

PR #39280 完整报告

vllm-project/vllm

[ROCm][Perf] Add Fused Shared Expert (FSE) support for Qwen3-Next

合并时间: 2026-05-09 03:38

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/39280>

执行摘要

- 一句话: 融合共享专家加速 Qwen3-Next 解码 16%-24%
- 推荐动作: 值得精读。展示了 ROCm 专有性能优化的一般模式: 利用环境变量开关、运行时版本检查、fallback 路径、以及通过独立路由器类避免污染通用代码。设计权衡 (门控融合位置、条件判断缓存) 值得参考。后续扩展 FSE 到其他模型时可以参考此框架。

功能与动机

Qwen3-Next 模型包含共享专家, 常规实现需要独立 SharedExpert MLP 前向, 带来额外 kernel 启动和显存读写。通过将共享专家融合到 MoE 内核, 可以减少 kernel 启动次数、提高 GPU 利用率, 显著提升解码吞吐。

实现拆解

实现包含以下主要步骤:

1. 门控权重融合: 在 `moe_runner.py` 的 `MoERunner._maybe_fuse_gate_weights()` 中, 将 `gate.weight` 与 `shared_expert_gate.weight` 沿第 0 维拼接为 `[num_experts + num_shared, hidden]` 的合并权重, 使得一次 `F.linear` 即可产生路由和共享专家的组合 logits。
2. AITER ops 扩展: 在 `_aiter_ops.py` 中为 `topk_softmax` 函数新增 `num_shared_experts` 和 `shared_expert_scoring_func` 参数; 类方法 `topk_softmax_supports_fused_sigmoid()` 动态检查 AITER 版本是否支持融合; `fuse_sigmoid_in_kernel()` 汇总全部条件 (共享专家启用、AITER 支持、元数据已初始化)。
3. 专用路由器: 新建 `aiter_shared_routed_fused_moe_router.py`, 包含 `AiterSharedRoutedFusedMoERouter`。在 `_compute_routing` 中根据 `fuse_sigmoid_in_kernel` 判断: 若支持则调用 7 参数 `topk_softmax` 同时计算 softmax 和 sigmoid; 否则回退到分离的 `topk_softmax + sigmoid + inject_shared_expert_weights` 合并。
4. 权重注入: `rocm_aiter_fused_moe.py` 新增 `inject_shared_expert_weights()`, 将路由 `topk` 结果拷贝到 AITER 预分配 `total_topk_weights/total_topk_ids` 缓冲区, 并将动态共享专家门控值写入对应槽位。
5. 模型适配: `qwen3_next.py` 中当 FSE 启用时 `self.shared_expert` 置为 `None` (避免模型侧重复计算); `load_weights` 中将共享专家权重重映射到第 `num_experts` 号的融合专家槽。

`qwen3_next_mtp.py` 相应调整。

6. 路由器工厂：`router_factory.py` 增加条件分支，当 `num_fused_shared_experts > 0` 且符合 ROCm AITER 条件时返回 `AiterSharedRoutedFusedMoERouter`。
7. 其他配套：`layer.py` 新增 `shared_expert_gate` 参数传递。

关键文件：

- `vllm/model_executor/layers/fused_moe/router/aiter_shared_routed_fused_moe_router.py`（模块 路由器；类别 source；类型 data-contract；符号 `AiterSharedRoutedFusedMoERouter`, `init`, `routing_method_type`, `_compute_routing`）：新增核心路由器，实现融合共享专家的路由逻辑，包含 AITER 融合和 fallback 两条路径。
- `vllm/_aiter_ops.py`（模块 AITER 层；类别 source；类型 dependency-wiring；符号 `topk_softmax_supports_fused_sigmoid`, `fuse_sigmoid_in_kernel`）：扩展 AITER `topk_softmax` 接口，添加融合 sigmoid 能力检查和运行时版本探测。
- `vllm/model_executor/layers/fused_moe/rocm_aiter_fused_moe.py`（模块 AITER MoE；类别 source；类型 data-contract；符号 `inject_shared_expert_weights`）：新增共享专家权重注入函数，确保 AITER 预分配缓冲区被正确写入。
- `vllm/model_executor/layers/fused_moe/runner/moe_runner.py`（模块 MoE 执行器；类别 source；类型 data-contract；符号 `_maybe_fuse_gate_weights`）：实现门控权重融合和条件判断，控制 FSE 是否生效。
- `vllm/model_executor/models/qwen3_next.py`（模块 模型；类别 source；类型 data-contract）：模型适配 FSE：禁用独立 `shared_expert`、权重重映射、环境变量检测。
- `vllm/model_executor/layers/fused_moe/router/router_factory.py`（模块 路由器工厂；类别 source；类型 data-contract）：路由器工厂新增条件分支来创建 `AiterSharedRoutedFusedMoERouter`。

