

PR #29613 完整报告

sgl-project/sglang

[DSA] Use cos_sin_cache for DSA indexer fusion

合并时间: 2026-06-30 14:49

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29613>

执行摘要

- 一句话: DSA indexer 融合直接使用 `cos_sin_cache` 避免缓存不同步
- 推荐动作: 建议相关开发者阅读此 PR, 了解如何通过简化缓存设计消除冗余和潜在 bug。对于使用 DSA indexer 的模型 (如 GLM-5.2) 部署可提升内存效率。

功能与动机

在 #29576 中, 修复了 DSA indexer fusion 导致内存消耗过大的 bug, 但当时使用了一个独立的 `_dsa_indexer_freqs_cis` 缓存, 与 `rotary_emb.cos_sin_cache` 分开存储。这需要手动同步, 当调用 `_ensure_cos_sin_cache_length` 替换 RoPE 缓存以适应更长上下文时可能导致 stale-cache 行为。本 PR 直接复用 `cos_sin_cache`, 彻底消除这一风险, 同时减少额外的显存占用。

实现拆解

1. 在 `dsa_indexer.py` 中删除 `_shared_indexer_freqs_cis` 函数和 `__init__` 中的缓存创建, 添加 `_indexer_cos_sin_cache` 属性直接返回 `self.rotary_emb.cos_sin_cache`。
2. 修改 `dsv32/elementwise.py` 和 `dsv4/elementwise.py` 中 `fused_k_indexer_norm_rope`、`fused_k_indexer_norm_rope_store` 和 `fused_q_indexer_rope_first_quant` 函数的参数从 `freqs_cis` 改为 `cos_sin_cache`, 移除内部的 `torch.view_as_real` 转换。
3. 更新 CUDA 核文件 `indexer_k.cuh` 和 `main_norm_rope.cuh`, 将参数结构体中的 `freqs_cis` 指针替换为 `cos_sin_cache` 指针, 并新增 `load_rope_first_cos_sin` 辅助函数从 `(cos, sin)` 连续布局加载数据。
4. 调整测试 `test_dsv32_indexer_fusion.py`, 使 `_make_inputs` 直接生成 `cos_sin_cache`, 并更新所有调用处; 新增 `test_indexer_uses_replaced_rope_cache_for_fused_kernels` 测试验证替换缓存后的正确性。

关键文件:

- `python/sglang/srt/layers/attention/dsa/dsa_indexer.py` (模块 索引器; 类别 source; 类型 core-logic; 符号 `_shared_indexer_freqs_cis`, `_indexer_cos_sin_cache`): 核心变更: 删除 `_shared_indexer_freqs_cis`, 添加 `_indexer_cos_sin_cache` 属性, 直接使用 `rotary_emb.cos_sin_cache`, 是 PR 的主要逻辑改动。
- `test/registered/jit/test_dsv32_indexer_fusion.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `test_indexer_uses_replaced_rope_cache_for_fused_kernels`,

DummyRotary)：测试配套：修改测试输入生成方式以匹配新接口，新增替换缓存正确性测试，确保变更正确。

- python/sglang/jit_kernel/dsv32/elementwise.py (模块 JIT 核；类别 source；类型 core-logic；符号 fused_k_indexer_norm_rope, fused_k_indexer_norm_rope_store)：关键 Python 封装：修改 fused_k_indexer_norm_rope 和 fused_k_indexer_norm_rope_store 函数的签名，从 freqs_cis 改为 cos_sin_cache，并移除中间转换。
- python/sglang/jit_kernel/dsv4/elementwise.py (模块 JIT 核；类别 source；类型 core-logic；符号 fused_q_indexer_rope_first_quant)：对称修改：将 fused_q_indexer_rope_first_quant 的参数从 freqs_cis 改为 cos_sin_cache，保持接口一致。
- python/sglang/jit_kernel/csrc/deepseek_v32/indexer_k.cuh (模块 CUDA 核；类别 other；类型 core-logic；符号 load_rope_first_cos_sin, FusedKIndexerNormRopeParams)：CUDA 核文件：修改参数结构体 FusedKIndexerNormRopeParams，使用 cos_sin_cache 指针替代 freqs_cis，新增 load_rope_first_cos_sin 辅助函数。
- python/sglang/jit_kernel/csrc/deepseek_v4/main_norm_rope.cuh (模块 CUDA 核；类别 other；类型 core-logic；符号 load_rope_first_cos_sin, FusedQIndexerRopeHadamardQuantParams)：CUDA 核文件：修改 FusedQIndexerRopeHadamardQuantParams 结构体，使用 rope_cache 字段替代 freqs_cis，并添加 load_rope_first_cos_sin 函数 (DSV4 路径的对称适配)。

