PR #23190 完整报告

sgl-project/sglang

[NPU] add split_qkv_tp_rmsnorm_rope ops for minimax2 & fix eagle3 hidden states capture in dp attn mode

合并时间：2026-04-30 08:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23190

执行摘要

- 一句话：为 MiniMax-M2 添加 NPU 融合算子支持并修复 dp attention bug
- 推荐动作：值得精读 forward_prepare_npu 的实现和平台分支设计，可作为未来 NPU 适配其他模型时的参考模式。注意后续需补充 NPU 集成测试，并关注 sgl_kernel_npu 的版本更新。

功能与动机

根据 PR body，该变更是为了在 NPU 上支持 MiniMax2 模型（需要融合的 QKV 投影 + RMSNorm + RoPE 算子），并修复 cudagraph + eagle3 + dp attention 组合下 batch size > 1 时的崩溃错误。

实现拆解

1. 导入 NPU 相关模块：在 minimax_m2.py 中新增 is_npu 判断和条件导入 sgl_kernel_npu.norm.split_qkv_tp_rmsnorm_rope。
2. 新增 forward_prepare_npu 方法：使用融合算子 split_qkv_tp_rmsnorm_rope 一次完成 QKV 投影、QK RMSNorm 和 RoPE 的计算，减少显存访问和 kernel 启动开销；同时加入空 hidden_states 的短回路保护。
3. 修改 forward 方法：根据全局 _is_npu 标志分支到 forward_prepare_npu 或原有的 forward_prepare，确保 GPU 路径不受影响。
4. 修复 eagle3 隐藏状态捕获：在 attention layer 的 forward 方法中新增 captured_last_layer_outputs 参数，使 eagle3 能够在 dp attention 模式下正确传递隐藏状态。

关键文件：

- python/sglang/srt/models/minimax_m2.py（模块 模型层；类别 source；类型 core-logic, data-contract；符号 forward_prepare_npu, is_npu, split_qkv_tp_rmsnorm_rope, captured_last_layer_outputs）：唯一修改文件，包含 NPU 融合算子的条件导入、forward_prepare_npu 方法实现、forward 方法平台分支逻辑、eagle3 隐藏状态捕获修复以及空 short-circuit 保护。

关键符号：forward_prepare_npu, forward, forward_prepare

关键源码片段

python/sglang/srt/models/minimax_m2.py

唯一修改文件，包含 NPU 融合算子的条件导入、forward_prepare_npu 方法实现、forward 方法平台分支逻辑、eagle3 隐藏状态捕获修复以及空 short-circuit 保护。

defforward_prepare_npu(self, positions: torch.Tensor, hidden_states: torch.Tensor, forward_batch: ForwardBatch,): # 空 hidden_states 短回路保护：防止在分布式环境中因部分 rank 跳过 all-reduce 导致挂起 if hidden_states.shape[0] == 0: assert (not self.o_proj.reduce_results), "short-circuiting allreduce will lead to hangs" return hidden_states, forward_batch, None # 通过 qkv_proj 计算 QKV 联合投影 qkv, _ = self.qkv_proj(hidden_states) if self.use_qk_norm: # 使用 NPU 融合算子合并 QK RMSNorm + RoPE cos_sin = self.rotary_emb.cos_sin_cache.index_select(0, positions.flatten()) cos, sin = cos_sin.chunk(2, dim=-1) q, k, v = split_qkv_tp_rmsnorm_rope(input=qkv, cos=cos, sin=sin, q_weight=self.q_norm.weight, k_weight=self.k_norm.weight, q_hidden_size=self.q_size, kv_hidden_size=self.kv_size, head_dim=self.head_dim, rotary_dim=self.rotary_dim, eps=self.q_norm.variance_epsilon, tp_world=self.q_norm.attn_tp_size, tp_group=get_attention_tp_group().device_group,) else: # 非 qk_norm 场景：走常规 QKV 分割 + 独立 RoPE q, k, v = qkv.split([self.q_size, self.kv_size, self.kv_size], dim=-1) q, k = q.contiguous(), k.contiguous() q, k = self.rotary_emb(positions, q, k) # 将 QKV 状态打包为 inner_state，供后续 forward_core 使用 inner_state = q, k, v, forward_batch return None, forward_batch, inner_state （该代码块已按盘古排版规范添加注释，中文与英文标识符间保留空格）

评论区精华

审查中 gemini-code-assist[bot] 指出 forward_prepare_npu 缺少空 hidden_states 的短回路逻辑，可能在分布式环境中导致某些 rank 跳过 all-reduce 而其他 rank 等待，造成挂起。同时建议移除注释掉的旧代码。该问题在后续 commit（"short circuiting logix"）中已修复，最终合并前由 iforgetmyname 批准。

- 缺少空 hidden_states 短回路逻辑 (correctness): 在后续 commit（"short circuiting logix"）中已添加 short-circuit 逻辑，并移除注释掉的旧代码。

风险与影响

- 风险：新增的 NPU 路径仅在 _is_npu 为 True 时启用，不影响 GPU 现有路径，回归风险低。但 NPU 特定代码缺少单元测试覆盖，依赖外部 sgl_kernel_npu 库的正确性和兼容性。此外，forward_prepare_npu 中的 short-circuit 逻辑涉及 o_proj.reduce_results 断言，若未来修改该属性行为可能引入硬失败。

- 影响：对 GPU 用户无影响。对 NPU 用户，MiniMax-M2 模型获得原生推理支持，性能提升显著。修复的 *eagle3 + dp attention* bug 影响使用 *cudagraph* 和 batch size > 1 的 Multi-Token Prediction 场景。
- 风险标记：NPU 路径缺少测试覆盖，依赖外部 sgl_kernel_npu 库，核心路径条件分支

关联脉络

- 暂无明显关联 PR