PR #25646 完整报告

sgl-project/sglang

fix deepseek v4 hisparse

合并时间：2026-05-21 08:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25646

执行摘要

- 一句话：修复 HiSparse C4 压缩时 out_loc 错误
- 推荐动作：建议合并。这是一个精确且低风险的修复，解决了 HiSparse 模式下 v2 压缩器的精度问题。变更仅 11 行，逻辑清晰，有精度 benchmark 佐证。

功能与动机

PR body 中对比了 HiSparse + compressor_v2 在 GSM8K 200 examples 20-shot 上的表现：修复前准确率 82.5%，修复后准确率 96.0%，接近不启用 compressor_v2 时的 96.5%。说明原有代码导致 v2 压缩器在 HiSparse 模式下写入错误的 C4 位置，严重损害精度。

实现拆解

1. 修改 _forward_compress_all_in_one 函数签名：在 compressor_v2.py 中为 _forward_compress_all_in_one 新增 out_loc 参数，替代内部调用 self._get_out_loc(compress_ratio)，将 out_loc 的决定权交给调用方。
2. 在 forward_unified 中提前获取并调整 out_loc：在非 indexer 分支中，通过 token_to_kv_pool.layer_mapping[layer_id] 获取 compress_kv_pool，并检查其是否具有 translate_loc_to_hisparse_device 方法（HiSparse 特有）。如果存在，则调用该方法将 out_loc 转换为物理 C4 位置。
3. 传递转换后的 out_loc：将调整后的 out_loc 传递给 _forward_compress_all_in_one，确保压缩结果写入正确的 C4 缓存地址。

关键文件：

- python/sglang/srt/layers/attention/dsv4/compressor_v2.py（模块 压缩器；类别 source；类型 core-logic；符号 _forward_compress_all_in_one, forward_unified）：核心修复文件，修改了 _forward_compress_all_in_one 和 forward_unified 方法，添加 out_loc 参数并引入 HiSparse 设备位置转换逻辑。

关键符号：_forward_compress_all_in_one, forward_unified, translate_loc_to_hisparse_device

关键源码片段

python/sglang/srt/layers/attention/dsv4/compressor_v2.py

核心修复文件，修改了 _forward_compress_all_in_one 和 forward_unified 方法，添加 out_loc 参数并引入 HiSparse 设备位置转换逻辑。

```
# python/sglang/srt/layers/attention/dsv4/compressor_v2.py
# 关键变更：将 out_loc 作为显式参数传入，而非在函数内部调用 self._get_out_loc，
# 从而允许调用方根据 HiSparse 设备进行位置转换。

class CompressorBackendMixin:
    # ... 其他代码 ...

    def _forward_compress_all_in_one(
        self,
        *,
        kv_score_buffer: torch.Tensor,
        kv_score_input: torch.Tensor,
        ape: torch.Tensor,
        head_dim: int,
        norm: RMSNorm,
        freqs_cis_cache: torch.Tensor,
        kv_cache: torch.Tensor,
        is_indexer: bool,
        rotate: bool,
        compress_ratio: int,
        page_size: int,
        out_loc: torch.Tensor, # <-- 新增参数，让调用方控制写入位置
    ) -> None:
        # ... 不变 ...
        compress_norm_rope_store(
            kv_compressed,
            plan,
            norm_weight=norm.weight,
            norm_eps=norm.variance_epsilon,
            freq_cis=freqs_cis_cache,
            out_loc=out_loc, # <-- 使用传入的 out_loc
            kvcache=kv_cache,
            page_size=page_size,
        )

    def forward_unified(
        self,
        x: torch.Tensor,
        forward_batch: ForwardBatch,
        layer_id: int,
        compressor: Compressor,
    ) -> None:
        # ... 前面代码不变 ...
        out_loc = self._get_out_loc(compressor.ratio) # <-- 提前获取默认位置
        if compressor.is_in_indexer:
            kv_cache = token_to_kv_pool.get_index_k_with_scale_buffer(layer_id)
```

```
        page_size = token_to_kv_pool.get_index_k_page_size()
    else:
        # 对于非 indexer 模式（即 core compress），获取 compress_kv_pool
        _, _, compress_kv_pool = token_to_kv_pool.layer_mapping[layer_id]
        assert compress_kv_pool is not None
        kv_cache = token_to_kv_pool.get_extra_key_buffer(layer_id)
        page_size = token_to_kv_pool.get_extra_key_page_size(layer_id)
        if hasattr(compress_kv_pool, "translate_loc_to_hisparse_device"):
            # v2 压缩器直接写入原始 C4 KV 张量。
            # HiSparse C4 因此需要物理 C4 位置。
            out_loc = compress_kv_pool.translate_loc_to_hisparse_device(out_loc)
    self._forward_compress_all_in_one(
        # ... 其他参数不变 ...
        out_loc=out_loc, # <-- 使用可能已转换的 out_loc
    )
```

评论区精华

该 PR 的 review 讨论较少，两位 reviewer（yhyang201, xiezhq-hermann）均直接批准，未提出争议或疑问。代码注释清晰解释了 HiSparse 场景下需要物理 C4 位置的原因。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅影响 HiSparse + compressor_v2 组合的非 indexer 分支，且通过 hasattr 安全检测确保仅在 HiSparse 设备存在时才执行位置转换。其他模式（indexer、非 HiSparse）不受影响。精度测试显示修复后效果与不启用 compressor_v2 基本一致，无回归。
- 影响：对 HiSparse + compressor_v2 用户：GSM8K 准确率从 82.5% 提升至 96.0%，接近无压缩的精度，是显著的质量修复。对其他用户：无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #25859 [DSA] Make MQA logits free memory ratio configurable: 同样修改了 DeepSeek V4 相关代码，属于同模块的持续优化。