

PR #26724 完整报告

sgl-project/sglang

[NPU] adaptation to support operator FA3 in deterministic inference on NPU.

合并时间: 2026-06-25 09:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/26724>

执行摘要

- 一句话: 在 NPU 上引入 FA3 算子支持确定性推理
- 推荐动作: 建议对 NPU 确定性推理感兴趣的读者仔细阅读此 PR, 尤其是 FA3 算子的集成方式和审查中的设计决策 (延迟导入、语义简化)。该 PR 为 NPU 后端提供了可选的确定性推理能力, 但需关注第三方依赖的稳定性。

功能与动机

需要在 Ascend NPU 上启用确定性推理以确保训练推理一致性, 当设置 `--enable-deterministic-inference` 参数且关闭图模式时, 保证多次运行输出一致。

实现拆解

实现按以下步骤进行:

1. 引入环境变量开关: 在 `AscendFlashAttentionBackend.__init__` 中添加 `self.use_fa = get_bool_env_var("ASCEND_USE_FA", "False")`, 默认关闭, 只有显式设置 `ASCEND_USE_FA=True` 才启用 FA3 算子。
2. `forward_extend` 分支: 在 `forward_extend` 方法的 `elif` 链中插入 `self.use_fa` 分支, 调用 `flash_attn_with_kvcache` 完成注意力计算。该分支需要构造 `cu_seqlens_q` 和 `max_seqlen_q` 参数, 并传递 `causal=True`、`window_size=[-1,-1]` 等参数。
3. `forward_decode` 分支: 在 `forward_decode` 方法中同样插入 `self.use_fa` 分支, 调用 `flash_attn_with_kvcache`, 参数相对简单。
4. 延迟导入依赖: 根据审查建议, 将 `from flash_attn_npu_v3 import flash_attn_with_kvcache` 移到各分支内部, 仅在运行时按需导入, 避免未安装第三方库时启动失败。
5. 使用简化语义: 根据审查建议, 使用 `self.forward_metadata.seq_lens` 替代复杂的 `seq_lens_cpu_int/seq_lens_cpu_list` 判断, 使代码更清晰。该 PR 仅修改一个文件, 未引入测试文件变更, 但 PR body 中提供了充分的确定性验证结果。

关键文件:

- `python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py` (模块 NPU 注意力; 类别 source; 类型 dependency-wiring): 唯一变更文件, 包含 FA3 算子支持的全部逻辑: 环境变量开关、`forward_extend` 和 `forward_decode` 新分支。

关键符号：forward_extend, forward_decode, init

关键源码片段

[python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py](#)

唯一变更文件，包含 FA3 算子支持的全部逻辑：环境变量开关、forward_extend 和 forward_decode 新分支。

```
# 在 __init__ 中获取环境变量，控制是否启用 FA3 后端
self.use_fa = get_bool_env_var("ASCEND_USE_FA", "False")

# forward_extend 中的 FA3 分支
elif self.use_fa:
    # 运行时导入，避免未安装该依赖时启动失败
    from flash_attn_npu_v3 import flash_attn_with_kvcache

    q = q.reshape(-1, layer.tp_q_head_num, layer.qk_head_dim)
    k = k_cache.view(-1, self.page_size, layer.tp_k_head_num, layer.qk_head_dim)
    v = v_cache.view(-1, self.page_size, layer.tp_v_head_num, layer.v_head_dim)
    extend_seq_lens = self.forward_metadata.extend_seq_lens_cpu_int.npu()
    cu_seqlens_q = torch.cat([
        torch.zeros(1, dtype=torch.int32).npu(),
        extend_seq_lens.cumsum(0).to(torch.int32),
    ])
    max_seqlen_q = extend_seq_lens.max().item()
    # 调用 FA3 的 flash_attn_with_kvcache 接口
    attn_output = flash_attn_with_kvcache(
        q, k, v,
        cache_seqlens=self.forward_metadata.seq_lens,
        page_table=self.forward_metadata.block_tables,
        cu_seqlens_q=cu_seqlens_q,
        max_seqlen_q=max_seqlen_q,
        softmax_scale=layer.scaling,
        causal=True,
        window_size=[-1, -1],
        softcap=0.0,
        rotary_interleaved=False,
        num_splits=0,
        sm_margin=0,
        return_softmax_lse=False,
    )
    attn_output = attn_output.view(-1, layer.tp_q_head_num * layer.v_head_dim)
```

评论区精华

审查中重点关注了两个问题：

- 审查者 Todobe 询问 FA3 算子是否需要额外安装包，建议在环境变量启用时才导入（评论："Does using this operator require installing additional packages? If so, it is

recommended to import it when the environment variable ASCEND_USE_FA is enabled.")。作者回复 "done", 最终实现将导入移到分支内部。

- 审查者 Todobe 在两个分支中建议检查是否可以使用 `self.forward_metadata.seq_lens` 替代额外的 CPU tensor 操作 (评论: "Check if `self.forward_metadata.seq_lens` can be used.")。作者均回复 "done", 最终代码采用了 `self.forward_metadata.seq_lens`。所有讨论已解决。
- flash-attention-npu 导入时机 (design): 已在最终实现中将 import 移到 elif 分支内部, 实现延迟导入。
- 使用 `self.forward_metadata.seq_lens` 简化代码 (correctness): 作者接受建议, 在最终代码中使用了 `self.forward_metadata.seq_lens`。

风险与影响

- 风险: 主要风险包括:
 1. 依赖第三方库: `flash_attn_npu_v3` 来自外部仓库 `flash-attention-npu`, 如果未安装, 运行时启用 `ASCEND_USE_FA` 会导致 `ImportError`。但延迟导入机制避免了未启用时的启动失败。
 2. 环境变量误用: 如果用户设定了 `ASCEND_USE_FA` 但未安装对应算子, 会导致运行时错误。应该在文档中明确说明依赖。
 3. 确定性保证: 仅针对非图模式, 且依赖 FA3 算子的 `batch-invariant` 实现。如果 FA3 版本更新引入不兼容, 可能破坏确定性。
 4. 性能影响: 添加的分支仅在启用时使用, 不影响现有路径性能。
 5. NPU 后端专属: 不影响 CUDA 等其他后端。 - 影响: 影响范围限定于使用 Ascend NPU 并且显式设置 `ASCEND_USE_FA=True` 的用户。其他用户和系统无影响。团队需要维护对 `flash_attn_npu_v3` 的依赖和兼容性。对确定性推理的需求 (如回归测试、训练推理一致性验证) 有正向作用。 - 风险标记: 依赖第三方库, 环境变量控制

关联脉络

- 暂无明显关联 PR