关键符号：`AiterSharedRoutedFusedMoERouter.init`,
`AiterSharedRoutedFusedMoERouter._compute_routing`,
`rocm_aiter_ops.topk_softmax_supports_fused_sigmoid`,
`rocm_aiter_ops.fuse_sigmoid_in_kernel`, `inject_shared_expert_weights`,
`MoERunner._maybe_fuse_gate_weights`

评论区精华

- 架构选择：@tjtanaa 和 @bnellnm 建议采用 DeepSeekV2 的方式而非注入方式，最终 PR 调整为独立路由器类，符合路由器扩展性设计。
- 路由器独立性：@robertgshaw2-redhat 提出将 FSE 逻辑从现有路由器中抽离，避免了条件膨胀，最终采纳。
- 运行时检查：@dllehr-amd 建议将多次条件检查合并为单一缓存方法，PR 实现了 `fuse_sigmoid_in_kernel()` 统一入口。
- Bug 修复：@gemini-code-assist[bot] 指出 `n_shared_experts` 传递 `None` 导致 `TypeError` 以及 `SharedFusedMoE` 返回解包错误，PR 已修正。

- 门控融合位置: @robertgshaw2-redhat 和 @bnellnm 讨论应在模型层还是 runner 层融合, 当前采用 runner 层融合, 但保留模型层可选路径的讨论空间。
- 采用 DeepSeek 方法替代注入 (design): PR 最终重构为独立路由器类, 接近 DeepSeek 模式。
- 抽取 num_fused_shared_experts 为构造函数参数 (design):
num_fused_shared_experts 作为 BaseRouter.__init__ 参数, per-instance 决定。
- 运行时条件检查缓存 (performance): 通过单一方法组合静态和动态检查, 减少重复计算。
- 门控融合位置选择 (design): 当前采用 runner 内融合, 保留后续优化空间。
- n_shared_experts 默认为 None 导致 TypeError (correctness): 修正为 default 0。
- SharedFusedMoE 返回解包错误 (correctness): 修正解包逻辑。

风险与影响

- 风险:
 - 缺少测试覆盖: 本 PR 未新增单元测试, 依赖现有 MoE 测试套件, 无法保证所有分支 (fallback 路径、异常条件) 都覆盖。
 - 权重重映射依赖: qwen3_next.py 将共享专家权重重映射到固定索引 num_routed 的专家槽, 若模型配置变化 (如 num_experts 动态调整) 可能导致加载错误。
 - AITER 版本兼容性: 融合 sigmoid 功能依赖 aiter.topk_softmax 的 7 参数版本, 若无则该功能透明 fallback, 但用户需同时更新 AITER 库才能获得性能收益。
 - 环境变量安全: VLLM_ROCM_USE_AITER_FUSION_SHARED_EXPERTS 为新增变量, 默认关闭, 误开启但条件不满足时会调用 FusedTopKRouter 回退, 行为与未开启一致, 无崩溃风险。
- 影响:
 - 用户影响: 仅 ROCm 用户受益, 需设置 VLLM_ROCM_USE_AITER_FUSION_SHARED_EXPERTS=1 获得吞吐提升; 非 ROCm 和默认配置无变化。
 - 系统影响: 新增 AiterSharedRoutedFusedMoERouter 路由选择和 ROCM AITER ops 扩展, 增加代码路径。
 - 团队影响: 需跟踪 AITER 库版本变更以维护 topk_softmax 签名检查; 可能引入与 DeepSeek 等其他共享专家实现的兼容问题。
 - 风险标记: 缺少测试覆盖, AITER 版本兼容, 权重重映射依赖, 核心路径变更

关联脉络

- 暂无明显关联 PR