PR #24826 完整报告

sgl-project/sglang

[spec decoding] support kimi-k2.5-eagle3-mla

合并时间：2026-05-10 14:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24826

执行摘要

- 一句话：支持 Kimi-K2.5 EAGLE3 MLA 推测解码模型
- 推荐动作：值得精读 EAGLE3 与 MLA 结合的设计，特别是如何通过替换 fused_qkv_a_proj_with_mqa 来适配拼接输入。建议后续补充单元测试覆盖前向逻辑与权重加载。

功能与动机

支持 Kimi-K2.5 模型的 EAGLE3 推测解码，利用 MLA 注意力缩小草稿模型的 KV 缓存占用，提升推理效率。PR 描述中提到 checkpoint 格式为 kimi-k2.5-eagle3-mla，旨在保持与目标模型一致的 MLA 布局。

实现拆解

1. 新增 EAGLE3+MLA 草稿模型：创建 kimi_k25_eagle3.py，包含 Eagle3MLADecoderLayer（复用 DeepseekV2AttentionMLA，但将 fused QKV 投影输入维度翻倍以处理拼接后的 [embed_norm, hidden_norm]）和 Eagle3DeepseekV2ForCausalLM 顶层模型类（继承自 DeepseekV2ForCausalLM 的 embed/head，注册为 Eagle3DeepseekV2ForCausalLM architecture）。
2. 配置注册：在 common.py 中添加 _KimiK2ConfigAlias，继承自 DeepseekV3Config，model_type = 'kimi_k2'，并注册到 _CONFIG_REGISTRY，使得加载权重时能正确识别配置。
3. 架构识别：在 model_config.py 的 MLA 架构判断条件中增加 'Eagle3DeepseekV2ForCausalLM'，保证其视为 MLA 架构并设置 head_dim=256、kv_lora_rank 等参数。
4. 测试与验证：PR 作者报告了在 GPQA 上的精度测试结果（accuracy 0.87, acc len 2.7），但未提交单元测试文件。

关键文件：

- python/sglang/srt/models/kimi_k25_eagle3.py（模块 新模型；类别 source；类型 core-logic；符号 _get_eagle_aux_layer_count, Eagle3MLADecoderLayer, init, forward）：核心新增文件，包含 EAGLE3 草稿层和顶层模型类，实现 MLA 投影替换的关键逻辑。
- python/sglang/srt/utils/hf_transformers/common.py（模块 配置注册；类别 source；类型 core-logic；符号 _KimiK2ConfigAlias）：注册 _KimiK2ConfigAlias 使模型配置可识别

，确保加载权重时指向正确的 config 类。

- python/sglang/srt/configs/model_config.py（模块 模型配置；类别 source；类型 data-contract）：将 Eagle3DeepseekV2ForCausalLM 加入 MLA 架构判断，保证注意力维度正确。

关键符号：_get_eagle_aux_layer_count, Eagle3MLADecoderLayer.init, Eagle3MLADecoderLayer.forward, Eagle3MLAModel.init, Eagle3DeepseekV2ForCausalLM.get_input_embeddings, Eagle3DeepseekV2ForCausalLM.get_embed_and_head, _KimiK2ConfigAlias

关键源码片段

python/sglang/srt/models/kimi_k25_eagle3.py

核心新增文件，包含 EAGLE3 草稿层和顶层模型类，实现 MLA 投影替换的关键逻辑。

```
class Eagle3MLADecoderLayer(nn.Module):
    """
    One EAGLE3 draft layer that uses DeepSeek-V2 multi-latent attention.
    Pre-attention concatenates the input embedding and the target hidden
    state along the channel dim, doubling the input width to MLA's fused
    QKV-down projection.
    """
    def __init__(
        self,
        config: PretrainedConfig,
        layer_id: int = 0,
        quant_config: Optional[QuantizationConfig] = None,
        prefix: str = "",
    ) -> None:
        super().__init__()
        self.hidden_size = config.hidden_size
        # 省略 self.self_attn = DeepseekV2AttentionMLA(...) 初始化

        # EAGLE3 将 embed_norm 和 hidden_norm 在特征维拼接，
        # 输入宽度变为 2 * hidden_size，因此需要替换 fused QKV 投影层
        attn = self.self_attn
        if attn.q_lora_rank is None:
            raise ValueError(
                "Eagle3 MLA layer requires q_lora_rank in the draft config"
            )
        # 使用双倍输入大小的投影层替换原有单输入投影
        attn.fused_qkv_a_proj_with_mqa = ReplicatedLinear(
            2 * config.hidden_size,
            attn.q_lora_rank + attn.kv_lora_rank + attn.qk_rope_head_dim,
            bias=False,
            quant_config=quant_config,
            prefix=add_prefix("self_attn.fused_qkv_a_proj_with_mqa", prefix),
        )
```

```
# 更新融合投影标志，禁用最小延迟模式
attn.has_fused_proj = True
attn.use_min_latency_fused_a_gemm = False
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 缺少单元测试：新模型文件 kimi_k25_eagle3.py 无对应测试用例，若后续重构可能引入回归。
  - 依赖现有 MLA 实现：直接复用 DeepseekV2AttentionMLA 并替换投影，任何对 MLA 层的改动可能影响此模型。
  - 配置兼容性：假设 q_lora_rank 在草稿配置中存在，否则抛出异常；若 checkpoint 格式稍有差异可能导致加载失败。
  - 领域特定性：仅适用于 Kimi-K2.5 EAGLE3 MLA checkpoint，普通模型不受影响。
- 影响：
  - 用户影响：启用 Kimi-K2.5 EAGLE3 推测解码的用户可直接使用新模型，提升推理速度；其他模型无影响。
  - 系统影响：新增模型类注册在 AutoModel 中，不干扰其他模型；配置文件别名的添加仅当试图加载 kimi_k2 配置时生效。
  - 团队影响：后续维护者需关注 MLA 相关改动对该草稿模型的兼容性。
  - 风险标记：缺少单元测试，依赖现有 MLA 实现，无回归测试覆盖

关联脉络

- PR #23976 Support Gemma3/4 + Eagle3: 该 PR 为 EAGLE3 模型支持奠定了基础（包括多层捕获、TP 分片等），本 PR 在此基础上针对 Kimi-K2.5 MLA 进行适配。