

PR #27283 完整报告

sgl-project/sglang

[NPU] Enable consistency checking for diffusion tests

合并时间: 2026-06-09 20:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27283>

执行摘要

- 一句话: 为 NPU 扩散测试启用一致性检查并回退旋转嵌入优化
- 推荐动作: 该 PR 展示了如何为特定硬件平台 (NPU) 配置测试基础设施, 包括根据平台选择实现和测试配置, 值得测试和基础设施团队关注。建议后续持续监控 NPU 旋转嵌入性能。

功能与动机

根据 PR body 描述, 在添加 GT 数据 (PR#24630) 后, 需要为扩散测试启用一致性检查来验证输出, 此前一致性检查被禁用 (`run_consistency_check=False`)。同时发现旋转嵌入优化在 NPU 上存在性能退化, 需要回退。

实现拆解

1. 回退旋转嵌入实现: 在 `python/sglang/multimodal_gen/runtime/layers/rotary_embedding/utils.py` 的 `apply_flashinfer_rope_qk_inplace` 函数中, 当平台为 NPU 时, 使用 `apply_rotary_embedding` 替代原有的 `apply_rope_prefix`, 因为后者在 NPU 上导致性能下降。
2. 更新测试 GT 配置: 在 `python/sglang/multimodal_gen/test/test_utils.py` 中, 导入 `current_platform`, 根据是否为 NPU 动态修改 `SGL_TEST_FILES_CI_DATA_REVISION` (使用新 commit hash) 和 `SGL_TEST_FILES_CONSISTENCY_GT_BASE` (指向 ascend 子目录), 确保 GT 数据路径正确。
3. 启用一致性检查: 在 `python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py` 中, 移除所有 `DiffusionTestCase` 的 `run_consistency_check=False` 参数, 使测试执行一致性检查。

关键文件:

- `python/sglang/multimodal_gen/runtime/layers/rotary_embedding/utils.py` (模块 旋转嵌入; 类别 source; 类型 core-logic; 符号 `apply_flashinfer_rope_qk_inplace`): 核心逻辑修改: 在 NPU 上回退旋转嵌入的实现, 避免性能退化
- `python/sglang/multimodal_gen/test/test_utils.py` (模块 测试工具; 类别 test; 类型 test-coverage): 测试配置: 根据平台设置不同的 GT 路径和 commit hash, 确保 NPU 使用正确的 GT 数据
- `python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py` (模块 NPU 测试配置; 类别 test; 类型 test-coverage): 测试用例配置: 移除

run_consistency_check=False, 启用一致性检查

关键符号: apply_flashinfer_rope_qk_inplace

关键源码片段

python/sglang/multimodal_gen/runtime/layers/rotary_embedding/utils.py

核心逻辑修改: 在 NPU 上回退旋转嵌入的实现, 避免性能退化

```
# 在 apply_flashinfer_rope_qk_inplace 函数中, 当 use_flashinfer 为 False 且平台为 NPU 时,  
# 直接使用 torch 的 apply_rotary_embedding 替代原有手动实现的 apply_rope_prefix  
# 以避免 NPU 上的性能退化。
```

```
if not use_flashinfer:
```

```
    if flashinfer_apply_rope_inplace is None:  
        _warn_about_missing_flashinfer()
```

```
    half_size = rope_dim // 2
```

```
    if positions is None:
```

```
        cos = cos_sin_cache[:seqlen, :half_size].to(q.dtype)  
        sin = cos_sin_cache[:seqlen, half_size:].to(q.dtype)  
        cos = cos.unsqueeze(0).expand(bsz, -1, -1).reshape(bsz * seqlen, -1)  
        sin = sin.unsqueeze(0).expand(bsz, -1, -1).reshape(bsz * seqlen, -1)
```

```
    else:
```

```
        positions = positions.to(device=q.device, dtype=torch.long).view(-1)  
        cos = cos_sin_cache[positions, :half_size].to(q.dtype)  
        sin = cos_sin_cache[positions, half_size:].to(q.dtype)
```

```
# NPU 专用路径: 使用 apply_rotary_embedding 提高性能
```

```
if current_platform.is_npu():
```

```
    q_flat = q.reshape(bsz * seqlen, q_heads, d)  
    k_flat = k.reshape(bsz * seqlen, k_heads, d)  
    q_rot = apply_rotary_embedding(q_flat, cos, sin, interleaved=not is_neox)  
    k_rot = apply_rotary_embedding(k_flat, cos, sin, interleaved=not is_neox)  
    return q_rot.view(bsz, seqlen, q_heads, d), k_rot.view(  
        bsz, seqlen, k_heads, d  
    )
```

```
# 非 NPU 路径保持原有逻辑
```

```
def apply_rope_prefix(x: torch.Tensor, num_heads: int) -> torch.Tensor:
```

```
    # ... 原有实现 ...
```

