PR #23809 完整报告

sgl-project/sglang

fix act fun for xpu

合并时间：2026-05-21 14:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23809

执行摘要

- 一句话：为 XPU 添加 SiluMul 和 Rotary Embedding 前向路径
- 推荐动作：变更简单且逻辑清晰，建议合并。值得关注的设计决策是：将 XPU 的 silu_and_mul 导入与 HIP 共享同一条件分支，暗示了 sgl_kernel 对两者均支持的意图。

功能与动机

在运行 Z Image Turbo 模型时，需要为 XPU 平台启用 Silu_Mul 激活函数。PR body 明确说明："Enable Silu_Mul act function for xpu while running Z Image Turbo model"。

实现拆解

1. 激活函数层 (activation.py): 添加 _is_xpu 标志检测当前平台是否为 XPU，修改导入条件使得 XPU 平台从 sgl_kernel 导入 silu_and_mul（与 HIP 共享同一路径）。新增 SiluAndMul.forward_xpu 方法，其实现与 forward_cuda 完全一致：计算输出形状、分配空张量、调用 silu_and_mul(x, out)。
2. 旋转位置编码层 (rotary_embedding/base.py): 在 RotaryEmbedding 类中添加 forward_xpu 方法，直接委托给 forward_native，因为 XPU 平台没有专门的加速库，使用 PyTorch 原生实现。
3. 测试与配置: 未涉及测试文件或配置变更。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/activation.py（模块 激活函数；类别 source；类型 core-logic；符号 forward_xpu）：核心变更：添加 forward_xpu 方法，修改导入逻辑以支持 XPU 从 sgl_kernel 导入 silu_and_mul。
- python/sglang/multimodal_gen/runtime/layers/rotary_embedding/base.py（模块 位置编码；类别 source；类型 core-logic；符号 forward_xpu）：次要变更：添加 forward_xpu 方法，直接委托给 forward_native。

关键符号：forward_xpu

关键源码片段

python/sglang/multimodal_gen/runtime/layers/activation.py

核心变更：添加 forward_xpu 方法，修改导入逻辑以支持 XPU 从 sgl_kernel 导入 silu_and_mul。

```
# 检测通用平台适配层
_is_xpu = current_platform.is_xpu()

# XPU 与 HIP 共用 sgl_kernel 中的 silu_and_mul
if _is_cuda:
    from sglang.jit_kernel.activation import silu_and_mul
elif _is_hip or _is_xpu:
    from sgl_kernel import silu_and_mul

# 在 SiluAndMul 类中新增 XPU 前向方法
class SiluAndMul(CustomOp):
    # ... 其他方法不变 ...
    def forward_xpu(self, x: torch.Tensor) -> torch.Tensor:
        """XPU 专用 SiluMul 前向，使用 sgl_kernel 加速"""
        d = x.shape[-1] // 2
        output_shape = x.shape[:-1] + (d,)
        out = torch.empty(output_shape, dtype=x.dtype, device=x.device)
        silu_and_mul(x, out)
        return out
```

评论区精华

无 review 评论。审核者 mingfeima 批准了 PR，无需进一步讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：变更仅新增 XPU 专用分支，不影响现有 CUDA/HIP/NPU 路径。但缺少针对 XPU 的单元测试或集成测试覆盖，如果 sgl_kernel 中的 silu_and_mul 在 XPU 上未经过充分测试，可能出现数值错误或崩溃。建议后续补充 XPU 精度测试。
- 影响：影响范围较小：仅影响 XPU 平台上的扩散模型推理，使得此前运行失败的 Z Image Turbo 等模型能够在 XPU 上正常工作。对其他平台无影响。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR