PR #27887 完整报告

sgl-project/sglang

[Model] Add HrmTextForCausalLM (Hierarchical Reasoning Model - Text)

合并时间：2026-06-30 13:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27887

执行摘要

- 一句话：添加 HRM-Text 层次循环推理模型
- 推荐动作：值得精读。本 PR 展示了如何集成一个具有特殊循环结构和 bidirectional attention 的模型到 SGLang，尤其是在缺少原生支持的情况下的妥协方案。对于关注推理框架扩展性的开发者有学习价值。

功能与动机

HRM-Text 是 HuggingFace transformers 5.9.0 中新增的层次循环推理模型。本 PR 将 SGLang 中已有的 vLLM 实现镜像移植，使得 SGLang 用户能够推理该模型。

实现拆解

1. 新增模型文件hrm_text.py：实现 HrmTextForCausalLM，包含 HrmTextStack（嵌套 H/L 循环）、HrmTextDecoderLayer、HrmTextAttention（支持 sigmoid gating 和融合 gqkv 投影）和 HrmTextMLP。每个循环步的注意力层使用唯一的 RadixAttention(layer_id) 分配独立 KV 槽。权重加载通过 MergedColumnParallelLinear 的 fused-on-disk 路径处理融合的 gqkv_proj 和 gate_up_proj。
2. 调整 KV 槽计数model_config.py：在 num_attention_layers 计算中添加 HrmTextForCausalLM 分支，使用显式公式 num_layers_per_stack * H_cycles * (L_cycles + 1) 保证无论配置来源均分配正确的注意力层数。
3. 运行时强制约束model_runner.py：在 model_specific_adjustment 中检测 HRM-Text 后强制设置 attention_backend=triton、chunked_prefill_size=-1、disable_radix_cache=True 和 disable_cuda_graph=True，确保双向 prefix attention 正确执行。
4. 版本检查model_config.py 的 _verify_transformers_version：检测到 HRM-Text 架构时要求 transformers >= 5.9.0，否则抛出明确错误，防止老版本静默加载错误权重。

关键文件：

- python/sglang/srt/models/hrm_text.py（模块 模型实现；类别 source；类型 core-logic；符号 _num_layers_per_stack, _steps_used, HrmTextMLP, HrmTextAttention）：新增模型核心文件，包含全部前向逻辑和权重加载。
- python/sglang/srt/configs/model_config.py（模块 配置；类别 source；类型 data-contract；符号 _derive_model_shapes, _verify_transformers_version）：修改

num_attention_layers 计算，确保 HRM-Text 的注意力层数（KV 槽数）正确；添加 transformers 版本检查。

- python/sglang/srt/model_executor/model_runner.py（模块 运行时；类别 source；类型 data-contract；符号 model_specific_adjustment）：在 model_specific_adjustment 中强制 HRM-Text 使用 Triton 后端并关闭不兼容优化。

关键符号：_num_layers_per_stack, _steps_used, HrmTextMLP, HrmTextAttention, HrmTextDecoderLayer, HrmTextStack, HrmTextForCausalLM, model_specific_adjustment

关键源码片段

python/sglang/srt/models/hrm_text.py

新增模型核心文件，包含全部前向逻辑和权重加载。

```
# 根据配置推导单个栈（H 或 L）的层数
# 原生 config 使用 num_layers_per_stack，否则从总层数反推

def _num_layers_per_stack(config: PretrainedConfig) -> int:
    nlps = getattr(config, "num_layers_per_stack", None)
    if nlps is not None:
        return int(nlps)
    # 公式：total = per_stack * H_cycles * (L_cycles + 1)
    return config.num_hidden_layers // (config.H_cycles * (config.L_cycles + 1))


# 返回栈在全部循环步骤中被激活的下标列表
# L 栈在 h*(L+1) + l 步执行，H 栈在 h*(L+1) + L 步执行

def _steps_used(config: PretrainedConfig, stack_kind: str) -> list[int]:
    H_cycles = config.H_cycles
    L_cycles = config.L_cycles
    if stack_kind == "L":
        return [
            h * (L_cycles + 1) + l
            for h in range(H_cycles)
            for l in range(L_cycles)
        ]
    # stack_kind == "H"
    return [h * (L_cycles + 1) + L_cycles for h in range(H_cycles)]
```

python/sglang/srt/model_executor/model_runner.py

在 model_specific_adjustment 中强制 HRM-Text 使用 Triton 后端并关闭不兼容优化。

```
# HRM-Text 需要双向 prompt attention，仅 Triton 后端支持且必须
# 关闭 cuda graph / chunked prefill / radix cache
hf_config = self.model_config.hf_config
is_hrm_text = getattr(hf_config, "model_type", None) == "hrm_text" or \
              "HrmTextForCausalLM" in getattr(hf_config, "architectures", [])
is_prefix_lm_recurrent = is_hrm_text and getattr(hf_config, "prefix_lm", True)
```

```
if is_prefix_lm_recurrent:
    if server_args.attention_backend not in (None, "triton"):
        logger.warning(...)
    server_args.attention_backend = "triton"
    server_args.chunked_prefill_size = -1
    server_args.disable_radix_cache = True
    server_args.disable_cuda_graph = True
    logger.warning(...)
```

评论区精华

PR body 提出了两个开放设计问题：

1. 双向 attention 后端局限：AttentionType.DECODER_BIDIRECTIONAL 仅 Triton 后端实现，FlashInfer/FA3 不识别，导致模型被绑定到 Triton。评审者 JustinTong0323 确认这是当前限制，并指出 vLLM 使用 PrefillPrefixLMAttention 层实现后端无关的方案。
2. 强制全局 flag 的折中：在 model_specific_adjustment 中全局关闭 chunked prefill、radix cache 和 cuda graph 被视为重手段。评审者表示目前 SGLang 缺乏优雅的 per-layer hook，多模态模型也采用类似方式，暂时接受现方案。

- 双向 attention 后端支持局限 (design): 评审者表示当前确实受限，vLLM 有 PrefillPrefixLMAttention 层更优雅，但 SGLang 尚未实现，暂时维持现状。
- 强制全局 flag 的折中 (design): 评审者确认目前多模态模型也采用同样模式，暂时接受此做法。

风险与影响

- 风险：
  1. 性能下降：强制关闭 chunked prefill 和 cuda graph 可能显著降低吞吐和延迟，影响 HRM-Text 用户的体验。
  2. 后端锁定：模型只能在 Triton 后端运行，无法利用 FlashInfer 或 FA3 的优化，可能成为长文本场景的瓶颈。
  3. Transformers 版本依赖：老旧 transformers 静默失败被转换为显式错误，但用户必须升级才能使用。
  4. 缺少测试：PR 未包含单元测试或集成测试文件，仅依靠手动验证 GSM8K，未来回归风险较高。 - 影响：用户端：SGLang 现在可以加载和推理 HRM-Text 系列模型（需 transformers >= 5.9.0）。系统端：引入了一个新的模型架构代码路径，增加了维护负担；强制使用 Triton 后端可能与其他系统配置冲突。团队端：需要关注后续双向 attention 后端无关化的进展，以便移除强制限制。 - 风险标记：Transformers 版本依赖，强制后端限制，缺少测试覆盖，双向注意力兼容性限制

关联脉络

- 暂无明显关联 PR