关键符号：_shared_indexer_freqs_cis, _indexer_cos_sin_cache, fused_k_indexer_norm_rope, fused_k_indexer_norm_rope_store, fused_q_indexer_rope_first_quant, test_indexer_uses_replaced_rope_cache_for_fused_kernels, load_rope_first_cos_sin

关键源码片段

python/sglang/srt/layers/attention/dsa/dsa_indexer.py

核心变更：删除 _shared_indexer_freqs_cis，添加 _indexer_cos_sin_cache 属性，直接使用 rotary_emb.cos_sin_cache，是 PR 的主要逻辑改动。

```
# python/sglang/srt/layers/attention/dsa/dsa_indexer.py
```

```
class Indexer(MultiPlatformOp):
    # ... 其他代码

    @property
    def _indexer_cos_sin_cache(self) -> torch.Tensor:
        """
        直接返回共享的 RoPE 缓存 `cos_sin_cache`，替代之前独立的 `_shared_indexer_freqs_cis`。
        这样当 `_ensure_cos_sin_cache_length` 更新缓存时，indexer 自动使用新值，
        避免手动同步导致的 stale-cache 问题。
        """
```

```

        return self.rotary_emb.cos_sin_cache

def _fused_k_prepare_and_store(self, ...):
    # ... 其他准备代码
    fused_k_indexer_norm_rope_store(
        k_to_compute,
        self.index_k_cache, # type: ignore[arg-type]
        loc,
        self.k_norm.weight,
        self.k_norm.bias,
        self.k_norm.variance_epsilon,
        self._indexer_cos_sin_cache, # 直接传入 cos_sin_cache
        positions,
        page_size,
    )

def _fused_q_prepare_and_store(self, ...):
    # ... 其他准备代码
    fused_q_indexer_rope_first_quant(
        q.contiguous(),
        weights_raw,
        q_scale_gate,
        self._indexer_cos_sin_cache, # 直接传入 cos_sin_cache
        positions,
    )

```

python/sglang/jit_kernel/dsv32/elementwise.py

关键 Python 封装：修改 fused_k_indexer_norm_rope 和 fused_k_indexer_norm_rope_store 函数的签名，从 freqs_cis 改为 cos_sin_cache，并移除中间转换。

```
# python/sglang/jit_kernel/dsv32/elementwise.py
```

```

def fused_k_indexer_norm_rope(
    k_input: torch.Tensor,
    weight: torch.Tensor,
    bias: torch.Tensor,
    eps: float,
    cos_sin_cache: torch.Tensor, # 直接传入 (max_pos, 64) 的浮点缓存
    positions: torch.Tensor,
) -> torch.Tensor:
    '''V3.2 indexer K: LayerNorm + RoPE on leading dims -> bf16. CUDA only.'''
    # k_input may be a non-contiguous wk slice; output is always contiguous.
    k_out = torch.empty(k_input.shape, dtype=k_input.dtype, device=k_input.device)
    module = _jit_k_indexer_norm_rope_module(k_input.dtype)
    module.forward(
        k_input,
        k_out,
        weight,

```

```
    bias,  
    cos_sin_cache, # 直接传递, 不再需要 view_as_real 转换  
    positions,  
    float(eps),  
)  
return k_out
```

评论区精华

Review 中 b8zhong 对 `indexer_k.cuh` 文件中的注释准确性提出疑问, 原注释写 'un-rotated activations'。作者 mmangkad 解释称 'activations' 指 LayerNorm K 输出, 'rotated' 指 RoPE 后而非 RHT。b8zhong 确认并同意。该问题已解决, 无未解决争议。

- 注释准确性: 'un-rotated activations' 应改为 'rotated activations' (question): 注释措辞确认, 无需修改。

风险与影响

- 风险: 直接引用 `rotary_emb.cos_sin_cache` 替代独立缓存, 若 `cos_sin_cache` 在其他路径被意外修改 (例如长度调整后未正确同步), 可能影响 `indexer` 结果。但 `cos_sin_cache` 是整个引擎共享的 RoPE 缓存, 其更新机制已经稳定; 本次变更反而消除了之前两缓存可能不一致的风险。总体风险低。
- 影响: 仅影响 DeepSeek-V3.2 模型的 DSA 融合 `indexer` 路径。减少显存占用 (约减少 0.25 GB/GPU 的独立缓存)。对原生 DSV4 `indexer` 路径无影响。功能正确性通过 AIME 2025 精度测试验证 (pass@1: 90.83%), 与基准一致。
- 风险标记: 缓存依赖变更, 低风险, DSA 专用

关联脉络

- PR #27705 DSA indexer fusion: 此 PR 是 DSA indexer fusion 工作的基础, 引入融合索引器使用了独立的 `freqs_cis` 缓存。
- PR #29576 Fix DSA indexer fusion bug causing excessive memory consumption.: 修复内存消耗 bug, 使用了共享缓存但仍需同步; 本 PR 在此基础上进一步消除独立缓存。