PR #28293 完整报告

sgl-project/sglang

[NPU] Add NPU fallback for fused Triton gating kernels

合并时间：2026-06-16 11:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28293

执行摘要

- 一句话：NPU 回退 fused Triton 门控内核
- 推荐动作：此 PR 是标准的平台兼容性修复，设计简洁，风险低。值得精读以了解 NPU 硬件限制（coreDim）对 Triton kernel launch 的影响，以及如何通过条件分支优雅回退。

功能与动机

Issue #28272 报告了 Qwen3.6-35B-A3B 模型在前缀缓存场景下因 fused_gate_sigmoid_mul_add_kernel 的 coreDim 参数超出 NPU 限制 (65535) 导致 KernelLaunch 崩溃。PR #26924 引入了这些 fused Triton 内核以减少 kernel launch 开销，但未考虑 NPU 的硬件限制。此 PR 旨在提供 NPU 兼容的 fallback。

实现拆解

1. qwen2_moe.py (MoE 门控融合回退)：在 forward 方法中，use_fused_gate 条件新增 and not is_npu() 检查。当运行在 NPU 上时，use_fused_gate 为 False，从而跳过 fused_gate_sigmoid_mul_add 调用，转而执行 _forward_shared_experts(apply_gate=True) 分支，该分支使用原生 PyTorch 操作，语义等价。
2. qwen3_5.py (注意力门控融合回退)：在 self_attention 方法中，对 attn_output_gate 的处理增加 _is_npu 条件判断。NPU 上不使用 fused_sigmoid_mul，而是将 gate 进行 reshape（若为 3D）并计算 torch.sigmoid，再通过 attn_output.mul_() 原地更新 attention 输出。
3. 测试与文档：PR body 确认无精度影响，但未新增单元测试或更新文档。CI 已触发（包含 NPU 测试）。

关键文件：

- python/sglang/srt/models/qwen3_5.py（模块 Qwen3.5 模型；类别 source；类型 data-contract；符号 self_attention）：修改 self_attention 方法，为 NPU 提供 fused_sigmoid_mul 的回退路径，使用 in-place mul_ 与 sigmoid。
- python/sglang/srt/models/qwen2_moe.py（模块 Qwen2-MoE 模型；类别 source；类型 data-contract；符号 forward）：修改 use_fused_gate 条件，新增 and not is_npu()，阻止 NPU 使用 fused_gate_sigmoid_mul_add kernel。

关键符号：self_attention, forward

关键源码片段

python/sglang/srt/models/qwen3_5.py

修改 self_attention 方法，为 NPU 提供 fused_sigmoid_mul 的回退路径，使用 in-place mul_ 与 sigmoid。

```
# python/sglang/srt/models/qwen3_5.py
# self_attention 方法中的门控融合操作
if self.attn_output_gate:
    if not _is_npu:
        # CUDA/HIP: 使用 fused Triton kernel（减少 kernel launch 开销）
        attn_output = fused_sigmoid_mul(attn_output, gate, inplace=True)
    else:
        # NPU: fused kernel 的 grid size（num_tokens）可能超过 NPU coreDim 限制 (65535)
        # 使用原生 PyTorch 操作回退，语义等价
        gate_val = gate.reshape(gate.shape[0], -1) if gate.ndim == 3 else gate
        attn_output.mul_(torch.sigmoid(gate_val))
```

python/sglang/srt/models/qwen2_moe.py

修改 use_fused_gate 条件，新增 and not is_npu()，阻止 NPU 使用 fused_gate_sigmoid_mul_add kernel。

```
# python/sglang/srt/models/qwen2_moe.py
# forward 方法中的 fused gate 条件判断
use_fused_gate = (
    self.shared_expert_gate is not None
    and not use_intel_amx_backend(self.shared_expert_gate)
    and not is_npu() # NPU 不支持 fused_gate_sigmoid_mul_add kernel（grid size 超限）
)
```

评论区精华

无 reviewer 评论或讨论线程。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险低：回退路径使用原生 PyTorch 运算（sigmoid、mul_），逻辑与 PR #26924 前的代码一致。CUDA/HIP 路径无改动。
  - 性能轻微差异：NPU 上 in-place mul_ 避免了 fused 内核的 kernel launch 开销优化，但 reduction 了一次分配（inplace 操作节省了中间张量），综合性能无回归。
  - 兼容性：仅影响 NPU 后端；其他硬件不受影响。
  - 测试覆盖不足：未添加针对 NPU 的单元测试验证回退路径。
- 影响：

- 用户：NPU 用户可正常运行 Qwen2-MoE 和 Qwen3.5 系列模型，解决此前大 batch size 下 KernelLaunch 崩溃问题。性能无退化。
- 系统：改动轻微（仅 6 行），不影响 CUDA/HIP 后端。
- 团队：低维护成本。
- 风险标记：缺少测试覆盖

关联脉络

- PR #26924 [Introduce fused Triton kernels for gating]: 引入了 fused_sigmoid_mul 和 fused_gate_sigmoid_mul_add 内核，本 PR 为其提供 NPU 回退路径。