

PR #32157 完整报告

sgl-project/sglang

[diffusion] fix: per-shard FP8 scale shape for single-GPU fused linears

合并时间: 2026-07-29 11:28

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32157>

执行摘要

- 一句话: 修复单 GPU 下 MergedColumnParallelLinear 的 FP8 scale 形状错误
- 推荐动作: PR 值得合并, 逻辑清晰且修复明确。建议审查者关注后续的单元测试 PR, 确保单 GPU 场景得到充分覆盖。

功能与动机

在单 GPU 上, 使用 MergedColumnParallelLinear 的 FP8 量化检查点 (如 HunyuanVideo 的 img_attn_qkv) 会加载失败, 回退到非 FP8 模式。原因是融合检查点将 3 个投影的 scale 堆叠为 (3,), 但层的 weight_scale/input_scale 被创建为 (1,), 导致断言失败。

实现拆解

1. 定位问题: 在 MergedColumnParallelLinear.__init__ 中, self.output_sizes 仅在 tp_size > 1 时设置。当 tp_size == 1 时, self.output_sizes 未定义, 导致 ColumnParallelLinear 基类的 __init__ 中通过 hasattr(self, "output_sizes") 判断失败, 进而使用 [sum(output_sizes)] 作为 output_partition_sizes, 使得 scale 的形状错误。
2. 修复方法: 将 self.output_sizes = output_sizes 无条件地移到 super().__init__() 之前, 并移除原来的条件判断 if get_group_size(tp_group) > 1:。同时删除了原来位于 super().__init__() 之后的重复赋值语句。
3. 影响范围: 仅修改 python/sglang/multimodal_gen/runtime/layers/linear.py 文件中的 MergedColumnParallelLinear.__init__ 方法, 共 3 行增删。该修改不影响 QKVParallelLinear, 因为它已经无条件设置了 self.output_sizes, 所以从未受影响。

关键文件:

- python/sglang/multimodal_gen/runtime/layers/linear.py (模块 扩散模型; 类别 source; 类型 core-logic; 符号 MergedColumnParallelLinear.init): 包含 MergedColumnParallelLinear 类的初始化逻辑, 是本 PR 唯一的修改文件。修复了 output_sizes 初始化顺序导致 FP8 scale 形状错误的 bug。

关键符号: MergedColumnParallelLinear.init

关键源码片段

python/sglang/multimodal_gen/runtime/layers/linear.py

包含 MergedColumnParallelLinear 类的初始化逻辑，是本 PR 唯一的修改文件。修复了 output_sizes 初始化顺序导致 FP8 scale 形状错误的 bug。

```
# python/sglang/multimodal_gen/runtime/layers/linear.py

class MergedColumnParallelLinear(ColumnParallelLinear):
    def __init__(
        self,
        input_size: int,
        output_sizes: list[int],
        bias: bool = True,
        gather_output: bool = False,
        skip_bias_add: bool = False,
        params_dtype: torch.dtype | None = None,
        quant_config: QuantizationConfig | None = None,
        prefix: str = "",
        tp_group: dist.ProcessGroup = None,
    ):
        tp_group = tp_group or get_tp_group()
        # 关键修复：无条件设置 output_sizes 在 super().__init__() 之前，
        # 这样 ColumnParallelLinear 的 __init__ 方法可以通过
        # hasattr(self, "output_sizes") 获取到正确的值，用于计算
        # per-shard 的 output_partition_sizes。当 tp_size == 1 时，
        # 之前的条件判断导致该属性未设置，使得 scale 形状错误。
        self.output_sizes = output_sizes
        super().__init__(
            input_size=input_size,
            output_size=sum(output_sizes),
            bias=bias,
            gather_output=gather_output,
            skip_bias_add=skip_bias_add,
            params_dtype=params_dtype,
            quant_config=quant_config,
            prefix=prefix,
            tp_group=tp_group,
        )
        assert all(output_size % self.tp_size == 0 for output_size in output_sizes)

# ... 后续方法不变
```

评论区精华

审查者 mickqian 批准了 PR 并建议添加单元测试。作者 AgainstEntropy 回应将作为后续 PR 添加单元测试。

- 缺少单元测试 (testing): 作者同意作为后续 PR 添加单元测试。

风险与影响

- 风险：风险极低：修改本质上是将赋值语句提前并移除条件，逻辑上保证了 `self.output_sizes` 始终在基类使用前被设置。由于 `QKVParallelLinear` 已经采用类似的无条件赋值，该修改的一致性很好。但缺少单元测试覆盖单 GPU 场景下的 `scale` 形状验证。
- 影响：直接影响：修复了单 GPU 场景下使用 `MergedColumnParallelLinear` 的 FP8 量化模型（如 `HunyuanVideo`）的加载失败问题。间接影响：无，因为修改不影响多 GPU 场景，且不影响非量化模型。
- 风险标记：缺少单 GPU 测试覆盖

关联脉络

- 暂无明显关联 PR