PR #21383 完整报告

sgl-project/sglang

[diffusion] [NPU] support ring attention on NPU with FA

合并时间：2026-03-31 01:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21383

执行摘要

- 一句话：新增 NPU Ascend Flash Attention 后端，支持 ring attention。
- 推荐动作：建议技术管理者关注此 PR，以了解如何为不同平台添加定制 attention backend 的架构模式。工程师可精读 ascend_fa.py 中的实现，学习如何集成硬件专用操作并遵循抽象基类设计，以及通过 review 讨论了解代码优化点。

功能与动机

PR body 中说明：'ring attention requires return_softmax_lse but sdpa backend cannot support this option.' 因此，需要为 NPU 实现一个支持该选项的新 attention backend，使用 torch.ops.npu.npu_fused_infer_attention_score，作为 PR #20248 的另一种方案。

实现拆解

使用 Markdown 按 4 个步骤拆解实现过程：

1. 新增 AscendFA 后端实现：在 python/sglang/multimodal_gen/runtime/layers/attention/backends/ascend_fa.py 中创建 AscendFAMetadata、AscendFAMetadataBuilder、AscendFABackend 和 AscendFAImpl 类。关键符号 AscendFAImpl.forward 调用 torch.ops.npu.npu_fused_infer_attention_score 以支持 return_softmax_lse，为 ring attention 提供基础。
2. 集成到 NPU 平台：修改 python/sglang/multimodal_gen/runtime/platforms/npu.py 中的 get_attn_backend_cls_str 方法，当 selected_backend 为 AttentionBackendEnum.FA 时返回 AscendFABackend 类路径，实现后端选择逻辑。
3. 更新测试配套：在 python/sglang/multimodal_gen/test/server/ascend/perf_baselines_npu.json 中添加 qwen_image_t2i_2npu 用例的性能基准数据，并在 python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py 中添加相应测试配置，验证 ring attention 功能。
4. 文档同步：更新 docs/diffusion/performance/attention_backends.md，修正 NPU 平台对 FA 的支持描述和兼容性表格，确保用户文档准确。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/attention/backends/ascend_fa.py（模块 注意力后端；类别 source；类型 core-logic；符号 AscendFAMetadata, AscendFAMetadataBuilder, init, prepare）：新增 AscendFA 后端实现，是支持 ring

attention 的核心逻辑文件，包含关键类和 forward 方法。

- python/sglang/multimodal_gen/runtime/platforms/npu.py（模块 平台集成；类别 source；类型 entrypoint；符号 get_attn_backend_cls_str）：修改 NPU 平台的后端选择逻辑，集成 AscendFA 后端，是功能启用的关键入口。
- python/sglang/multimodal_gen/test/server/ascend/perf_baselines_npu.json（模块 性能测试；类别 test；类型 test-coverage）：新增性能基准数据用于测试 qwen_image_t2i_2npu 用例，确保新 backend 性能可验证。
- python/sglang/multimodal_gen/test/server/ascend/testcase_configs_npu.py（模块 测试配置；类别 test；类型 test-coverage）：新增测试用例配置，启用 ring attention 测试（ulysses_degree=1, ring_degree=2），验证功能。
- docs/diffusion/performance/attention_backends.md（模块 文档；类别 docs；类型 documentation）：更新文档以反映 NPU 对 FA 的支持，提升用户指南准确性。

关键符号：AscendFAImpl.forward, AscendFABackend.get_enum, AscendFABackend.get_metadata_cls, AscendFAMetadataBuilder.build, NPUPlatform.get_attn_backend_cls_str

