

PR #29121 完整报告

sgl-project/sglang

Handle input-embed-only batches in eager runner

合并时间: 2026-06-25 01:33

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29121>

执行摘要

- 一句话: 修复 eager runner 对无 input_ids 的 batch 支持
- 推荐动作: 建议快速合并。这是一个简单且必要的修复, 完善了对嵌入输入的支持。

功能与动机

根据 PR body: "Use input_embeds to derive the token count when input_ids is absent", 支持输入嵌入 (embedding) 直接传入而不需要 token ID 的场景。这是对 eager runner 的一个缺漏修复。

实现拆解

1. 在 python/sglang/srt/model_executor/runner/eager_runner.py 的 load_batch 方法中, 将原先直接使用 forward_batch.input_ids.shape[0] 的代码改为三路条件判断: 优先取 input_ids 的长度, 若为 None 则取 input_embeds 的长度, 若两者均为 None 则置为 0。
2. 对应测试文件 test_attention_patching.py 和 test_mlx_runner_pool_contract.py 中仅有格式化调整 (行折叠、EOF 换行符), 不涉及功能变更。

关键文件:

- python/sglang/srt/model_executor/runner/eager_runner.py (模块 执行器; 类别 source; 类型 data-contract): 核心变更文件, 修复 load_batch 方法中对 input_ids 为 None 时的处理。
- test/registered/unit/hardware_backend/mlx/test_attention_patching.py (模块 测试; 类别 test; 类型 test-coverage): 仅格式化调整, 无功能变化。
- test/registered/unit/hardware_backend/mlx/test_mlx_runner_pool_contract.py (模块 测试; 类别 test; 类型 test-coverage): 仅添加尾随换行符。

关键符号: 未识别

关键源码片段

[python/sglang/srt/model_executor/runner/eager_runner.py](#)

核心变更文件, 修复 load_batch 方法中对 input_ids 为 None 时的处理。

```
def load_batch(
    self, forward_batch: ForwardBatch, pp_proxy_tensors=None, **kwargs
```

```

) -> ForwardBatch:
    # ... 前面的代码不变 ...
    raw_bs = forward_batch.batch_size
    # 优先使用 input_ids 确定 token 数量
    if forward_batch.input_ids is not None:
        raw_num_tokens = forward_batch.input_ids.shape[0]
    elif forward_batch.input_embeds is not None:
        # 若 input_ids 为空但提供了 input_embeds, 则使用嵌入张量长度
        raw_num_tokens = forward_batch.input_embeds.shape[0]
    else:
        # 兜底：两者均无则标记为 0 (零 token batch)
        raw_num_tokens = 0
    # 后续代码不变, 使用 raw_num_tokens 填充缓冲区
    registry = self._eager_registry
    registry.fill_from(
        forward_batch,
        raw_bs=raw_bs,
        padded_bs=raw_bs,
        raw_num_tokens=raw_num_tokens,
        padded_num_tokens=raw_num_tokens,
        pp_proxy_tensors=pp_proxy_tensors,
    )
    return registry.extract_buffer(
        padded_bs=raw_bs,
        padded_num_tokens=raw_num_tokens,
        forward_batch_template=forward_batch,
    )

```

评论区精华

该 PR 无讨论 comment。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。更改仅影响 eager runner 中一个分支（input_ids 为 None 时），且逻辑直白（回退到 input_embeds 或 0）；不会影响 CUDA graph 或其他 runner。测试文件仅格式化调整，无实质性改动。
- 影响：影响范围局限在 eager runner 的 load_batch 流程，使得使用 input_embeds 直接输入的 batch（例如某些嵌入模型或多模态输入）能够在 eager 模式下正常运行。对已有功能无负面影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR