PR #29640 完整报告

sgl-project/sglang

Add Qwen3 MoE tests for PP compatibility with CP and DP

合并时间：2026-06-29 18:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29640

执行摘要

- 一句话：为 Qwen3 MoE 添加 PP×CP 和 PP×DP 的 GSM8K 精度测试
- 推荐动作：建议开发者在修改并行策略相关代码（如调度器、通信层、PP proxy）时，关注此测试的 CI 结果。此 Mixin 模式简洁，值得在其他组合测试中复用。可考虑后续扩展更多模型和评估集以进一步增强覆盖。

功能与动机

We have CI coverage for Pipeline Parallelism (PP) on its own and for Context Parallelism (CP) / DP attention on their own, but no test that exercises PP composed with another parallel strategy. Regressions in the cross-strategy plumbing would currently slip through.

实现拆解

实现分为 5 个步骤：

1. 在 test/registered/pp/test_pp_parallel_compat.py 中定义 Mixin 类 _Qwen3MoePPCompatMixin，封装 SGLang 服务启动（通过 popen_launch_server）和 GSM8K 精度评估逻辑（通过 run_eval），该类不继承 TestCase，避免被独立收集。
2. 定义两个具体测试子类继承 Mixin 和 CustomTestCase：TestQwen3MoePPxCP 配置 PP×CP（--pp-size 2 --attn-cp-size 2 --cp-strategy zigzag 等，并设置环境变量 SGLANG_ENABLE_CP_V2=1）；TestQwen3MoePPxDP 配置 PP×DP（--pp-size 2 --dp-size 2 --enable-dp-attention）。
3. 通过 register_cuda_ci 将测试注册到 CI 的 extra-b 阶段，指定 runner_config 为 4-gpu-h100，预估执行时间 600 秒。
4. 设置 GSM8K 门控阈值 GSM8K_BASELINE_ACCURACY = 0.93，采样参数 temperature=0.6, top_p=0.95, top_k=20，使用 200 个示例，与现有 CP 测试保持一致（参考 test/registered/cp/test_gqa_preill_cp.py）。
5. 服务启动的通用参数包括 --cuda-graph-max-bs-decode 32 --max-running-requests 32 --trust-remote-code --disable-piecewise-cuda-graph 以及多线程模型加载配置，确保测试环境稳定。

关键文件：

- test/registered/pp/test_pp_parallel_compat.py（模块 并行测试；类别 test；类型 test-coverage；符号 _Qwen3MoePPCompatMixin, setUpClass, tearDownClass, test_gsm8k）：唯一变更文件，新增整个测试套件，覆盖 PP×CP 和 PP×DP 两种组合并执行 GSM8K 精度门控。

关键符号：setUpClass, tearDownClass, test_gsm8k

评论区精华

仅有一条来自 gemini-code-assist[bot] 的评论，指出 SimpleNamespace 中的 host 和 port 属性在 run_eval 中未被使用，且 port 解析可能因 URL 格式不同而失败，建议移除。作者在后续 fix commit 中已移除这两个属性，最终代码干净简洁，问题已解决。

- 移除未使用的 host 和 port 参数 (correctness): 作者在后续 commit 中移除了这两个属性，最终代码干净简洁。

风险与影响

- 风险：
  1. 随机波动风险：GSM8K 精度门控基于随机采样（200 条），不同运行可能因种子或模型加载顺序波动，偶发低于阈值 0.93 导致 CI 失败。
  2. 外部依赖风险：测试需要从 HuggingFace 下载 Qwen/Qwen3-30B-A3B-FP8 模型，网络不稳定或模型更新可能导致超时或失败。
  3. 资源消耗：使用 4 块 GPU，CI 排队时间可能增加，且占用较多集群资源。
  4. 覆盖有限：仅测试一个模型和一个数据集，无法保证所有跨并行组合场景无回归。 - 影响：对 CI 流程增加约 600 秒的额外执行时间（extra-b 阶段），但填补了关键的组合并行测试空白。开发者修改涉及 PP、CP、DP 调度的代码时，应确保此测试通过，否则可能引入回归。对最终用户无直接交互影响，但长期可提高系统在组合并行下的稳定性。 - 风险标记：随机精度门控，外部模型依赖，4-GPU 资源消耗

关联脉络

- 暂无明显关联 PR