

PR #35450 完整报告

sgl-project/sglang

Update deeppep for SBO feature

合并时间: 2026-08-20 14:37

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35450>

执行摘要

- 一句话: 升级 sgl-deep-ep 至 0.1.2 并启用 SBO 测试
- 推荐动作: 该 PR 是常规依赖升级和测试启用, 简单明确, 无需精读。可关注 sgl-deep-ep 后续版本是否引入新的 API 变化。

功能与动机

PR 标题和代码注释表明, 之前 DeepEP 不支持 `--enable-single-batch-overlap` 标志, 导致相关测试被跳过。现在上游 DeepEP PR (<https://github.com/sgl-project/DeepEP/pull/10>) 已合并, 新版本 0.1.2 支持该功能, 因此需要升级依赖并重新启用测试。

实现拆解

1. 升级依赖: 在 `python/pyproject.toml` 中将 `sgl-deep-ep` 从 0.1.1 改为 0.1.2, 确保使用支持 SBO 的版本。
2. 移除测试跳过: 在 `test/registered/4-gpu-models/test_deepseek_v3_cutedsl_4gpu.py` 中删除 `TestDummyWithSBO` 类上的 `@unittest.skipIf(True, ...)` 装饰器, 恢复该测试类的正常运行。
3. 验证: 通过 `/rerun-test` 重新运行测试 `test_deepseek_v3_cutedsl_4gpu.py`, 测试通过。

关键文件:

- `python/pyproject.toml` (模块 依赖配置; 类别 `config`; 类型 `configuration`): 升级 `sgl-deep-ep` 到 0.1.2, 启用 SBO 支持的依赖基础。
- `test/registered/4-gpu-models/test_deepseek_v3_cutedsl_4gpu.py` (模块 测试用例; 类别 `test`; 类型 `test-coverage`): 移除 SBO 测试的 `skip` 装饰器, 恢复测试覆盖。

关键符号: 未识别

关键源码片段

`test/registered/4-gpu-models/test_deepseek_v3_cutedsl_4gpu.py`

移除 SBO 测试的 `skip` 装饰器, 恢复测试覆盖。

```
# test/registered/4-gpu-models/test_deepseek_v3_cutedsl_4gpu.py
# 移除 skip 装饰器, 恢复 SBO 测试
class TestDummyWithSBO(CustomTestCase):
```

```
@classmethod
def setUpClass(cls):
    cls.model = try_cached_model(DEFAULT_DEEPEEP_NVFP4_MODEL_FOR_TEST)
    cls.base_url = DEFAULT_URL_FOR_TEST
    other_args = [
        "--trust-remote-code",
        "--disable-radix-cache",
        # ... 其他参数
        "--quantization",
        "modelopt_fp4",
        # ...
    ]
```

评论区精华

作者 Fridge003 在 review 中提出将版本从 0.1.2rc0 改为 0.1.2（最终采纳）。无其他争议。

- 依赖版本选择 (design): 采纳使用正式版本 0.1.2。

风险与影响

- 风险：依赖升级可能导致 sgl-deep-ep 的 API 行为变化，但该版本升级是修复型（新增 SBO 支持），且测试重新启用后已通过，风险较低。测试仅在 4 GPU 环境运行，其他硬件或配置可能未覆盖。
- 影响：影响范围：仅限 DeepSeek 模型使用 sgl-deep-ep 的场景。启用 SBO 测试后，可验证新功能正确性，潜在改善性能。团队方面，需要确保 CI 环境能正确安装 sgl-deep-ep 0.1.2。
- 风险标记：依赖升级，测试环境特定

关联脉络

- PR #35593 [Fix] Support 128-aligned hidden sizes in the W4AFP8 DeepEP
low-latency requant kernel: 涉及 DeepEP 内核的修改，与本 PR 的 sgl-deep-ep 依赖升级相关。
- PR #32327 [DeepSeek-V4] Add Q8KV8 sparse MLA prefill runtime backend: 涉及 DeepSeek 模型和 DeepEP 相关后端。