PR #25399 完整报告

sgl-project/sglang

Add NPU condition for cosine and sine caching

合并时间：2026-05-15 20:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25399

执行摘要

- 一句话：NPU 条件化缓存 cos/sin 节省约 230MB
- 推荐动作：该 PR 属于性能微优化，变更简单直接，适合快速合并。可关注 review 中的优化建议，在后续迭代中进一步减少冗余计算。

功能与动机

PR body 明确指出 'Conditionally compute cached cosine and sine values based on NPU flag. ~230 MB saving.'，旨在减少非 NPU 场景下的显存占用。

实现拆解

1. 修改文件：python/sglang/srt/layers/rotary_embedding/rope_variant.py 中的 _compute_cos_sin_cache 方法。
2. 变更点：将原来无条件计算的 emb、cos_cached_total、sin_cached_total 三行代码包裹在 if _is_npu: 条件块内。
3. 效果：非 NPU 设备不再计算和存储 cos_cached_total 和 sin_cached_total，从而节省约 230MB 显存。

关键文件：

- python/sglang/srt/layers/rotary_embedding/rope_variant.py（模块 旋转编码；类别 source；类型 core-logic）：核心变更文件，修改了 _compute_cos_sin_cache 方法，添加 NPU 条件判断以节省显存。

关键符号：_compute_cos_sin_cache

关键源码片段

python/sglang/srt/layers/rotary_embedding/rope_variant.py

核心变更文件，修改了 _compute_cos_sin_cache 方法，添加 NPU 条件判断以节省显存。

```
def _compute_cos_sin_cache(self) -> torch.Tensor:
    inv_freq = self._compute_inv_freq(self.scaling_factor)
    t = torch.arange(
        self.max_position_embeddings * self.scaling_factor,
        device=self.device,
```

```
        dtype=torch.float32,
    )
    freqs = torch.einsum("i,j -> ij", t, inv_freq)
    cos = freqs.cos() * self.mscale
    sin = freqs.sin() * self.mscale
    cache = torch.cat((cos, sin), dim=-1)
    # 仅在 NPU 环境下计算并缓存完整的 cos/sin 张量
    # 非 NPU 场景下跳过，节省约 230MB 显存
    if _is_npu:
        emb = torch.cat((freqs, freqs), dim=-1)
        self.cos_cached_total = torch.cos(emb) * self.mscale
        self.sin_cached_total = torch.sin(emb) * self.mscale
    return cache
```

评论区精华

Review 评论来自 gemini-code-assist[bot]，建议优化 NPU 块内复用已计算的 cos 和 sin 张量，避免重复的 torch.cos/torch.sin 调用和中间 emb 张量的内存分配。该建议未被采纳，PR 已合并。

- 复用已计算的 cos/sin 张量以避免冗余调用 (performance): 建议未被采纳，PR 已按原始方案合并。

风险与影响

- 风险：风险较低。变更仅添加条件判断，不影响非 NPU 路径。NPU 路径的行为保持不变。但需确认 _is_npu 变量在上下文中已正确初始化，否则可能因变量未定义导致运行时错误。
- 影响：影响范围小，仅针对 NPU 环境下的旋转位置编码缓存。非 NPU 用户无感知，NPU 用户显存节省约 230MB，推理时内存效率提升。
- 风险标记：变量 _is_npu 未定义风险

关联脉络

- 暂无明显关联 PR