PR #44122 完整报告

vllm-project/vllm

[Refactor] Remove dead code fp quant

合并时间：2026-06-04 02:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44122

执行摘要

- 一句话：移除 FPQuant 中的死代码
- 推荐动作：建议合并，属于常规代码清理，无技术风险，有助于保持代码库整洁。

功能与动机

根据 PR 标题和 body，目的是清理不再使用的代码（dead code），提升代码可维护性。

实现拆解

1. 核心源码路径 - vllm/model_executor/layers/quantization/fp_quant.py（data-contract）：源码主路径；包含 控制流调整、配置键调整、异常路径调整；+0/-21

关键文件：

- vllm/model_executor/layers/quantization/fp_quant.py（模块 量化层；类别 source；类型 data-contract；符号 FPQuantConfig, FPQuantLinearMethod）：移除了 pseudoquantization 参数及其配置解析逻辑，以及 backward_hadamard_matrix 权重注册。

关键符号：未识别

关键源码片段

vllm/model_executor/layers/quantization/fp_quant.py

移除了 pseudoquantization 参数及其配置解析逻辑，以及 backward_hadamard_matrix 权重注册。

```
class FPQuantConfig(QuantizationConfig):
    """Config class for FPQuant."""
    # 移除 pseudoquantization 参数，该参数始终为 False 且对应 ValueError 分支已删除
    def __init__(
        self,
        hadamard_group_size: int = 32,
        forward_dtype: str = "mxfp4",
        forward_method: str = "abs_max",
        modules_to_not_convert: list[str] | None = None,
    ) -> None:
        super().__init__()
```

```
        self.hadamard_group_size = hadamard_group_size
        self.forward_dtype = forward_dtype
        self.forward_method = forward_method
        self.modules_to_not_convert = modules_to_not_convert
        # 原 pseudoquantization 检查已被移除，因为该功能从未被使用

class FPQuantLinearMethod(LinearMethodBase):
    # ...
    def create_weights(self, layer, input_size_per_partition, output_partition_sizes, ...):
        # ...
        layer.register_parameter("forward_hadamard_matrix", forward_hadamard_matrix)
        # backward_hadamard_matrix 已移除，因为其在 forward 中从未被使用
        # layer.register_parameter("backward_hadamard_matrix", backward_hadamard_matrix) #
        已删除
```

评论区精华

该 PR 没有 review 评论，仅获得一名 reviewer 的 Approval，表明变更清晰无争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。移除的参数和权重在代码中已无实际使用路径，且通过移除 pseudoquantization 检查（原为立即 raise ValueError）进一步清理了死分支。但需确认无下游代码隐性依赖 backward_hadamard_matrix 属性。
- 影响：影响范围仅限 FPQuantConfig 和 FPQuantLinearMethod 类的内部实现，外部调用者不受影响，因为移除的均为内部未使用字段。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR