PR #1928 完整报告

radixark/miles

[fix] DSA indexer on Blackwell: send the DSA indexer wk unquantized

合并时间：2026-07-30 08:56

原文链接：http://prhub.com.cn/radixark/miles/pull/1928

执行摘要

- 一句话：DSA indexer wk 改发 bf16，修复 Blackwell 在线权重同步
- 推荐动作：该 PR 值得精读，尤其关注在线权重同步与量化格式匹配的问题。设计决策是选择直接发送 bf16 而非尝试适配 ue8m0，简化了路径且成本可接受。建议后续增加针对该场景的回归测试。

功能与动机

在 sglang >= 0.5.16 中，引擎将 DSA indexer 的 wk 与 weights_proj 融合为单个 bf16 的 indexer.wk_weights_proj，并在线用 plain fp32 block scales 对 fp8 wk 反量化。Blackwell 上转换器为 non-MoE linear 生成 ue8m0-packed scales，导致 block_quant_dequant 形状不匹配，每次在线权重同步都失败。作者在 GLM-5.2 744B-A40B 上复现，修复前所有 update_weights_from_distributed 均失败。

实现拆解

1. 修改文件 miles/backends/megatron_utils/megatron_to_hf/processors/quantizer_fp8.py，在 quantize_params_fp8 的量化白名单中删除 self_attention.wk.weight 和 self_attention.indexer.linear_wk.weight 两项。
2. 删除后，wk 参数不再经过 _quantize_param，而是直接以 bf16 原始格式输出，与 sglang 目标参数格式一致。
3. 保留 self_attention.wq_b.weight 等其余 DSA 参数量化，因为引擎仍以 fp8 消费这些参数。
4. 未新增测试，但作者在 GB300 集群上验证：修复后 35 次权重更新调用零失败，训练正常推进。

关键文件：

- miles/backends/megatron_utils/megatron_to_hf/processors/quantizer_fp8.py（模块 量化器；类别 source；类型 core-logic；符号 quantize_params_fp8）：核心修复文件：从 FP8 量化白名单移除 DSA indexer 的 wk 权重，使其以 bf16 在线发送，避免 ue8m0 scale 不兼容导致的在线权重同步崩溃。

关键符号：quantize_params_fp8

关键源码片段

miles/backends/megatron_utils/megatron_to_hf/processors/quantizer_fp8.py

核心修复文件：从 FP8 量化白名单移除 DSA indexer 的 wk 权重，使其以 bf16 在线发送，避免 ue8m0 scale 不兼容导致的在线权重同步崩溃。

```
# 文件：miles/backends/megatron_utils/megatron_to_hf/processors/quantizer_fp8.py
# 函数 quantize_params_fp8 决定哪些 Megatron 参数需要量化
# 下面展示 DSA indexer 相关的白名单决策（head 版本）

if rest in [
    "self_attention.linear_proj.weight",
    "self_attention.linear_qkv.weight",
    "mlp.linear_fc1.weight",
    "mlp.linear_fc2.weight",
    # mla
    "self_attention.linear_q_proj.weight",
    "self_attention.linear_q_down_proj.weight",
    "self_attention.linear_q_up_proj.weight",
    "self_attention.linear_kv_down_proj.weight",
    "self_attention.linear_kv_up_proj.weight",
    # DSA indexer
    "self_attention.wq_b.weight",
    # 移除了 self_attention.wk.weight 与 self_attention.indexer.linear_wk.weight
    # 原因：sglang >= 0.5.16 将 wk 融合为 bf16 的 indexer.wk_weights_proj
    # 并使用 ue8m0 scale 反量化 fp8 的 wk，在 Blackwell 上会 shape 不匹配
    # linear attention
    "self_attention.linear_attn.in_proj_qkv.weight",
    "self_attention.linear_attn.in_proj_z.weight",
    "self_attention.linear_attn.out_proj.weight",
    # DeepSeek V4 attention
    "self_attention.wq_a.weight",
    "self_attention.wkv.weight",
    "self_attention.wo_b.weight",
    "self_attention.indexer.linear_wq_b.weight",
]:
    # 对这些参数执行 FP8 量化并返回
    quantize_named_params = []
    for converted_name, param in converted_named_params:
        quantize_named_params.extend(_quantize_param(args, converted_name, param, weight_
        block_size))
    return quantize_named_params

# 其他参数直接返回原样
return converted_named_params
```

评论区精华

Review 中 Zhichenzzz 直接 APPROVED，无讨论。Issue 评论中作者 yueming-yuan 发布了验证结果：在 GLM-5.2 744B-A40B、16x GB300 的 fully-async RL 配置上，修复前每次 update_weights_from_distributed 都失败，修复后 35 次调用零失败，且确认与 shared_experts fusion 无关，问题仅出现在 _load_fused_indexer_wk 路径。

- 修复验证结果 (testing): 修复有效，问题解决。

风险与影响

- 风险：风险点：1）该修复仅针对 DSA indexer wk，如果未来 sglang 版本恢复 fp8 消费 wk 或改变融合方式，需要重新评估；2）wk 以 bf16 发送会增加约 60 MB 同步流量，对数百 GB 级同步可忽略；3）改动无自动化测试覆盖，依赖手动验证；4）如果其他模型结构也有类似融合但未在列表中，可能触发同类问题。
- 影响：影响范围：仅影响 Blackwell 上使用 DSA（如 DeepSeek 架构）模型且 sglang >= 0.5.16 的在线权重同步路径。修复后这些配置的 RL 训练可正常启动。对离线 checkpoint 转换、Hopper 等平台无影响。团队可在相关模型测试中回补该场景。
- 风险标记：缺少测试覆盖，在线权重同步关键路径，特定硬件条件

关联脉络

- PR #2014 fix: quantize non-interleaved DSA indexer wk: 同样修改 quantizer_fp8.py 中 DSA indexer 的量化逻辑，与本 PR 属于同一功能线，可合并理解 DSA indexer 量化历史。