PR #46986 完整报告

vllm-project/vllm

[Bugfix] Fix DeepseekV2Model hidden_size

合并时间：2026-06-30 00:44

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46986

执行摘要

- 一句话：修复 DeepseekV2Model hidden_size 引用错误
- 推荐动作：值得合并，属于规范化修复，提升代码一致性和可维护性。建议后续考虑在相似模型中推广此模式。

功能与动机

DeepseekV2Model 的 __init__中多处直接使用 config.hidden_size，但未显式保存为 self.hidden_size，导致子类或后续逻辑若覆盖 hidden_size 时引用不一致。PR body 引用 #46983 说明此问题。

实现拆解

1. 在 DeepseekV2Model.__init__ 中增加 self.hidden_size = config.hidden_size，将 hidden_size 存储为实例属性。
2. 将 embed_tokens 的构造参数从 config.hidden_size 改为 self.hidden_size。
3. 将 norm（RMSNorm）的输入维度从 config.hidden_size 改为 self.hidden_size。
4. 将 make_empty_intermediate_tensors 调用中的 hidden_size 参数从 config.hidden_size 改为 self.hidden_size。 以上变更仅涉及 vllm/model_executor/models/deepseek_v2.py，共 +4/-4 行，无测试或配置配套改动。

关键文件：

- vllm/model_executor/models/deepseek_v2.py（模块 模型层；类别 source；类型 data-contract；符号 DeepseekV2Model.init）：唯一变更文件，修复 DeepseekV2Model 中 hidden_size 的引用方式。

关键符号：DeepseekV2Model.init

关键源码片段

vllm/model_executor/models/deepseek_v2.py

唯一变更文件，修复 DeepseekV2Model 中 hidden_size 的引用方式。

```
class DeepseekV2Model(nn.Module):
    def __init__(self, *, vllm_config: VllmConfig, prefix: str = ""):
        super().__init__()
```

```
config = vllm_config.model_config.hf_config
quant_config = vllm_config.quant_config
self.config = config
self.device = current_platform.device_type
# 新增：将 hidden_size 存储为实例属性，避免直接引用 config
self.hidden_size = config.hidden_size
self.vocab_size = config.vocab_size
self.is_v32 = hasattr(config, "index_topk")
if self.is_v32:
    topk_tokens = config.index_topk
    topk_indices_buffer = torch.empty(
        vllm_config.scheduler_config.max_num_batched_tokens,
        topk_tokens,
        dtype=torch.int32,
        device=self.device,
    )
else:
    topk_indices_buffer = None

if get_pp_group().is_first_rank:
    # 使用 self.hidden_size 替代 config.hidden_size
    self.embed_tokens = VocabParallelEmbedding(
        config.vocab_size,
        self.hidden_size,
        quant_config=quant_config,
        prefix=f"{prefix}.embed_tokens",
    )
else:
    self.embed_tokens = PPMissingLayer()
# ... layers creation ...

if get_pp_group().is_last_rank:
    # 使用 self.hidden_size
    self.norm = RMSNorm(self.hidden_size, eps=config.rms_norm_eps)
else:
    self.norm = PPMissingLayer()
self.make_empty_intermediate_tensors = make_empty_intermediate_tensors_factory(
    ["hidden_states", "residual"], self.hidden_size
)
# ... other attributes ...
```

评论区精华

Review 中 cjackal 提出了一个建议，指出差分中 `self.hidden_size = self.hidden_size` 应改为 `self.hidden_size = config.hidden_size`，但该评论最终未被采纳（实际提交已修正为正确赋值）。此外，yewentao256 批准了该 PR，无其他讨论。

- self.hidden_size 赋值错误 (correctness): 该评论在后续提交中已修正为正确赋值，PR 已合并。

风险与影响

- 风险：风险极低：变更仅将 config.hidden_size 的引用方式改为通过 self.hidden_size 间接引用，值完全一致，无逻辑行为改变。但若子类覆盖了 hidden_size 且未正确处理，可能引入不一致，但当前代码无子类覆盖此属性。
- 影响：直接影响 DeepseekV2 系列模型的初始化流程，但无功能变化。对用户透明，不改变加载行为或推理结果。
- 风险标记：低风险，无测试覆盖

关联脉络

- PR #46983 关联 issue 或 PR（未明确）：PR body 中引用了该链接，但内容未提供，可能为同一问题的根源讨论。
- PR #46973 [Bugfix] Capture final-layer aux hidden state in deepseek_v2 backbone: 最近对 deepseek_v2 模型的另一个 bugfix，涉及类似文件，表明 deepseek 模型相关修复较多。