PR #25107 完整报告

sgl-project/sglang

perf(nvfp4): free unused source scales after weight processing

合并时间：2026-05-13 07:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25107

执行摘要

- 一句话：释放 NVFP4 权重量化中不再使用的源缩放张量
- 推荐动作：该 PR 设计清晰，内存收益显著，风险可控，建议合并。值得关注其删除张量的策略和保留 w13_weight_scale_2/w2_weight_scale_2 的权衡。

功能与动机

NVFP4 权重加载后，多个源端缩放张量（如 layer.input_scale、layer.weight_scale_2、layer.weight_scale）不再被前向推理使用，却一直占用 GPU 显存。PR 描述中详细分析了每张量在 apply() 中的使用情况，确认删除的安全性。

实现拆解

1. Linear 方法 (ModelOptNvFp4LinearMethod.process_weights_after_loading)：在推导出 alpha 和 input_scale_inv 后，del layer.input_scale, layer.weight_scale_2。在 FlashInfer-TRTLLM 分支，del layer.weight_scale；在 CUTLASS 分支末尾同样 del layer.weight_scale。
2. MoE 方法 (ModelOptNvFp4FusedMoEMethod.process_weights_after_loading)：无条件删除 w13_input_scale 和 w2_input_scale。TRTLLM 分支删除 w13_blockscale_swizzled 和 w2_blockscale_swizzled。非 TRTLLM 分支（CUTLASS/CuteDSL）末尾删除 w13_weight_scale 和 w2_weight_scale。
3. 保留 w13_weight_scale_2 和 w2_weight_scale_2：因为 flashinfer_cutedsl 通过 hasattr 读取它们并存在后备路径，为避免精度漂移，保留并留下 TODO。
4. 删除过时注释：移除了“Keep per-shard scales intact for hot reload”注释，因为热加载会在 process_weights_after_loading 重新运行前重新绑定这些参数。
5. 测试：未新增单元测试，但触发了多个 CI 测试（如 test_nvidia_nemotron_3_super_nvfp4.py、test_deepseek_v4_flash_fp4_b200.py 等）。

关键文件：

- python/sglang/srt/layers/quantization/modelopt_quant.py（模块 量化；类别 source；类型 core-logic；符号 ModelOptNvFp4LinearMethod.process_weights_after_loading, ModelOptNvFp4FusedMoEMethod.process_weights_after_loading）：唯一修改文件，包含 Linear 和 MoE 方法的 process_weights_after_loading 中删除不再使用的缩放张量的逻辑。

关键符号：process_weights_after_loading

评论区精华

该 PR 没有 Review 评论，但作者在 PR body 中详细讨论了内存节省的计算和保留 w13_weight_scale_2/w2_weight_scale_2 的技术理由。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险：如果 apply() 方法未来需要访问已删除的张量（如 input_scale），会导致 AttributeError。目前确认所有路径均不再使用这些张量，但需通过 CI 验证。
  2. 精度风险：保留 w13_weight_scale_2/w2_weight_scale_2 避免了潜在精度漂移，但删除其他张量可能影响未来兼容性。
  3. 热加载风险：热加载机制在重新运行 process_weights_after_loading 前会重新绑定参数，因此删除是安全的。
- 影响：
  1. 用户：显存显著降低（如 Kimi-K2.5 约 15 GiB/rank），可能允许部署更大模型或增加批处理大小。
  2. 系统：无延迟影响，因为这些张量不在热路径上。
  3. 团队：需确保未来对 apply() 的修改不依赖于已删除张量。 - 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- 暂无明显关联 PR