PR #21408 完整报告

sgl-project/sglang

[NPU] Support GLM-4.7-Flash on NPU

合并时间：2026-04-02 17:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21408

执行摘要

本次 PR 实现了 GLM-4.7-Flash 模型在 NPU 硬件上的支持，通过添加注意力头填充和适配 rotary embedding，解决了模型头数不满足算子限制的问题，扩展了 SGLang 的硬件兼容性。变更涉及核心注意力后端逻辑，对使用 NPU 和该模型的用户有直接影响。

功能与动机

动机是支持 GLM-4.7-Flash 在 NPU 上运行，因为该模型有 20 个注意力头，而 NPU 的 FIA 算子目前只支持头数为 2 的幂，因此需要填充适配。PR body 中明确表述为 "Support GLM-4.7-Flash for NPU."。

实现拆解

主要改动分为两个部分：

1. ascend_backend.py：
   - 添加 q_head_num_padding 属性和 padding_size_list，计算所需填充大小。
   - 在 init_forward_metadata_capture_cuda_graph 中创建填充张量，动态获取模型 dtype。
   - 在 forward_extend 和 forward_decode_graph 中应用填充逻辑，使用 npu_fused_infer_attention_score 适配 qk_head_dim 等于 v_head_dim 的情况。
   - 代码片段示例：
2. rotary_embedding/base.py：
   - 修改 forward_npu 方法，扁平化输入张量以匹配算子要求，然后重塑回原始形状。
   - 代码片段示例：

评论区精华

review 讨论中的关键点：

- gemini-code-assist[bot]指出：

"The dtype for padding tensors is hardcoded to torch.bfloat16. This can cause dtype mismatches... This should be resolved by dynamically getting the model's data type from the model configuration."

- Hexq0210强调：

"take a look for gemini comments, Do not hardcode torch.bfloat16."

- Estrella-xx回复 "done"，表明已修复 dtype 问题。 此外，关于修改实例状态的建议未完全解决，但提供了代码改进方向。

风险与影响

风险：

- 填充逻辑引入额外计算和内存开销，可能影响性能。
- 若未来 NPU 算子支持非 2 的幂头数，代码需要更新以保持兼容性。
- rotary embedding 中张量重塑可能带来细微的正确性风险，但已通过测试验证。

影响：

- 用户：现在可以在 NPU 上运行 GLM-4.7-Flash 模型，扩展了使用场景。
- 系统：增强了硬件后端的模型支持能力，但增加了维护复杂性。
- 团队：需要关注 NPU 特定代码的测试和未来优化。

关联脉络

从近期历史 PR 看，本 PR 与以下 PR 相关：

- PR 21914：设置 TRTLLM 内核为 Blackwell 默认，同属硬件后端优化类别。
- PR 20394：启用 FP8 MoE 以提升性能，涉及模型特定适配。 这些关联表明仓库在持续扩展硬件兼容性和性能优化，本 PR 是 NPU 支持演进的一部分。