PR #25155 完整报告

sgl-project/sglang

[perf] avoid hidden states d2h when return_hidden_states=false

合并时间：2026-05-14 14:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25155

执行摘要

- 一句话：避免不必要的 hidden states D2H 拷贝
- 推荐动作：建议精读，这是一个典型的小型性能优化案例，展示了如何通过传递控制参数避免不必要的 GPU-CPU 数据传输。

功能与动机

在 overlap scheduling 中，即使 return_hidden_states=false，copy_to_cpu 仍然会无条件执行 hidden states 的 D2H 拷贝，造成 GPU 带宽浪费。PR 作者提供了性能对比截图，显示改动后调度器 D2H 延迟显著降低。

实现拆解

1. 修改 GenerationBatchResult.copy_to_cpu 方法（python/sglang/srt/managers/utils.py）：新增 return_hidden_states: bool = True 参数，将原来 if self.logits_output.hidden_states is not None: 改为 if return_hidden_states and self.logits_output.hidden_states is not None:，条件化执行 hidden states 的 D2H 拷贝。默认值为 True 保持向后兼容。
2. 更新 scheduler.py 中的调用点（python/sglang/srt/managers/scheduler.py）：在两个调用 copy_to_cpu 的地方显式传入当前 batch 的 return_hidden_states 标志，一处是 overlap 路径中的 run_batch 方法（第 3036-3039 行），另一处是 launch_batch_sample_if_needed 方法（第 3144-3147 行）。

关键文件：

- python/sglang/srt/managers/utils.py（模块 调度器；类别 source；类型 core-logic；符号 copy_to_cpu）：核心逻辑所在，修改了 copy_to_cpu 方法，新增 return_hidden_states 参数控制 hidden states 的 D2H 拷贝条件。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic）：调用 copy_to_cpu 的两个位置被更新，传入 return_hidden_states 标志，确保调度器根据 batch 配置传递条件。

关键符号：copy_to_cpu

关键源码片段

python/sglang/srt/managers/utils.py

核心逻辑所在，修改了 copy_to_cpu 方法，新增 return_hidden_states 参数控制 hidden states 的 D2H 拷贝条件。

```
def copy_to_cpu(self, return_logprob: bool, return_hidden_states: bool = True):
    """Copy tensors to CPU in overlap scheduling.
    Only the tensors which are needed for processing results are copied,
    e.g., next_token_ids, logits outputs
    """
    if return_logprob:
        # 复制 logprobs 相关张量到 CPU（省略具体代码）
        ...
    # 关键变更：只有当 return_hidden_states 为 True 时才复制 hidden_states
    if return_hidden_states and self.logits_output.hidden_states is not None:
        self.logits_output.hidden_states = self.logits_output.hidden_states.to(
            "cpu", non_blocking=True
        )
    self.next_token_ids = self.next_token_ids.to("cpu", non_blocking=True)
    # 其他复制操作 ...
    self.copy_done.record()
```

python/sglang/srt/managers/scheduler.py

调用 copy_to_cpu 的两个位置被更新，传入 return_hidden_states 标志，确保调度器根据 batch 配置传递条件。

```
# run_batch 方法中（约第 3036 行）
batch_result.copy_to_cpu(
    return_logprob=batch.return_logprob,
    return_hidden_states=batch.return_hidden_states, # 新增参数
)

# launch_batch_sample_if_needed 方法中（约第 3144 行）
batch_result.copy_to_cpu(
    return_logprob=self.cur_batch.return_logprob,
    return_hidden_states=self.cur_batch.return_hidden_states, # 新增参数
)
```

评论区精华

无人工 review 讨论。gemini-code-assist[bot] 自动评论描述了变更内容，ispobock 批准了 PR。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅新增一个控制条件，默认值 True 保持了与旧行为的完全兼容；变更集中于 scheduler 中的两个调用点，不涉及其他模块。

- 影响：影响范围小，仅对 overlap scheduling 模式下 return_hidden_states=false 的场景有性能提升；不影响显式请求 hidden states 的用例。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR