PR #25292 完整报告

sgl-project/sglang

[Quant] Support asymmetric weight quant in compressed-tensors WNA16

合并时间：2026-06-05 04:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25292

执行摘要

- 一句话：支持 compressed-tensors WNA16 非对称权重量化
- 推荐动作：值得精读 dispatch 逻辑，了解如何与 vLLM 对齐。尽管测试被移除，但核心逻辑变更经过 review，且 kernel 路径有间接覆盖。建议未来在类似修复中保留最小单元测试。

功能与动机

用户报告（Issue #25291）在加载 AWQ 风格的非对称权重量化模型（symmetric=false）时抛出 NotImplementedError: No compressed-tensors compatible scheme was found.，而相同的量化在 vLLM 中工作正常。需要使 SGLang 的 dispatch 支持非对称权重量化。

实现拆解

1. 放宽对称性检查：在 _is_wNa16_group_channel 方法中移除对 weight_quant.symmetric 的硬性要求，仅保留策略（CHANNEL/GROUP）、静态量化等条件，非对称配置即可通过 dispatch。
2. 传递对称性参数：在 _get_scheme_from_parts 方法中构造 CompressedTensorsWNA16 时，新增传入 symmetric=weight_quant.symmetric 参数，使下游 kernel 能正确处理零点。
3. 测试变更：作者最初添加单元测试覆盖非对称分发，但 review 后认为不必要，最终移除了测试文件。

关键文件：

- python/sglang/srt/layers/quantization/compressed_tensors/compressed_tensors.py（模块 量化层；类别 source；类型 core-logic；符号 _is_wNa16_group_channel, _get_scheme_from_parts）：唯一修改文件，包含 dispatch 逻辑的两处关键调整：移除对称量化要求并传递 symmetric 参数

关键符号：_is_wNa16_group_channel, _get_scheme_from_parts

关键源码片段

python/sglang/srt/layers/quantization/compressed_tensors/compressed_tensors.py

唯一修改文件，包含 dispatch 逻辑的两处关键调整：移除对称量化要求并传递 symmetric 参数

```
def _is_wNa16_group_channel(self, weight_quant, input_quant):
    input_quant_none = input_quant is None
    is_channel_group = (
        weight_quant.strategy == QuantizationStrategy.CHANNEL.value
        or weight_quant.strategy == QuantizationStrategy.GROUP.value
    )
    is_static = not weight_quant.dynamic
    # Both symmetric and asymmetric weight quant are handled by
    # CompressedTensorsWNA16 via the Marlin kernel path; asymmetric
    # checkpoints carry a weight zero-point.
    return is_channel_group and input_quant_none and is_static

def _get_scheme_from_parts(self, weight_quant, input_quant):
    # ... other detection ...
    if self._is_wNa16_group_channel(weight_quant, input_quant):
        if (
            self.quant_format == CompressionFormat.pack_quantized.value
            and weight_quant.num_bits in WNA16_SUPPORTED_BITS
        ):
            return CompressedTensorsWNA16(
                num_bits=weight_quant.num_bits,
                strategy=weight_quant.strategy,
                group_size=weight_quant.group_size,
                symmetric=weight_quant.symmetric, # new: pass symmetry flag
                actorder=weight_quant.actorder,
            )
        else:
            raise ImportError('Other method ... not supported now')
```

评论区精华

审核者 b8zhong 在 review 中表示同意修改，但认为不需要单元测试："LGTM, but we don't need this UT. Could you please remove it"。作者随后移除了测试文件。核心逻辑修改无争议。

- 单元测试必要性 (testing): 作者移除了测试文件，PR 只保留核心逻辑变更。
- AWQ 模型加载失败报告 (question): PR 通过修改 dispatch 解决了问题。

风险与影响

- 风险：风险较低。主要风险是回归：对称量化模型仍能正确加载，因为移除的检查不是必要条件，且 symmetric 参数传递对对称量化为 True 无影响。非对称路径依赖 Marlin kernel 对 zero-point 的支持，该支持已在 vLLM 验证。潜在风险是组大小或策略组合兼容性，但本 PR 仅扩展了原有检查的允许范围。
- 影响：对用户：可以加载 AWQ 风格的非对称 INT4 量化模型（如 Qwen3.6-27B AWQ），扩大模型兼容性。对系统：仅影响 dispatch 路径，推理时无变化。对团队：小变更，容易维护。
- 风险标记：核心路径变更，缺少测试覆盖，对称量化回归风险

关联脉络

- 暂无明显关联 PR