PR #1978 完整报告

THUDM/slime

[docker] Fix qwen3 30B + deepep with sglang 0.5.12

合并时间：2026-05-29 17:38

原文链接：http://prhub.com.cn/THUDM/slime/pull/1978

执行摘要

- 一句话：修复 Qwen3 30B 与 Deepep 在 sglang 0.5.12 下的运行问题
- 推荐动作：值得合并。该 PR 解决了特定模型与运行时的兼容性问题，修复逻辑清晰，风险可控。建议在合并后增加相关 CI 测试覆盖，确保后续版本不会再次出现类似问题。

功能与动机

修复 Qwen3 30B 模型与 Deepep 在 sglang 0.5.12 环境下的兼容性和运行正确性问题，确保模型可以正常加载和推理。

实现拆解

1. 调整 DeepGEMM JIT EP activation 条件- 在 sglang/srt/layers/moe/moe_runner/deep_gemm.py 中，将原来硬性的 assert N % 4 == 0 and G % 4 == 0 改为有条件地启用 SGLANG_OPT_USE_JIT_EP_ACTIVATION：仅在满足维度对齐约束时才使用 JIT 路径，否则跳过，避免因维度不匹配导致 crash。
2. 移除 MoE 权重加载中的 zero 过滤- 在 sglang/srt/layers/moe/fused_moe_triton/layer.py 中，删除了 and "zero" not in weight_name 条件，使 FusedMoE 在加载权重时不再跳过名称中包含 zero 的权重。这可能是为了支持某些量化方案中 zero 权重参与合并或处理。
3. 放宽 CompressedTensors 量化检查- 在 sglang/srt/layers/quantization/compressed_tensors/compressed_tensors.py 中，移除了 is_symmetric 条件，使得原本要求对称量化的检查现在也接受非对称量化，以兼容 Qwen3 30B 的权重格式。
4. 新增 restore_weights_before_loading 方法- 在相同文件中为 CompressedTensorsFusedMoEMethod 类增加了 restore_weights_before_loading 方法，委托给 layer.scheme 实现，以确保在加载前正确恢复权重（可能涉及去量化或格式转换）。

关键文件：

- docker/patch/latest/sglang.patch（模块 部署脚本；类别 test；类型 test-coverage）：PR 中唯一修改的文件，包含了所有针对 sglang 0.5.12 的修复调整，是核心变更载体。

关键符号：未识别

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 兼容性风险：放宽量化检查（移除 is_symmetric）可能影响其他原本依赖对称性约束的模型，需确保该变更不破坏其他量化方案的加载逻辑。
  2. 性能影响：DeepGEMM 中 JIT EP activation 的条件变更可能导致在某些模型上前向路径发生变化，但影响范围较小。
  3. 回归风险：移除 zero 权重过滤可能引入之前因过滤而规避的权重处理问题，需验证模型收敛性。 - 影响：影响范围：仅影响 docker 环境下的 sglang 补丁文件，直接影响使用 sglang 0.5.12 并需要运行 Qwen3 30B 或 Deepep 的用户。影响程度：低到中等。对不涉及这些模型的用户无影响，但对目标用户至关重要。
- 风险标记：潜在回归风险，测试覆盖不足

关联脉络

- PR #1973 [docker] fix patch: 同为 docker 补丁修复，涉及相同文件 docker/patch/latest/sglang.patch，且本次 PR 在其基础上进一步修改。
- PR #1972 [docker] fix mooncake offload in sglang v0.5.12: 同样修复 sglang v0.5.12 的问题，与本次 PR 有相同的工作文件，属于同一修复系列。