PR #38770 完整报告

vllm-project/vllm

[CPU] Support gelu act in cpu_fused_moe

合并时间：2026-04-02 14:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38770

执行摘要

本 PR 为 CPU fused MoE 内核新增 gelu 激活函数支持，通过扩展 C++ 枚举、实现 gelu_and_mul 函数、更新 Python 映射和测试，提升了 CPU 后端对 gelu 激活模型的兼容性。同时，PR 意外包含了 attention KV split 的环境变量配置修改。核心讨论围绕 erf 计算的性能优化和函数命名准确性展开，建议关注向量化优化点以提升性能。

功能与动机

本 PR 的主要动机是扩展 CPU fused MoE 内核支持的激活函数类型，以覆盖需要 gelu 激活的模型。根据 PR body 描述，目的是“Add gelu act in cpu_fused_moe”，解决特定模型或用户需求。没有关联 Issue，表明这可能是一个内部改进或预研功能。

实现拆解

实现分为三个核心部分：

1. C++ 内核扩展：在 csrc/cpu/cpu_fused_moe.cpp 中，添加 FusedMOEAct::GeluAndMul 枚举值，并实现 gelu_and_mul 函数。该函数使用门控乘法模式，但 erf 计算当前为标量循环，存在优化空间。
2. Python 层映射：在 vllm/model_executor/layers/fused_moe/cpu_fused_moe.py 中，新增 _gelu_and_mul 函数（原名经 review 修正），使用 PyTorch 的 F.gelu(approximate="none") 实现，并注册到 _CPU_MOE_ACT_FN 字典。
3. 测试更新：修改 tests/kernels/moe/test_cpu_fused_moe.py，将 GELU 加入测试激活列表，确保新功能正确性。

额外修改包括在 vllm/envs.py 中添加 VLLM_CPU_ATTN_SPLIT_KV 环境变量，并在 vllm/v1/attention/backends/cpu_attn.py 中使用该变量控制 attention KV split，这与 gelu 功能无关，可能是一个独立的配置改进。

评论区精华

review 中 gemini-code-assist[bot] 提供了关键反馈：

> “erf 计算使用标量循环，应使用向量化 er() 方法提升性能。”“函数名 _gelu_and_mul_tanh 与实际非近似 gelu 不符，建议重命名为 _gelu_and_mul。”

这些讨论突出了性能优化和设计一致性的重要性。函数名在提交中已修正，但 erf 向量化优化状态未知，需后续验证。

风险与影响

- 性能风险：C++ 代码中 erf 计算未向量化，可能导致 CPU fused MoE 性能瓶颈，尤其在批量推理时。
- 兼容性风险：新增 gelu 激活需确保与现有模型和其他激活函数行为一致，避免回归问题。
- 配置风险：环境变量 VLLM_CPU_ATTN_SPLIT_KV 的引入可能影响 CPU attention 后端行为，默认值设置需谨慎。
- 影响范围：用户受益于更广泛的模型支持，但团队需维护新代码并监控性能指标。

关联脉络

从历史 PR 看，本 PR 与以下趋势相关：

- 性能融合内核演进：如 PR #32996（SiLU 融合量化）和 PR #34664（MXFP8 支持），显示 vllm 在持续优化融合内核以提升推理性能。
- CPU 后端增强：近期 PR 如 #38743（CPU 清理）和 #38750（ROCm 修复），表明团队在扩展和稳定 CPU 支持。
- v1 标签集中：多个 PR（如 #36836 RayExecutorV2）带有 v1 标签，指向 v1 版本的架构迭代，本 PR 作为一部分贡献了 CPU 功能完善。

整体上，本 PR 是 CPU 后端功能扩展的一个环节，未来可能带动更多激活函数或优化内核的开发。