PR #1822 完整报告

THUDM/slime

Revert no_grad for entropy to prevent comm stuck in dsa

合并时间：2026-04-09 19:20

原文链接：http://prhub.com.cn/THUDM/slime/pull/1822

执行摘要

- 一句话：移除熵计算中的 no_grad 上下文，修复 DSA 模式下通信卡死问题。
- 推荐动作：建议技术管理者和核心工程师精读此 PR，重点关注：
  1. 熵计算梯度保留的设计决策，理解 DSA 通信机制的特殊要求。
  2. 分布式张量重建逻辑中对 None 值的处理方式，确保边缘场景覆盖。
  3. 结合近期 PR #1788（修复 loss oom）和 #1762（修复 grad_norm 初始化）一起分析，这些 PR 都涉及损失计算和梯度处理的底层优化。

功能与动机

根据 PR 标题和提交信息，变更动机是修复 DSA（Distributed Shared Architecture）模式下通信卡死问题。具体表现为熵计算使用 torch.no_grad() 时，在分布式环境中可能导致通信操作无法正常完成。PR body 未提供详细描述，但从代码变更可推断需要确保熵计算张量具有梯度信息以维持分布式通信的连续性。

实现拆解

实现方案分为两个关键文件修改：

1. slime/backends/megatron_utils/loss.py：修改 _allgather_cp_redistribute 函数，增加对 None 值的跳过逻辑，并统一使用参考张量的 dtype/device 创建零张量，避免因 value 为 None 导致的属性访问错误。同时移除 need_entropy_grad 参数及相关逻辑。
2. slime/utils/ppo_utils.py：重构 calculate_log_probs_and_entropy 函数，完全移除 need_entropy_grad 参数和 torch.no_grad() 上下文管理，确保熵计算始终使用可计算梯度的 logits.clone() 输入，避免梯度信息丢失。

关键文件：

- slime/backends/megatron_utils/loss.py（模块 megatron_utils）：修改了分布式张量重建的核心函数 _allgather_cp_redistribute，增加 None 值跳过逻辑并统一 dtype/device 引用，直接影响损失计算的通信稳定性。
- slime/utils/ppo_utils.py（模块 ppo_utils）：重构了 calculate_log_probs_and_entropy 函数，彻底移除梯度控制参数和 no_grad 上下文，这是修复通信卡死的核心变更点。

关键符号：_allgather_cp_redistribute, calculate_log_probs_and_entropy, compute_entropy_from_logits

评论区精华

该 PR 没有 review 评论，属于直接合并的修复。从代码变更看，核心决策是彻底移除熵计算中的梯度控制逻辑，统一使用可计算梯度的张量，这可能是基于DSA环境下通信机制的特殊要求。

- 熵计算梯度保留的必要性 (correctness): 决定完全移除梯度控制，确保 DSA 环境下通信连续性。
- 分布式张量重建中的 None 值处理 (correctness): 实现更健壮的 None 值处理机制，确保分布式通信的鲁棒性。

风险与影响

- 风险：技术风险包括：
  1. 性能影响：移除 no_grad 可能增加显存占用和计算开销，因为熵计算现在会保留梯度信息。
  2. 兼容性风险：变更可能影响非 DSA 环境下的训练行为，特别是当 entropy_coef=0 时，原本不需要梯度计算，现在可能产生不必要的开销。
  3. 逻辑一致性：loss.py 中增加对 None 值的跳过逻辑，需确保在所有分布式场景下都能正确处理 None 值，避免遗漏边缘情况。 关键风险点在于熵计算梯度保留对整体训练稳定性的影响，需验证是否会导致梯度爆炸或内存溢出。
- 影响：影响范围：
  1. 对用户：修复 DSA 环境下的训练卡死问题，提升分布式训练的稳定性，但可能略微增加显存使用。
  2. 对系统：影响所有使用 Megatron 损失计算和 PPO 熵计算的训练流程，特别是涉及分布式通信的场景。
  3. 对团队：变更涉及核心训练逻辑，需要团队关注后续性能监控和回归测试。 影响程度中等，主要针对特定架构（DSA）的问题修复，但改动触及分布式通信和梯度计算的基础层。
- 风险标记：核心路径变更，梯度计算调整，分布式通信依赖

关联脉络

- PR #1788 [WIP] fix loss oom: 同样修改了 slime/backends/megatron_utils/loss.py 文件，优化损失计算内存使用，与本 PR 的 loss 修改有直接关联。
- PR #1762 [Fix] Initialize grad_norm before found_inf skip path: 涉及 Megatron 训练中的梯度处理问题修复，与本 PR 的梯度计算调整属于同一技术领域。
- PR #1807 sync from internal: 同样修改了 megatron_utils 模块，优化多模态训练兼容性，显示该模块近期活跃度较高。