PR #31294 完整报告

sgl-project/sglang

Skip no-op EAGLE sampling renormalization

合并时间：2026-07-16 13:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31294

执行摘要

- 一句话：跳过 EAGLE 默认采样归一化
- 推荐动作：建议精读核心变更逻辑，但无需深入 review 其他文件。该 PR 展示了如何利用已有的批量状态信息消除冗余计算，是一个干净的性能优化模式。可关注类似优化是否也可应用于其他采样路径。

功能与动机

EAGLE 验证阶段对形状为 [batch_size * draft_token_num, vocab_size] 的概率张量执行冗余的 top-k 和 top-p 重归一化，默认配置下这两步是恒等变换但耗费 GPU 时间。标准采样路径已跳过这些变换，EAGLE 应保持一致。PR body 指出 'The redundant work is larger than in the standard sampling path'，因为 EAGLE 处理的词汇行数是 draft_token_num 倍。

实现拆解

1. 检查批量采样需求：在 *eagle_sample* 函数的 target_probs 计算后，利用 sampling_info.need_top_k_sampling 和 sampling_info.need_top_p_sampling 两个布尔标志，判断当前批次是否需要执行 top-k 或 top-p 重归一化。这些标志由 SamplingBatchInfo 维护，已在其他采样路径中使用。
2. 条件执行 top-k 重归一化：仅当 sampling_info.need_top_k_sampling 为 True 时，调用 top_k_renorm_prob 并保留原有的 NaN 检测；否则跳过整个操作。
3. 条件执行 top-p 重归一化：仅当 sampling_info.need_top_p_sampling 为 True 时，调用 top_p_renorm_prob 并保留 NaN 检测；否则跳过。
4. 保持其余逻辑不变：温度缩放 softmax、draft_probs 校验、rejection sampling 分支、tree_speculative_sampling_target_only 调用等均未修改。

关键文件：

- python/sglang/srt/speculative/eagle_utils.py（模块 验证引擎；类别 source；类型 core-logic；符号 *eagle_sample*）：核心变更文件，在 *eagle_sample* 函数中为 top-k 和 top-p 重归一化添加条件判断，避免默认配置下的冗余计算。

关键符号：*eagle_sample*

关键源码片段

python/sglang/srt/speculative/eagle_utils.py

核心变更文件，在 eagle_sample 函数中为 top-k 和 top-p 重归一化添加条件判断，避免默认配置下的冗余计算。

```
# python/sglang/srt/speculative/eagle_utils.py, eagle_sample 函数片段

# Apply temperature and get target probs
expanded_temperature = torch.repeat_interleave(
    sampling_info.temperatures, verify_input.draft_token_num, dim=0
) # (bs * num_draft_tokens, 1)

target_probs = F.softmax(
    next_token_logits / expanded_temperature, dim=-1
) # (bs * num_draft_tokens, vocab_size)
maybe_detect_nan(target_probs, "v2 verify: target_probs after softmax")

# Only run top-k renormalization if at least one request in the batch
# uses a non-default top-k value. When all requests use top_k=all
# (the default), this step is an identity transform but adds GPU cost.
if sampling_info.need_top_k_sampling:
    target_probs = top_k_renorm_prob(
        target_probs,
        torch.repeat_interleave(
            sampling_info.top_ks, verify_input.draft_token_num, dim=0
        ),
    ) # (bs * num_draft_tokens, vocab_size)
    maybe_detect_nan(target_probs, "v2 verify: target_probs after top_k_renorm")

# Only run top-p renormalization if at least one request in the batch
# uses a non-default top-p value. Same reasoning as above.
if sampling_info.need_top_p_sampling:
    target_probs = top_p_renorm_prob(
        target_probs,
        torch.repeat_interleave(
            sampling_info.top_ps, verify_input.draft_token_num, dim=0
        ),
    )
    maybe_detect_nan(target_probs, "v2 verify: target_probs after top_p_renorm")

# The rest of the function (reshape, draft_probs, rejection sampling, etc.)
# remains unchanged.
target_probs = target_probs.reshape(bs, verify_input.draft_token_num, -1)
```

评论区精华

PR 讨论极少，CI 触发 spec 测试组后全部通过，合并者 kpham-sgl 直接批准，没有设计争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅添加两个条件检查，使用已存在的 SamplingBatchInfo 布尔标志，不影响混合批次的正确性（需要过滤的请求仍然执行重归一化）。NaN 检测在条件内部保留。风险在于若 need_top_k_sampling 或 need_top_p_sampling 在某些边缘场景未正确维护，可能导致应执行的重归一化被跳过，但该标志已在其他路径使用，可靠性高。
- 影响：对使用默认采样参数（top_k=all, top_p=1）的 EAGLE 推理用户有正向性能收益，基准测试显示 ITL 平均降低 2.8%，P99 降低 4.9%。对混合批次（部分请求使用非默认过滤）无影响。影响范围限于 speculative decoding 中的 EAGLE 验证路径。
- 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- PR #31488 Overlap grammar (constrained decoding) with speculative decode verify: 涉及 EAGLE speculative decoding 路径的性能优化，与本 PR 同属 speculative-decoding 性能改进系列。
- PR #31677 [Spec] Extract DFlash compact draft-cache rebuild helpers: 同样在 speculative-decoding 模块内进行重构优化。
- PR #31738 Fix stop boundaries for grammar-constrained speculative decoding: 关联 speculative-decoding 正确性修复。