PR #47448 完整报告

vllm-project/vllm

[Bugfix] Recycle post-final-norm hidden in GLM MTP (single norm)

合并时间：2026-07-06 16:07

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47448

执行摘要

- 一句话：修复 GLM MTP 回收错误 hidden states
- 推荐动作：值得精读，特别是关注 MTP 实现中 hidden states 回收与 norm 应用一致性的设计原则。

功能与动机

DeepSeekV32 MTP draft 层原先返回 pre-final-norm hidden 作为回收的 previous_hidden_states，与 draft 模型的 hnorm 不匹配，导致 MTP acceptance 下降。参考 deepseek_mtp.py (PR #45895) 的做法，应回收 post-final-norm hidden。

实现拆解

1. 修改 DeepseekV32MultiTokenPredictorLayer.forward 返回值：在 tensor_model_parallel_all_reduce 后，调用 self.shared_head.norm(hidden_states, residual) 计算 post-final-norm hidden，并将 (hidden_states, hidden_states) 以元组形式返回。
2. 简化 compute_logits：由于 forward 已返回 post-final-norm hidden，compute_logits 直接使用该 hidden 通过 logits_processor(mtp_layer.shared_head.head, hidden_states) 计算 logits，不再重复调用 mtp_layer.shared_head(hidden_states)。
3. 保持兼容性：元组返回被 V2 speculator（通过 isinstance-tuple 检查）和 legacy proposer（model_returns_tuple 为 True）识别。

关键文件：

- vllm/models/deepseek_v32/nvidia/mtp.py（模块 模型；类别 source；类型 data-contract；符号 DeepseekV32MultiTokenPredictorLayer.forward, DeepseekV32MultiTokenPredictor.compute_logits）：核心修复文件，修改了 draft 层 forward 返回值和 compute_logits 逻辑。

关键符号：DeepseekV32MultiTokenPredictorLayer.forward, DeepseekV32MultiTokenPredictor.compute_logits

关键源码片段

vllm/models/deepseek_v32/nvidia/mtp.py

核心修复文件，修改了 draft 层 forward 返回值和 compute_logits 逻辑。

```
# vllm/models/deepseek_v32/nvidia/mtp.py
# 关键修改：forward 返回 post-final-norm hidden
class DeepseekV32MultiTokenPredictorLayer(nn.Module):
    # ...
    def forward(
        self,
        input_ids: torch.Tensor,
        positions: torch.Tensor,
        previous_hidden_states: torch.Tensor,
        inputs_embeds: torch.Tensor | None = None,
        spec_step_index: int = 0,
    ) -> torch.Tensor:
        assert inputs_embeds is not None
        # 融合：位置 0 的 embeds 归零 + enorm(embeds) + hnorm(prev) + concat -> [N, 2H]
        eh_input = fused_eh_norm(
            positions,
            inputs_embeds,
            previous_hidden_states,
            self.enorm.weight,
            self.hnorm.weight,
            self.enorm.variance_epsilon,
        )
        hidden_states = self.eh_proj(eh_input)
        hidden_states, residual = self.mtp_block(
            positions=positions, hidden_states=hidden_states, residual=None
        )
        # mtp_block 的 MoE 输出未被 all-reduce（skip_final_all_reduce）
        # 主模型会将 all-reduce 融合到下一个 norm，但这里 recycle hidden 直接消费
        hidden_states = tensor_model_parallel_all_reduce(hidden_states)
        # 关键修复：将 residual 融合到最终 RMSNorm，计算 post-final-norm hidden
        # 原先返回 residual + hidden_states（pre-final-norm），导致 hnorm 不匹配
        hidden_states, _ = self.shared_head.norm(hidden_states, residual)
        # 返回元组 (hidden_states, hidden_states) 作为 draft-logits hidden 和 recycled previous_
        hidden_states
        # 该元组形式兼容 V2 speculator（isinstance-tuple 检查）和 legacy proposer
        return hidden_states, hidden_states

# vllm/models/deepseek_v32/nvidia/mtp.py
# compute_logits 简化：hidden_states 已经是 post-final-norm，直接应用 LM head
class DeepseekV32MultiTokenPredictor(nn.Module):
    # ...
    def compute_logits(
        self,
        hidden_states: torch.Tensor,
        spec_step_idx: int = 0,
    ) -> torch.Tensor:
        current_step_idx = spec_step_idx % self.num_mtp_layers
        mtp_layer = self.layers[str(self.mtp_start_layer_idx + current_step_idx)]
```

```
# hidden_states 已在 layer.forward 中经过 post-final-norm，不再重复应用 shared_head
return self.logits_processor(mtp_layer.shared_head.head, hidden_states)
```

评论区精华

review 评论仅由 claude[bot] 自动评论，无人工讨论。PR 由 WoosukKwon 直接批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，变更仅涉及单一文件的 forward 和 compute_logits 方法，且与参考实现 (deepseek_mtp.py) 行为对齐。需注意：如果其他模块依赖 pre-final-norm hidden 的原始返回，可能需要调整；但 PR 说明 tuple 形式已被 V2 和 legacy proposer 兼容。
- 影响：影响范围限定在 DeepSeekV32 MTP 的 draft 模型生成流程，具体修复 MTP 验收率问题。对用户而言，使用 DeepSeekV32 模型进行 speculative decoding 时，MTP draft 的接受率将提升，推理速度受益。系统其他部分不受影响。
- 风险标记：暂无

关联脉络

- PR #45895 [Core] Support Multi-Token Prediction (MTP) for DeepSeek V3 / GLM (compile-based): 参考实现，deepseek_mtp.py 中已采用 post-final-norm 回收 hidden states