PR #21633 完整报告

sgl-project/sglang

[Diffusion][NPU] Add support for MOVA

合并时间：2026-04-03 10:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21633

执行摘要

- 一句话：为 MOVA 扩散模型添加 NPU 硬件支持，优化 RoPE 和内存格式兼容性。
- 推荐动作：建议涉及 NPU 开发或 Diffusion 模型的工程师精读此 PR，关注平台检测设计（如 current_platform.is_npu）、RoPE 优化策略（torch_npu.npu_rotary_mul 集成）以及如何处理硬件不兼容性（如 channels_last_3d 和 complex128 临时修复）。这些设计决策值得学习，以应对类似跨平台支持场景。

功能与动机

根据 PR body，动机是 'This PR adds NPU support to the MOVA model.'，并引用 Issue #18967 '[Roadmap] [NPU] Sglang Diffusion on Ascend'，作为 NPU 支持路线图的一部分，旨在扩展硬件覆盖并解决现有 NPU 平台上的执行失败问题。

实现拆解

实现分三部分：1) 在 python/sglang/jit_kernel/diffusion/triton/npu_fallback.py 中添加 NPU 特定的 Rotary Embedding 优化，引入 torch_npu.npu_rotary_mul 以提高性能，并设置常量限制；2) 在 python/sglang/multimodal_gen/runtime/models/dits/mova_video_dit.py 中修改 patchify 函数，根据 current_platform.is_npu 选择 contiguous 方式，避免 NPU 不支持的 torch.channels_last_3d；3) 在 python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/mova.py 中调整 visual_freqs 和 audio_freqs 准备逻辑，处理 NPU 上 torch.cat 对 torch.complex128 的不兼容，临时使用 torch.complex64。

关键文件：

- python/sglang/jit_kernel/diffusion/triton/npu_fallback.py（模块 JIT-kernel/Diffusion）：核心优化点，添加 NPU-specific Rotary Embedding fallback 使用 torch_npu.npu_rotary_mul，提高性能并设置常量限制。
- python/sglang/multimodal_gen/runtime/models/dits/mova_video_dit.py（模块 Diffusion/Models）：处理 NPU 上 channels_last_3d 不兼容，修改 patchify 函数以避免性能下降，影响模型前向逻辑。
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/mova.py（模块 Diffusion/Pipelines）：调整频率准备逻辑以处理 NPU 上 complex128 不兼容，关键推理步骤，涉及 visual 和 audio 频率张量处理。

关键符号：apply_rotary_embedding_native, patchify, inference_single_step

评论区精华

review 中核心讨论点：1) gemini-code-assist[bot] 建议提取共享函数以减少 visual_freqs 和 audio_freqs 准备的冗余代码，但未直接采纳；2) ping1jing2 和 Makcum888e 建议减少代码中的 is_npu 检查，推动改动到 npu_fallback.py 中，作者 LLThomas 响应并修改相关代码；3) Makcum888e 要求添加注释说明临时修复（如 'remove it when torch.cat on NPU is supported torch.complex128'），作者已添加；4) ping1jing2 关注 CI 触发问题，指出 Diffusion PR 意外触发 LLM CI，合并后计划优化 CI 工作流。讨论结论是问题已解决，代码改进并合并。

- 代码冗余和共享函数提取 (design): 代码已修改，但未实现共享函数，保持原有逻辑。
- 减少 is_npu 检查并集中优化 (design): 作者响应并修改了相关代码，推动优化到统一位置。
- 添加注释和代码一致性 (documentation): 作者添加了所需注释，提高了代码清晰度。
- CI 工作流触发问题 (infra): PR 已合并，但问题未完全解决，计划在后续 PR 中优化 CI 工作流。

风险与影响

- 风险：技术风险包括：1) NPU 特定优化（如 torch_npu.npu_rotary_mul）可能引入 GPU 兼容性问题，依赖条件判断准确性；2) 临时使用 complex64 处理 complex128 不兼容可能影响数值精度，需未来移除；3) 添加的平台检测代码增加维护负担，注释中 TODO 表示临时修复；4) 性能优化依赖外部 NPU 库版本，可能不稳定；5) CI 工作流因修改 jit_kernel 文件意外触发 LLM 测试，需后续优化。
- 影响：影响范围：1) 用户：MOVA 模型现支持 NPU 硬件，扩展了部署选项和硬件生态；2) 系统：Diffusion 模块增加 NPU 支持代码，提升硬件覆盖率，但引入平台特定逻辑，可能增加复杂度；3) 团队：推进 NPU 路线图，为后续 NPU 优化奠定基础，促进跨团队协作。影响程度中等，主要限于 MOVA 模型和 NPU 平台。
- 风险标记：平台特定代码，临时修复依赖，性能优化不确定性，CI 工作流影响

关联脉络

- PR #21408 [NPU] Support GLM-4.7-Flash on NPU: 同为 NPU 支持功能，涉及硬件后端和注意力适配，共享 NPU 优化策略。
- PR #21955 [diffusion] chore: fix stage profiler for multi-stage denoising: 同属 Diffusion 模块的 bugfix，相关性能分析和阶段计时，可能影响整体性能评估。
- PR #21922 Revert "Rollback flashmla to older version [1/2]": 涉及 JIT-kernel 和 CUDA 内核支持，与本 PR 的 jit-kernel 优化有技术关联。