PR #47780 完整报告

vllm-project/vllm

[Bugfix] [Quantization] Fix loading for CT DSV2

合并时间：2026-07-07 10:28

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47780

执行摘要

- 一句话：修复 DSV2 模型 FP8 索引器权重加载失败
- 推荐动作：变更简单且必要，已快速合并。建议了解该仓库的量化层权重命名规范，未来应避免对特定后缀的硬编码，采用更通用的命名匹配模式。

功能与动机

修复 DeepSeek-V2 模型在加载带有压缩张量的 checkpoint 时的加载失败问题。PR body 指出原代码对 FP8 量化布局有大量硬编码，而本次修复仅针对 weight_scale 命名适配。测试表明，此前 RedHatAI/GLM-5.2-NVFP4-FP8 的精度分数会完全崩溃。

实现拆解

在 vllm/model_executor/models/deepseek_v2.py 文件的 _try_load_fp8_indexer_wk 函数中，将第 832 行判断 is_scale 的条件从 "weight_scale_inv" in name 改为 "weight_scale" in name。该函数负责加载 FP8 格式的 WK 权重并去量化到 BF16 后存入融合参数 wk_weights_proj.weight。

关键文件：

- vllm/model_executor/models/deepseek_v2.py（模块 模型执行器；类别 source；类型 data-contract；符号 _try_load_fp8_indexer_wk）：包含被修复的 _try_load_fp8_indexer_wk 函数，该函数负责 FP8 WK 权重的加载和去量化融合。改动将 is_scale 检测从硬编码的 "weight_scale_inv" 改为更通用的 "weight_scale" 子串匹配，使支持压缩张量的 checkpoint 能正确加载。

关键符号：_try_load_fp8_indexer_wk

关键源码片段

vllm/model_executor/models/deepseek_v2.py

包含被修复的 _try_load_fp8_indexer_wk 函数，该函数负责 FP8 WK 权重的加载和去量化融合。改动将 is_scale 检测从硬编码的 "weight_scale_inv" 改为更通用的 "weight_scale" 子串匹配，使支持压缩张量的 checkpoint 能正确加载。

```
# vllm/model_executor/models/deepseek_v2.py
def _try_load_fp8_indexer_wk(
```

```
    name, tensor, buf, params_dict, loaded_params, pp_missing_layer_names
):
    """
    We fuse the WK and weights_proj projections, but in some checkpoints WK is stored
    in FP8 with a separate weight_scale（或 weight_scale_inv）, while weights_proj is stored in
    BF16.
    Upcasting to BF16 during loading enables the fusion. This function loads the FP8 WK
    weights and scale, and when both are available, dequantizes to BF16 and stores into
    the fused wk_weights_proj.weight parameter.
    """
    if "indexer.wk." not in name or "wk_weights" in name:
        return False # Weight is not an isolated WK weight for the indexer, ignore.
    is_weight = name.endswith(".weight") and tensor.dtype == torch.float8_e4m3fn
    # 修复：将 "weight_scale_inv" 改为 "weight_scale"，兼容两种命名风格
    is_scale = "weight_scale" in name # 原为 "weight_scale_inv" in name
    if not is_weight and not is_scale:
        return False # WK is not in FP8 format, ignore.
    # ...（后续代码不变）
```

评论区精华

无 review 讨论。PR 由 claude[bot] 自动评论因来自 fork 而跳过自动审查，最后由 mgoin 直接批准（无评论）。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅一行字符串匹配条件，从具体的 "weight_scale_inv" 放宽为 "weight_scale"。这可以兼容原有命名 weight_scale_inv（包含子串），同时支持新的 weight_scale 命名。不会引入回归，但需确保未来 checkpoint 的 scale 参数名符合 weight_scale 子串（当前未见异常）。
- 影响：影响范围仅限于 DeepSeek-V2 模型在使用 FP8 量化且带有压缩张量时的加载流程。直接修复了 RedHatAI/GLM-5.2-NVFP4-FP8 等 checkpoint 的加载正确性，使这些模型可正常部署。对不涉及 FP8 WK 权重的 checkpoint 无影响。
- 风险标记：单行变更，缺少测试覆盖

关联脉络

- PR #46168 [Bugfix] Preserve FP8 indexer WK pairs across incremental load_weights: 同样涉及 _try_load_fp8_indexer_wk 函数的修改，修复增量加载时 WK 权重丢失的问题，属于同一 function 的连续 bugfix。