

PR #28223 完整报告

sgl-project/sglang

[NPU] Add MiMo-V2-Flash manual testcases

合并时间: 2026-06-15 19:57

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28223>

执行摘要

- 一句话: 新增 NPU MiMo-V2-Flash 手动测试用例
- 推荐动作: 该 PR 为纯测试补充, 对生产代码无影响。值得关注的是测试中使用的 EAGLE 投机解码参数配置, 可作为 NPU 上类似模型的参考模板。

功能与动机

在 Ascend NPU 上添加 MiMo-V2-Flash 模型的 GSM8K 精度验证手动测试用例, 以验证 cuda graph 和 MTP 投机解码的推理准确性。

实现拆解

1. 新增权重路径常量: 在 `python/sglang/test/ascend/test_ascend_utils.py` 中添加 `MIMO_V2_FLASH_WEIGHTS_PATH` 常量, 指向 `XiaomiMiMo/MiMo-V2-Flash` 模型权重目录。
2. 创建手动测试文件: 新增 `test/manual/ascend/llm_models/test_npu_mimo_v2_flash.py`, 定义 `TestMiMoV2FlashGraphWithMTP` 测试类。
3. 测试类结构: 继承 `GSM8KAscendMixin` (提供 GSM8K 评估逻辑) 和 `CustomTestCase`, 设置 `model` 为权重路径, `accuracy=0.9` 作为通过阈值。
4. 配置参数: 指定 `other_args` 包含 `--trust-remote-code`、`--mem-fraction-static 0.8`、`--attention-backend ascend`、`--tp-size 16`, 以及 EAGLE 投机解码参数 (`--speculative-algorithm EAGLE`、`--speculative-num-steps 3`、`--speculative-eagle-topk 1`、`--speculative-num-draft-tokens 4`、`--enable-multi-layer-eagle`)。
5. 后续讨论简化: 根据 Review 反馈, 将原先的 `_BASE_ARGS` 和 `_MTP_ARGS` 合并简化, 只保留核心 Graph+MTP 配置。

关键文件:

- `test/manual/ascend/llm_models/test_npu_mimo_v2_flash.py` (模块 NPU 模型测试; 类别 test; 类型 test-coverage; 符号 `TestMiMoV2FlashGraphWithMTP`): 新增的手动测试文件, 包含 MiMo-V2-Flash 模型在 NPU 上启用 cuda graph 和 EAGLE 投机解码的精度测试用例。
- `python/sglang/test/ascend/test_ascend_utils.py` (模块 测试工具; 类别 test; 类型 test-coverage): 新增变量 `MIMO_V2_FLASH_WEIGHTS_PATH`, 为测试提供模型权重路径。

关键符号：TestMiMoV2FlashGraphWithMTP

关键源码片段

[test/manual/ascend/llm_models/test_npu_mimo_v2_flash.py](#)

新增的手动测试文件，包含 MiMo-V2-Flash 模型在 NPU 上启用 cuda graph 和 EAGLE 投机解码的精度测试用例。

```
import unittest

from sglang.test.ascend.gsm8k_ascend_mixin import GSM8KAscendMixin
from sglang.test.ascend.test_ascend_utils import MIMO_V2_FLASH_WEIGHTS_PATH
from sglang.test.test_utils import CustomTestCase

class TestMiMoV2FlashGraphWithMTP(GSM8KAscendMixin, CustomTestCase):
    """Testcase: Verify the inference accuracy of MiMo-V2-Flash on GSM8K with cuda graph and
    MTP (speculative decoding).

    [Test Category] Model
    [Test Target] XiaomiMiMo/MiMo-V2-Flash
    [Test Config] Prefill+Decode, cuda graph enabled, EAGLE speculative decoding
    """

    # 模型权重路径，通过常量引入
    model = MIMO_V2_FLASH_WEIGHTS_PATH
    # GSM8K 准确率阈值，0.9 表示 >= 90% 通过
    accuracy = 0.9
    # 启动参数：启用 cuda graph、指定 ascend 后端、16 卡 TP、EAGLE 投机解码
    other_args = [
        "--trust-remote-code",
        "--mem-fraction-static",
        "0.8",
        "--attention-backend",
        "ascend",
        "--tp-size",
        "16",
        "--speculative-algorithm",
        "EAGLE",
        "--speculative-num-steps",
        "3",
        "--speculative-eagle-topk",
        "1",
        "--speculative-num-draft-tokens",
        "4",
        "--enable-multi-layer-eagle",
    ]

    if __name__ == "__main__":
```

unittest.main()

评论区精华

Review 中 Hexq0210 建议简化测试用例，只保留核心 Graph + MTP 特性。最终 patch 移除了 `_BASE_ARGS` 和 `_MTP_ARGS` 的分离，直接以 `other_args` 列表表达。

- 测试用例简化 (design): 最终实现将 `_BASE_ARGS` 和 `_MTP_ARGS` 合并为 `other_args` 列表，保持简洁。

风险与影响

- 风险：低风险。本次变更仅涉及测试文件新增和权重常量定义，不影响现有生产代码逻辑。但测试依赖特定权重路径，若权重目录不可用则测试无法通过。
- 影响：影响范围限于 Ascend NPU 手动测试流程。新增的测试为 MiMo-V2-Flash 模型的精度验证提供了入口，未来模型更新或配置变更需同步维护此测试。
- 风险标记：外部依赖权重路径

关联脉络

- 暂无明显关联 PR