PR #20717 完整报告

sgl-project/sglang

[CI] Add Per-Tensor, Blockwise FP8 Tests on SM120

合并时间：2026-04-02 09:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20717

执行摘要

本 PR 为 SM120 架构 GPU（如 RTX 5090）新增了 FP8 量化模型的 CI 测试，覆盖逐张量（Llama-3.1-8B-Instruct-FP8）和分块（Qwen3-4B-Instruct-2507-FP8）两种量化格式，通过 GSM8K 准确性评估验证功能。这是解决 #20600 测试覆盖率问题的一部分，填补了单 5090 GPU 上 FP8 测试的空白，风险较低但提升了特定硬件的测试完整性。

功能与动机

动机：根据 PR body，主要目标是“改进 SM120 上量化模型测试的覆盖率”。现有测试已包含 Blackwell 多 GPU 上的 FP4 覆盖和其他套件中的 FP8 测试，但缺少在单 5090/SM120 上的 FP8 量化模型测试。PR 作者指出这是“解决 #20600 的众多措施之一”。

关键表述：

> "Missing single-5090 coverage for these tests. This PR addresses coverage for FP8-quantized models in particular."

实现拆解

实现集中在一个新增的测试文件中：

文件路径	关键内容	作用
test/registered/quant/test_fp8_gemm_sm120.py	定义基类 FP8GemmSM120Base 和两个测试子类	封装服务器启动、GSM8K 评估和清理逻辑

核心逻辑：

1. 基类设计：FP8GemmSM120Base 提供通用的 setUpClass（启动服务器）和 test_gsm8k（运行评估）方法。
2. 测试注册：通过 register_cuda_ci(est_time=120, suite="stage-b-test-small-1-gpu") 将测试集成到 CI 流水线。
3. 两种量化格式测试：
   - TestFP8PerTensorGemmSM120Auto：测试逐张量量化的 Llama 模型，准确性阈值 ≥0.73。

- TestFP8BlockwiseGemmSM120Auto：测试分块量化的 Qwen 模型，准确性阈值≥0.87。

4. 硬件检查：使用 @unittest.skipIf(get_device_sm() < 100, "Test requires CUDA SM 100 or higher") 确保仅在 SM100+ 设备运行。

评论区精华

review 讨论非常简短但关键，体现了对测试覆盖完整性的关注：

> b8zhong: "Looks good. Can we also add Qwen/Qwen3-4B-Instruct-2507-FP8. The quantization format will be a bit different (blockwise vs per-tensor scale)."

作者 DerekY2 在 issue 评论中回应“刚刚添加了😉”，并在第二次提交中实现了该建议，确保了测试同时覆盖两种主要的 FP8 量化格式。

风险与影响

风险：

1. 外部依赖：测试依赖 Hugging Face 模型仓库，网络或模型可用性问题可能导致 CI 失败。
2. 阈值设置：准确性阈值（0.73 和 0.87）可能因模型更新或评估波动导致 CI 不稳定。
3. 执行时间：预估 120 秒可能增加 CI 流水线负担。

影响：

- 对用户：无直接影响。
- 对系统：提升了 SM120 设备上 FP8 量化功能的测试覆盖率，有助于早期发现回归。
- 对团队：CI 将更全面验证 FP8 量化，特别是针对 RTX 5090 等硬件，作为 #20600 整体测试改进的一部分。

关联脉络

与历史 PR 的关联：

1. #21888 和 #21576：都涉及 FP8 量化相关修改，属于同一技术领域，本 PR 的测试可能覆盖这些变更。
2. #21233：涉及量化代码重构，本 PR 的测试有助于验证重构后的功能。

演进方向：本 PR 是 #20600 测试覆盖率提升计划的一部分，近期多个 PR（如 #21888、#21576）都在加强量化相关测试，表明团队正系统性地完善量化功能的验证体系，特别是在不同硬件架构上的覆盖。