PR #43677 完整报告

vllm-project/vllm

[Perf] Optimize Fp8BlockScaledMMLinearKernel input_scale tensor using new_empty()

合并时间：2026-05-27 10:55

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43677

执行摘要

- 一句话：FP8 块缩放矩阵乘中占位张量优化，吞吐提升 2%
- 推荐动作：值得合并的微小性能优化。设计思路（避免不必要的张量初始化）对其他类似占位符场景有参考价值。建议维护者关注后续是否会有子类误用 As 参数的风险，可考虑在 apply_block_scaled_mm 接口文档中强调 As 在未量化时可能为未初始化值。

功能与动机

作者在分析 DeepSeek V4 推理 Profiling 时发现 at::native::vectorized_elementwise_kernel 在 tensorrt_llm::kernels::fp8_blockscale_gemm::scale_1x128_kernel 之前被启动，原因是 apply_input_quant=False 时占位 input_scale 使用了 new_ones()。由于该占位张量实际未被下游 apply_block_scaled_mm 使用，替换为 new_empty() 可消除该不必要的核启动。

实现拆解

仅修改一处：在 vllm/model_executor/kernels/linear/scaled_mm/BlockScaledMMLinearKernel.py 的 apply_weights 方法内，将第 129 行 input_2d.new_ones(1) 替换为 input_2d.new_empty(1)。该改动位于 apply_input_quant=False 分支，此时 input_scale 仅作为占位符传递给 apply_block_scaled_mm，其值不会被读取（具体可参考 flashinfer.py 中对应实现）。

关键文件：

- vllm/model_executor/kernels/linear/scaled_mm/BlockScaledMMLinearKernel.py（模块推理内核；类别 source；类型 data-contract）：所有变更均在此文件。将占位 input_scale 张量的创建从 new_ones(1) 改为 new_empty(1)，消除了一次不必要的逐元素核启动，从而提升吞吐。

关键符号：未识别

关键源码片段

vllm/model_executor/kernels/linear/scaled_mm/BlockScaledMMLinearKernel.py

所有变更均在此文件。将占位 input_scale 张量的创建从 new_ones(1) 改为 new_empty(1)，消除了一次不必要的逐元素核启动，从而提升吞吐。

```
# vllm/model_executor/kernels/linear/scaled_mm/BlockScaledMMLinearKernel.py

if self.apply_input_quant:
    q_input, input_scale = self.quant_fp8(
        input_2d, input_scale, scale_up, use_triton=self.use_triton
    )
else:
    q_input = input_2d
    # Provide a concrete placeholder so apply_block_scaled_mm args are
    # always Tensors. Subclasses with apply_input_quant=False must not
    # use As in apply_block_scaled_mm.
    input_scale = (
        input_scale if input_scale is not None else input_2d.new_empty(1)
        # 改为 new_empty(1) 以跳过不必要的元素级初始化 —— 该占位张量不会被读取
    )
```

评论区精华

无实质性讨论。仅有一条 gemini-code-assist[bot] 的自动代码审查，指出无待处理意见；随后 mgoin 批准了该 PR，评价为“Nice find, makes sense to me!”。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。input_scale 在 apply_input_quant=False 时仅作占位符传递，不会被读取。new_empty() 分配未初始化内存，但后续不会访问其值，因此不会引发未定义行为。需确保所有子类在 apply_input_quant=False 时确实不使用 As 参数，目前代码注释已声明此约定。建议添加明确注释或在将来重构为更安全的模式（如传递 None 并从 apply_block_scaled_mm 方处理）。
- 影响：影响范围仅限 Fp8BlockScaledMMLinearKernel 及其子类（如 Fp8BlockScaledDynamicMMLinearKernel），具体为当 apply_input_quant=False 时的代码路径。收益：H200 上 DeepSeek V4-Pro 推理 e2e 输出 token 吞吐提升约 2%。对准确性无影响（GSM8K 基准持平）。由于是单行改动且语义清晰，回归风险极低。
- 风险标记：占位常量语义约定（注释约定）

关联脉络

- 暂无明显关联 PR