PR #1748 完整报告

THUDM/slime

fix: resolve SP/CP gradient inflation in FLA (linear attention) layers

合并时间：2026-03-22 14:31

原文链接：http://prhub.com.cn/THUDM/slime/pull/1748

执行摘要

该 PR 修复了在序列并行和模型并行设置中，线性注意力层反向传播时梯度计算错误导致梯度范数膨胀的问题。通过自定义 autograd 函数和修改 gather 操作，确保梯度正确分割而非重复累加，已用实际模型验证修复效果。

功能与动机

在分布式训练中，序列并行（SP）和模型并行（CP）用于线性注意力层时，gather_from_sequence_parallel_region 和 dist.nn.all_gather 的 backward 错误地执行了 reduce-scatter 操作，导致梯度被错误地乘以 TP×CP 倍。这可能导致训练不稳定或失败。PR body 中说明修复后使用 Qwen3.5-27B 和 Qwen3Next-80B 模型验证了梯度范数恢复正常，解决了训练中的潜在问题。

实现拆解

主要改动集中在 slime_plugins/models/hf_attention.py 文件：

1. 配置加载增强：添加 _load_hf_config 函数作为 fallback，当 transformers 无法识别模型类型时，直接从 config.json 加载配置。
2. 自定义 Autograd 函数：定义 _AllGatherForDuplicatedComputation 类，继承 torch.autograd.Function，其 forward 执行 all-gather，backward 仅返回本地梯度切片，避免 reduce-scatter。
3. 序列并行修复：在 HuggingfaceAttention.forward 中，设置 tensor_parallel_output_grad=False 以指示 backward 执行分割而非 reduce-scatter。

在 qwen3_5.py 和 qwen3_next.py 中：

- 集成 _load_hf_config 替换原有的配置加载。
- 添加逻辑以计算 layer_types 当配置类缺少该属性时，确保模型层类型识别。

次要改动：slime/utils/reloadable_process_group.py 调整了内存清理阈值，可能优化资源管理。

评论区精华

该 PR 没有 review 评论，因此没有讨论记录。

风险与影响

风险：

- 自定义 autograd 函数 _AllGatherForDuplicatedComputation 可能未处理所有边界情况，如梯度形状或数据类型不匹配。
- 修改了核心并行路径，可能影响其他模型或配置，需要充分测试。
- 配置 fallback 逻辑可能不覆盖所有模型类型，导致兼容性问题。

影响：

- 对用户：修复后，在 SP/CP 下训练线性注意力模型将更稳定，避免梯度膨胀导致的训练问题。
- 对系统：梯度计算逻辑更准确，提升分布式训练的正确性。
- 对团队：此修复展示了在分布式环境中处理重复计算梯度问题的设计模式，值得学习。

关联脉络

从近期历史 PR 看，此 PR 独立修复了特定 bug，没有明显直接相关的其他 PR。但涉及 Qwen 模型的支持（如 PR #1742 添加了 Qwen3.5 loss mask），表明团队在持续优化 Qwen 系列模型的兼容性和性能，本修复是这一趋势中的一部分。