PR #37975 完整报告

vllm-project/vllm

[Model] Extract GatedDeltaNetAttention into shared layer for Qwen3Next and Qwen3.5

合并时间：2026-03-27 14:13

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37975

执行摘要

本 PR 将 GatedDeltaNetAttention 层从 qwen3_next.py 提取到新文件 gdn_linear_attn.py，统一 Qwen3Next 和 Qwen3.5 模型的实现，旨在支持 XPU/NPU 平台并减少代码重复。这是一个重要的重构，修复了潜在 bug 并涉及跨平台兼容性设计，值得工程师关注其参数化共享层实现。

功能与动机

为什么做：作者在 Issue 评论中解释，XPU 和 NPU 平台不支持 key-value 不连续操作，需要重写 GatedDeltaNetAttention 层。因此，通过重构提取共享层，以支持跨平台兼容性和代码复用。具体动机引用：'Since key-value in-contiguous are not supported in xpu and npu, the operators of the GatedDeltaNetAttention layer must be rewritten in xpu and npu'。

实现拆解

做了什么：按模块拆解关键改动：

- 新增模块：vllm/model_executor/layers/mamba/gdn_linear_attn.py – 包含参数化的 GatedDeltaNetAttention 类，处理 GQA 布局、LoRA 兼容性，并集成 FlashInfer 和 Triton 内核。
- 模型文件修改：
  - qwen3_5.py – 删除原 Qwen3_5GatedDeltaNet 类，导入并使用共享层，减少约 151 行代码。
  - qwen3_next.py – 删除原 Qwen3NextGatedDeltaNet 类，导入共享层，减少约 975 行代码。

关键代码逻辑示例（来自 review）：fix_query_key_value_ordering 方法修复后：

```
new_tensor_shape_ba = mixed_ba.size()[:-1] + (
    self.num_k_heads // self.tp_size,
    2 * self.num_v_heads // self.num_k_heads,
)
```

评论区精华

讨论了什么：提炼 review 中的核心交锋：

- Critical bug 修复：gemini-code-assist[bot] 发现 fix_query_key_value_ordering 中形状推导错误，作者及时修复。引用：'new_tensor_shape_ba is incorrectly derived from mixed_qkvz.size() instead of mixed_ba.size()'。
- 平台兼容性设计：jikunshang 讨论 forward_native 命名，认为应为 torch-native 实现以支持 CPU 平台。结论：暂保留，等待未来 IR PR。引用：'forward_native should be a torch-native impl'。
- 代码清理：claude[bot] 指出 gdn_in_proj 是死代码，作者确认已移除。
- 测试要求：ZJY0516 要求测试 qwen3.5、qwen3 next 和 lora，作者进行了测试并报告通过。

风险与影响

风险：具体技术风险包括：

- 回归风险：新共享层可能引入未覆盖 bug，需确保与旧实现行为一致，特别是在注意力计算路径。
- 平台兼容性：XPU/NPU 的键值不连续支持仍需验证，可能影响推理正确性或性能。
- 性能影响：Triton 内核使用可能不适用于 CPU 平台，需监控跨平台性能。

影响：

- 对用户：使用 Qwen3Next/Qwen3.5 模型时，底层实现更统一，可能提升跨平台支持，但用户无感知接口变化。
- 对系统：代码冗余减少约 2000 行，便于维护和扩展新模型变体。
- 对团队：需加强测试覆盖，确保模型准确性，并关注平台特异性集成。

关联脉络

与历史 PR 的关系：本 PR 是 Qwen 模型系列重构的一部分，与近期 PR 如 #38155（添加 Qwen3.5 模型测试）相关联，共同推进模型兼容性和测试完善。更大的功能演进方向是支持多平台（如 XPU/NPU）和代码模块化，减少模型特定实现。