关键源码片段

python/sglang/multimodal_gen/runtime/layers/attention/backends/ascend_fa.py

新增 AscendFA 后端实现，是支持 ring attention 的核心逻辑文件，包含关键类和 forward 方法。

```
from dataclasses import dataclass
from typing import Any
import torch
from sglang.multimodal_gen.runtime.layers.attention.backends.attention_backend import (
    AttentionBackend, AttentionImpl, AttentionMetadata, AttentionMetadataBuilder
)
from sglang.multimodal_gen.runtime.platforms import AttentionBackendEnum

@dataclass
class AscendFAMetadata:
    pass # 元数据类，当前无需额外字段

class AscendFAMetadataBuilder(AttentionMetadataBuilder):
    def __init__(self) -> None:
        pass
    def prepare(self) -> None:
        pass
    def build(self, **kwargs: Any) -> AttentionMetadata:
        return AscendFAMetadata() # 构建并返回元数据实例

class AscendFABackend(AttentionBackend):
    @staticmethod
```

```
    def get_enum() -> AttentionBackendEnum:
        return AttentionBackendEnum.FA # 返回后端枚举标识
    @staticmethod
    def get_impl_cls() -> type["AscendFAImpl"]:
        return AscendFAImpl # 返回实现类
    @staticmethod
    def get_metadata_cls() -> type["AttentionMetadata"]:
        return AscendFAMetadata # 实现元数据类返回，修复原 NotImplementedError
    @staticmethod
    def get_builder_cls() -> type["AttentionMetadataBuilder"]:
        return AscendFAMetadataBuilder # 返回构建器类

class AscendFAImpl(AttentionImpl):
    def __init__(self, num_heads: int, head_size: int, causal: bool, softmax_scale: float,
                 num_kv_heads: int | None = None, prefix: str = "", **extra_impl_args) -> None:
        self.causal = causal
        self.softmax_scale = softmax_scale
        self.num_heads = num_heads
        self.num_kv_heads = num_kv_heads or num_heads
        # 注意：head_size 和 prefix 参数未使用，保留以保持接口兼容性
    def forward(self, query: torch.Tensor, key: torch.Tensor, value: torch.Tensor,
                attn_metadata: AttentionMetadata, return_softmax_lse: bool = False) -> torch.
                Tensor:
        mask = None
        if self.causal:
            seq_len = query.shape[1]
            mask = torch.triu(torch.ones(seq_len, seq_len, device=query.device), diagonal=1).bool()
            # 构建因果掩码
        query = query.transpose(1, 2) # 转置为 BSHD 布局
        key = key.transpose(1, 2)
        value = value.transpose(1, 2)
        output, lse = torch.ops.npu.npu_fused_infer_attention_score(
            query, key, value, num_heads=self.num_heads,
            num_key_value_heads=self.num_kv_heads, scale=self.softmax_scale,
            input_layout="BNSD", softmax_lse_flag=return_softmax_lse, atten_mask=mask
        ) # 调用 NPU 专用 fused attention 操作，支持返回 softmax LSE
        output = output.transpose(1, 2) # 转置回 BSHD 布局
        if return_softmax_lse:
            return output, lse # 返回输出和 LSE，支持 ring attention
        return output
```

评论区精华

review 评论中，gemini-code-assist[bot] 指出了多个关键问题：AscendFABackend.get_metadata_cls 方法应实现而非抛出 NotImplementedError；文档措辞需更清晰以提升可读性；为保持代码一致性，建议将 metadata 类重命名为 AscendFAMetadata；修复类型提示和移除未使用参数（如 head_size、attn_metadata）。所有建议均被采纳，作者 Makcum888e 回应“done”，审核者 ping1jing2 批准，表明讨论已解决。

- 修复 NotImplementedError in get_metadata_cls (correctness): 作者 Makcum888e 修复为返回 AscendFAMetadata，审核者 ping1jing2 认可。
- 改进文档措辞清晰度 (documentation): 文档更新采纳建议，提升可读性。
- 代码一致性和未使用参数优化 (design): 作者实施重命名和参数清理，审核者批准。

风险与影响

- 风险：技术风险具体包括：新 backend 依赖 NPU 特定操作 torch.ops.npu.npu_fused_infer_attention_score，在其他平台不可用可能导致兼容性问题；AscendFAImpl.__init__ 中未使用的参数（如 head_size）可能影响未来扩展性和代码清晰度；测试覆盖虽新增性能基准，但新 backend 的完整功能验证依赖于 NPU 硬件环境，可能存在环境特定问题。
- 影响：对用户而言，现在可以在 NPU 上使用 ring attention，可能提升 diffusion 模型的推理性能和效率。系统层面，扩展了 attention backend 生态系统，增强了多平台支持，但增加了 NPU 特定代码的维护负担。团队需要确保新代码与现有 backend 接口兼容，并可能影响后续 NPU 相关开发。
- 风险标记：NPU 依赖风险，未使用参数，测试覆盖有限

关联脉络

- PR #20248 [需参考外部 PR，未在提供列表中]: PR body 提及本 PR 是 'another way for https://github.com/sgl-project/sglang/pull/20248'，表明关联于同一问题的不同解决方案。
- PR #22979 [Diffusion] [NPU] Fix multimodal gen CI: 同属 NPU 和 diffusion 模块，涉及 CI 和测试调整，与本 PR 的测试配套更新相关。
- PR #23041 [Docs] [npu] change the feature support status: 同属 NPU 文档更新，与本 PR 的文档修改类似，反映 NPU 功能支持演进。