PR #46389 完整报告

vllm-project/vllm

Humming support for 2/3/5/6/7-bit pack-quantized weight-only inference

合并时间：2026-06-25 01:53

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46389

执行摘要

- 一句话：扩展 WNA16 量化方案支持 2/3/5/6/7-bit pack-quantized 推理
- 推荐动作：值得精读，尤其是 pack_factor 改为 Fraction 以及 packed 维度计算的变化。这是 Humming kernel 支持的必要基础，也是量化路由重构的一步。建议关注后续对 Humming kernel 的集成 PR。

功能与动机

在 PR #45185 的讨论中需要将 wNa16 支持分离出来，以便 Humming kernel 能够处理 2/3/5/6/7-bit 的 pack-quantized weight-only 推理。同时避免与 Marlin kernel 的路由冲突（Marlin 不支持 per-tensor 量化），因此不启用 per-tensor 路径。

实现拆解

1. 扩展位宽映射(compressed_tensors_wNa16.py): 将 WNA16_SUPPORTED_TYPES_MAP 从 {4,8} 扩展到 {2,3,4,5,6,7,8}，并对应添加 scalar_types。pack_factor 改为 Fraction(32, num_bits) 以正确处理非 2 的幂次位宽。
2. 非对称量化校验(compressed_tensors_wNa16.py): 在 __init__中增加对 asymmetric 量化的校验：只有当 num_bits 在 WNA16_ZP_SUPPORTED_TYPES_MAP（即 4 或 8）时才允许，否则抛出 ValueError。
3. Packed 维度计算(compressed_tensors_wNa16.py): 在 create_weights 中用 math.ceil(input_size_per_partition * num_bits / 32) 替代之前的 input_size_per_partition // pack_factor，确保 packed 整数维度的正确性。
4. Layer 属性设置(compressed_tensors_wNa16.py): 在 create_weights 中设置 layer.input_size_per_partition、output_size_per_partition 等属性，以兼容 ParallelLMHead 等需要这些属性的下游 kernel。
5. 路由逻辑简化(compressed_tensors.py 和 schemes/init.py): 移除对 WNA16_SUPPORTED_BITS 的导入和依赖，_is_wNa8o8_int 不再单独判断 sub-byte 情况，统一由 pack_quantized 格式条件路由。
6. 新 scalar 类型(scalar_type.py): 添加 uint5b16、uint6b32、uint7b64 三个 ScalarType，用于 5、6、7-bit 的 pack-quantized 表示。
7. 测试配套(无): 本次未添加新测试，但 PR 作者用 lm_eval 对 Gemma 等模型进行了评估验证。

关键文件：

- vllm/model_executor/layers/quantization/compressed_tensors/schemes/compressed_tensors_wNa16.py（模块 量化方案；类别 source；类型 core-logic；符号 CompressedTensorsWNA16, WNA16_SUPPORTED_TYPES_MAP, WNA16_SUPPORTED_BITS）：核心实现文件，扩展位宽映射、修改 pack_factor 和 packed 维度计算、添加 layer 属性设置，是变更的主体
- vllm/model_executor/layers/quantization/compressed_tensors/compressed_tensors.py（模块 量化路由；类别 source；类型 data-contract；符号 CompressedTensors._is_wNa8o8_int, CompressedTensors._get_scheme_from_parts）：修改量化路由逻辑，移除对 WNA16_SUPPORTED_BITS 的依赖，简化 scheme 选择
- vllm/model_executor/layers/quantization/compressed_tensors/schemes/__init__.py（模块 导出列表；类别 source；类型 configuration；符号 WNA16_SUPPORTED_BITS）：调整公开接口，移除 WNA16_SUPPORTED_BITS 的导出
- vllm/scalar_type.py（模块 标量类型；类别 source；类型 core-logic；符号 uint5b16, uint6b32, uint7b64）：新增 5/6/7-bit 的 scalar 类型定义，支撑新位宽的量化表示

关键符号：CompressedTensorsWNA16.init, CompressedTensorsWNA16.create_weights, CompressedTensors._is_wNa8o8_int, CompressedTensors._get_scheme_from_parts

评论区精华

mgoin 要求用 google/gemma-4-E2B-it-qat-mobile-ct 和 E4B 模型做评估，确保路径不被破坏；HDCharles 后续提供了评估日志。dsikka 询问能否添加 / 更新 smoke 测试，整体评价 LGTM。

- Gemma 模型评估验证 (testing): HDCharles 提供了评估日志，mgoin 认可结果并批准
- 添加 smoke 测试 (testing): 未在本 PR 中实现，计划后续另开 PR 添加

风险与影响

- 风险：
  1. pack_factor 改为 Fraction: 在涉及整数索引的场景中，需要确保所有使用 pack_factor 的地方都已兼容 Fraction。patch 中已调整 packed_input_dim 计算，但需确认其他间接使用者（如 weight_loader）是否正确。
  2. 路由逻辑变更: 取消 num_bits 检查后，所有 pack_quantized 格式都会进入 WNA16 scheme，若存在尚未支持的位宽（如 1-bit）则可能提前暴露不支持的错误。当前 map 覆盖 2-8，暂时无害。
  3. Asymmetric 量化限制: 2/3/5/6/7-bit 非对称量化会直接抛出异常，用户需知道此限制。
  4. 缺少测试覆盖: dsikka 指出未更新 smoke 测试，存在回归风险。 - 影响：用户现在可以加载和推理 2/3/5/6/7-bit pack-quantized 模型，拓宽了模型格式兼容性。系统端，WNA16 scheme 的位宽范围扩大，不会影响已支持的 4/8-bit 模型。对 Marlin kernel 无影响，因为该路径仍由 choose_mp_linear_kernel 根据配置选择。团队后续需要为新位宽增加测试和基准。 - 风险标记：核心路径变更（量化路由），缺少测试覆盖 ，

Fraction 类型可能引入精度 / 性能问题

关联脉络

- PR #45185 Original Humming support PR (not in this context, but referenced): 该 PR 的讨论基础和设计来源，PR body 明确提及