PR #22425 完整报告

sgl-project/sglang

[HiSparse]: Add HiSpares-DSA Model's nightly CI

合并时间：2026-04-09 16:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22425

执行摘要

此 PR 新增了 HiSparse-DSA 模型的夜间 CI 测试，验证 GLM-5-FP8 在 8-GPU H200 环境下的 HiSparse 稀疏注意力功能。测试配置了数据并行和 HiSparse 参数，并通过 GSM8K 评估验证准确率。review 中发现了资源配置不匹配和预估时间过长的问题，但 CI 测试已通过，表明风险得到缓解。这是一个中等重要的基础设施扩展，为 HiSparse 功能提供自动化验证。

功能与动机

PR 标题和测试内容表明，动机是为 HiSparse-DSA 模型建立持续的集成测试，确保其在特定硬件配置（8-GPU H200）下的功能正确性和稳定性。PR body 未明确说明动机，但从关联 Issue 评论中作者执行测试并验证通过的行为推断，这是基础设施扩展的一部分，旨在自动化验证 HiSparse 功能。

实现拆解

实现集中于单个测试文件，按模块拆解如下：

模块	关键改动	说明
测试注册	register_cuda_ci(est_time=720, suite="stage-c-test-8-gpu-h200", nightly=True)	将测试注册到 8-GPU H200 夜间套件，预估时间 720 分钟。
服务器配置	--tp 8 --dp 8 --enable-hisparse --hisparse-config '{"top_k": 2048, ...}'	配置 8 路张量并行、8 路数据并行，启用 HiSparse 并设置参数。
评估逻辑	run_eval 运行 GSM8K 评估，500个样本，断言准确率>0.94。	验证模型在 HiSparse 模式下的输出质量。

关键代码逻辑：

```
other_args = [
    "--trust-remote-code",
    "--tp", "8",
    "--dp", "8",
    "--enable-dp-attention",
```

```
    "--enable-hisparse",
    "--hisparse-config", '{"top_k": 2048, "device_buffer_size": 4096, "host_to_device_ratio": 5}',
]
```

评论区精华

review 中 gemini-code-assist[bot] 提出了两个关键问题：

1. 资源配置不匹配：

"The current configuration specifies --tp 8 and --dp 8, which requires a total of 64 GPUs. However, this test is registered in the stage-c-test-8-gpu-h200 suite, which is intended for 8-GPU environments."

建议调整配置以适应 8-GPU 环境（如 --tp 4 --dp 2）。作者未直接回应，但 CI 测试通过，暗示配置可能已调整或环境支持。

1. 预估时间优化：

"The est_time=720 (12 hours) seems excessively high for a nightly CI test... Consider reducing this value to a more realistic estimate."

建议减少到 180-240 分钟以提升 CI 效率。此问题未在讨论中明确解决。

风险与影响

- 资源配置风险：原始配置需要 64 个 GPU，而测试套件仅支持 8 个，可能导致资源不足失败。但 CI 通过表明风险已缓解（可能配置已调整或环境特殊支持）。
- 性能风险：预估时间 720 分钟可能影响 CI 调度效率，但实际测试时间可能较短，需监控实际运行时间。
- 测试覆盖风险：仅测试 GSM8K 数据集，覆盖范围有限，可能遗漏其他推理场景或边缘情况。
- 影响范围：对系统影响中等，新增夜间测试扩展了 CI 覆盖；对团队影响正面，自动化验证 HiSparse 功能减少手动测试负担。

关联脉络

与此 PR 相关的历史 PR 包括：

- PR #22418：将 Runai 模型加载测试移至夜间套件，类似 CI 基础设施调整。
- PR #22399：新增 GLM-5.1 夜间测试，同为扩展大模型测试覆盖。
- PR #22353：新增 Torch Profiler 分析工作流程，涉及测试和性能分析改进。

这些 PR 共同反映了仓库在扩展测试覆盖、优化 CI 基础设施方面的持续演进，特别是针对新硬件（如 H200）和新功能（如 HiSparse）的验证。本 PR 是这一趋势的一部分，旨在确保 HiSparse-DSA 模型在特定环境下的稳定性。