py 的 fusion_passes_to_disable 字典中，删除了 "fuse_norm_quant" 和 "fuse_act_quant" 两条记录。同时，将原有的无条件循环禁用改为仅在 compilation_config.mode != CompilationMode.NONE 时打印警告并禁用，避免在编译模式为 NONE 时产生不必要的警告。
2. 添加条件导入：在 vllm/compilation/passes/pass_manager.py 中，增加了一个条件分支 if current_platform.is_xpu(): 来导入 ActivationQuantFusionPass 和 RMSNormQuantFusionPass，使得这些 pass 在 XPU 平台下可用。
3. 通用化设备类型：在 rms_quant_fusion.py 和 act_quant_fusion.py 中，将辅助函数（如 empty_bf16、empty_quant）和设备创建时的 device="cuda" 改为 device=current_platform.device_type，确保这些融合操作在不同平台（CUDA 和 XPU）上都能正确分配张量。

关键文件：

- vllm/platforms/xpu.py（模块 平台配置；类别 source；类型 dependency-wiring；符号 check_and_update_config）：移除 norm/act 量化融合的禁用配置，并调整警告触发条件。
- vllm/compilation/passes/pass_manager.py（模块 编译管理；类别 source；类型 entrypoint）：为 XPU 平台添加 ActivationQuantFusionPass 和 RMSNormQuantFusionPass 的条件导入。
- vllm/compilation/passes/fusion/rms_quant_fusion.py（模块 编译融合；类别 source；类型 core-logic；符号 empty_bf16, empty_fp32, empty_i32, empty_i64）：将辅助函数中的硬编码 device="cuda" 替换为 current_platform.device_type。

- vllm/compilation/passes/fusion/act_quant_fusion.py（模块 编译融合；类别 source；类型 core-logic；符号 ActivationQuantPattern.empty_quant）：将 empty_quant 方法中的硬编码 device="cuda" 替换为 current_platform.device_type。

关键符号：check_and_update_config, empty_bf16, empty_fp32, empty_i32, empty_i64, empty_quant

关键源码片段

vllm/platforms/xpu.py

移除 norm/act 量化融合的禁用配置，并调整警告触发条件。

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

# 移除了 "fuse_norm_quant" 和 "fuse_act_quant" 条目
fusion_passes_to_disable = {
    "enable_sp": "Sequence parallelism",
    "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",
}
# 仅在编译模式非 NONE 时打印警告并禁用
if compilation_config.mode != CompilationMode.NONE:
    for flag, feature_name in fusion_passes_to_disable.items():
        if getattr(pass_config, flag):
            logger.warning(
                "Feature %r is not yet supported on XPU and will be disabled.",
                feature_name,
            )
            setattr(pass_config, flag, False)
```

vllm/compilation/passes/fusion/rms_quant_fusion.py

将辅助函数中的硬编码 device="cuda" 替换为 current_platform.device_type。

```
# vllm/compilation/passes/fusion/rms_quant_fusion.py 设备通用化片段

def empty_bf16(*args: Any, **kwargs: Any) -> torch.Tensor:
    # 使用 current_platform.device_type 代替硬编码的 "cuda"
    return torch.empty(
        *args, **kwargs, dtype=torch.bfloat16, device=current_platform.device_type
    )

def empty_fp32(*args: Any, **kwargs: Any) -> torch.Tensor:
    return torch.empty(
        *args, **kwargs, dtype=torch.float32, device=current_platform.device_type
    )

def empty_i32(*args: Any, **kwargs: Any) -> torch.Tensor:
```

```
    return torch.empty(
        *args, **kwargs, dtype=torch.int32, device=current_platform.device_type
    )

def empty_i64(*args: Any, **kwargs: Any) -> torch.Tensor:
    return torch.empty(
        *args, **kwargs, dtype=torch.int64, device=current_platform.device_type
    )
```

vllm/compilation/passes/fusion/act_quant_fusion.py

将 empty_quant 方法中的硬编码 device="cuda" 替换为 current_platform.device_type。

```
# vllm/compilation/passes/fusion/act_quant_fusion.py ActivationQuantPattern.empty_quant 方法

def empty_quant(self, *args: Any, **kwargs: Any) -> torch.Tensor:
    kwargs = {
        "dtype": self.quant_dtype,
        "device": current_platform.device_type, # 替换 "cuda"
        **kwargs,
    }
    return torch.empty(*args, **kwargs)
```

评论区精华

此 PR 未见 review 讨论或评论。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险：将 device='cuda' 改为 device=current_platform.device_type 可能引入问题，若 current_platform.device_type 返回的值不是 CUDA 支持的设备（如 xpu），在 CUDA 上也可能使用此分支，但当前代码仅在 is_cuda_alike() 或 is_xpu() 下导入，因此风险较低。
  - 缺失测试：该 PR 未包含任何测试，无法保证融合 pass 在 XPU 上的正确性和性能收益。建议添加针对性的单元测试或集成测试。
- 影响：
  - 用户：使用 XPU 平台的用户将自动获得 norm 和 act 量化融合优化，可能带来推理性能提升。
  - 系统：仅影响 XPU 平台的编译流程，CUDA 等平台行为保持不变。
  - 团队：变更范围较小，易于回退。
  - 风险标记：缺少测试覆盖

关联脉络

- PR #44308 [ROCm] Fix AITER RMSNormQuantFusion for Kimi-Linear: 同样涉及 RMSNormQuantFusion，但属于 ROCm 平台修复。