python/sglang/multimodal_gen/test/test_utils.py

测试配置: 根据平台设置不同的 GT 路径和 commit hash, 确保 NPU 使用正确的 GT 数据

```
# 在文件开头导入 current_platform
```

```
from sglang.multimodal_gen.runtime.platforms import current_platform
```

```
# 原有默认 revision
```

```
SGL_TEST_FILES_CI_DATA_REVISION = 'caa56302ccf2d289e4488ed06d952edf5d2314cf'
```

```
# NPU 平台使用不同的 revision (指向 ascend GT 数据)
if current_platform.is_npu():
    SGL_TEST_FILES_CI_DATA_REVISION = '670d66a8a290b62c0c3c077b3e9b0f4a4d9a44e7'

# ... 其他常量定义 ...

# 设置默认 GT 基础路径
SGL_TEST_FILES_CONSISTENCY_GT_BASE = SGL_TEST_FILES_SGLANG_CONSISTENCY_GT_BASE

# NPU 平台使用 ascend 子目录
if current_platform.is_npu():
    SGL_TEST_FILES_CONSISTENCY_GT_BASE = SGL_TEST_FILES_SGLANG_CONSISTENCY_GT_BASE_ASCEND

# 在 _find_remote_consistency_gt_files 中, 非官方 GT cases 使用变量 bases
# 原先是硬编码 (SGL_TEST_FILES_SGLANG_CONSISTENCY_GT_BASE,), 现在改为 (SGL_TEST_FILES_CONSISTENCY_GT_BASE,)
# 这样当 NPU 时自动使用 ascend 路径
```

评论区精华

无实质性 Review 讨论, 直接获得批准。

- 暂无高价值评论线程

风险与影响

- 风险:
 1. 回退旋转嵌入优化: 可能降低 NPU 上旋转嵌入的整体性能, 但修复了更严重的性能退化, 需要持续监控。
 2. 一致性检查启用: 可能因 GT 数据不匹配导致测试失败, 尤其当 ci-data 仓库更新时。
 3. 配置路径依赖: SGL_TEST_FILES_CI_DATA_REVISION 和 SGL_TEST_FILES_CONSISTENCY_GT_BASE 依赖于外部 ci-data 仓库的路径和可用性, 可能因网络问题影响 CI 稳定性。- 影响: 直接影响 NPU 扩散测试的 CI 流程: 启用一致性检查增加了测试验证力度, 但可能增加失败概率; 回退优化可能影响旋转嵌入速度, 但整体有利于 NPU 性能。- 风险标记: 核心路径变更, 配置依赖外部仓库, 可能因 GT 数据不匹配导致测试失败

关联脉络

- PR #24630 Add GT for diffusion consistency: 该 PR 添加了 Ground Truth 数据, 本 PR 在此基础上启用一致性检查
- PR #27096 Rotary embedding optimization: 本 PR 回退了该优化以避免 NPU 上的性能退化