PR #42822 完整报告

vllm-project/vllm

add gelu_tanh to xpu moe backend supported activations

合并时间：2026-05-29 14:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42822

执行摘要

- 一句话：XPU MoE 支持 gelu_tanh 激活函数
- 推荐动作：该 PR 为简单的兼容性修复，建议快速合并，但需确保关联的 xpu-kernels PR 已合入并更新依赖。

功能与动机

Gemma4-26B-A3B 模型在 Intel XPU 上运行时失败，报错 NotImplementedError: No Unquantized MoE backend supports the deployment configuration，原因是 Gemma4 的 MoE 激活函数已由 gelu 改为 gelu_tanh（见 PR#41574），而 XPU MoE 后端未支持该激活函数。当前暂用方案是手动修改 gemma4.py 将激活函数改回 gelu。

实现拆解

1. 在 vllm/model_executor/layers/fused_moe/experts/xpu_moe.py 的 _supports_activation 静态方法中，于支持列表中添加 MoEActivation.GELU_TANH 枚举值。
2. 该修改仅一行，使得 XPU 后端在 MoE 激活函数为 gelu_tanh 时也能正确匹配。
3. 配套的 XPU 内核支持已由关联 PR（vllm-xpu-kernels#354）实现，需同步升级 xpu kernels 版本。

关键文件：

- vllm/model_executor/layers/fused_moe/experts/xpu_moe.py（模块 模型层；类别 source；类型 data-contract；符号 _supports_activation）：在 XPU MoE 后端的激活函数支持列表中添加 MoEActivation.GELU_TANH，是修复的唯一变更文件。

关键符号：_supports_activation

关键源码片段

vllm/model_executor/layers/fused_moe/experts/xpu_moe.py

在 XPU MoE 后端的激活函数支持列表中添加 MoEActivation.GELU_TANH，是修复的唯一变更文件。

```
# vllm/model_executor/layers/fused_moe/experts/xpu_moe.py（变更后）
```

```
@staticmethod
def _supports_activation(activation: MoEActivation) -> bool:
    """判断 XPU MoE 后端是否支持给定的激活函数。
    当前支持 SILU、GELU、GELU_TANH (新增)、SWIGLUOAI、RELU2_NO_MUL。
    """
    return activation in [
        MoEActivation.SILU,
        MoEActivation.GELU,
        MoEActivation.GELU_TANH, # 新增：支持 Gemma4 等使用 gelu_tanh 的模型
        MoEActivation.SWIGLUOAI,
        MoEActivation.RELU2_NO_MUL,
    ]
```

评论区精华

无实质讨论。jikunshang 评论表示此 PR 需等待 vllm-xpu-kernels v0.1.9 发布并升级依赖后合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅为在支持列表中添加一个枚举值，不影响现有逻辑。但需注意，若内核层未正确实现 gelu_tanh 的计算，则可能产生错误结果。此风险由关联 PR 覆盖。
- 影响：影响范围：修复 Intel XPU 上 Gemma4 模型的运行问题，仅影响 XPU 平台使用 MoE 且激活函数为 gelu_tanh 的场景。
- 风险标记：暂无

关联脉络

- PR #41574 Change Gemma4 MoE activation to gelu_tanh: 该 PR 将 Gemma4 的 MoE 激活从 gelu 改为 gelu_tanh，直接导致本 PR 修复的问题。