PR #18617 完整报告

sgl-project/sglang

[NPU] GLM-5 optimize with fused kernels

合并时间：2026-03-30 22:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/18617

执行摘要

本 PR 通过引入 fused kernels 和缓存机制，优化了 GLM-5 模型在 NPU 硬件上的推理性能。关键改动包括旋转位置嵌入缓存、Triton 内核替换原生操作，以及量化权重处理，预计提升吞吐量和准确性。讨论中解决了线程安全问题，但遗留了硬编码值等维护性顾虑，整体为有意义的性能改进。

功能与动机

PR 动机明确为“优化 GLM-5 在 NPU 的推理性能”，旨在通过硬件特定优化减少计算开销，提升模型推理效率。背景可能是 NPU 平台上 DeepSeek 模型的性能瓶颈，基准测试数据显示输出 token 吞吐量可达 30.34 token/s，以验证优化效果。

实现拆解

- 模型配置模块：在 python/sglang/srt/configs/model_config.py 中添加 draft model 检查，避免量化配置错误。
- NPU attention 模块：在 python/sglang/srt/hardware_backend/npu/modules/deepseek_v2_attention_mla_npu.py 中引入 fused_split_qk_norm，根据 token 数量（<65535）选择优化路径，否则回退到原生操作。
- NSA indexer：在 python/sglang/srt/layers/attention/nsa/nsa_indexer.py 中将 sin/cos 缓存从全局变量迁移到 forward_batch.npu_indexer_sin_cos_cache，确保线程安全。
- Rotary embedding：在 python/sglang/srt/layers/rotary_embedding/base.py 中添加 fused_rope_qk_mqa 调用，对小规模输入使用融合内核，并引入 sin_cos_cache 缓存机制。
- DeepSeek 模型：在 python/sglang/srt/models/deepseek_nextn.py 中加载 rot_weight（从 rot.safetensors）并调整环境变量（如 SGLANG_DEEPEP_BF16_DISPATCH），以支持 NPU 推理。

评论区精华

review 中，gemini-code-assist[bot] 提出关键建议：

> “使用全局变量 SIN 和 COS 进行缓存不线程安全，可能导致 race conditions。建议缓存到 forward_batch 对象。”代码已采纳此建议，修改为使用 forward_batch.npu_indexer_sin_cos_cache。此外，评论还指出硬编码 epsilon 值和魔法数字问题：“eps 值硬编码为 1e-6，应使用 m.q_a_layernorm.variance_epsilon。”“魔法数字 65535 应替换为命名常

> 量。”这些建议在 patch 中未完全实现，可能存在维护风险。iforgetmyname 的 approval 表明变更整体可接受。

风险与影响

技术风险：新 fused kernels（如 fused_split_qk_norm 和 fused_rope_qk_mqa）可能未充分测试，存在回归风险；硬编码 epsilon 值和魔法数字影响代码维护性和跨配置兼容性；NPU 优化增加平台依赖，可能影响其他硬件支持。 影响：用户端 GLM-5 推理性能提升，基准测试显示输出 token 吞吐量改善；系统优化核心注意力路径，减少计算延迟；团队需维护 NPU 特定代码，增加复杂度，但提升了硬件适配能力。

关联脉络

与近期 PR 如 #21315（AMD fused rope 优化）和 #21255（NPU 接受率修复）相关联，显示跨硬件性能优化的趋势。结合 #21468（DeepSeek-V3.2 NPU 文档），此 PR 是 NPU 上 DeepSeek 模型性能改进的一部分，反映仓库在硬件特定优化和模型支持上的持续演进。提交历史中的多次合并（如解决冲突）表明这是协作迭代的结果。