PR #43926 完整报告

vllm-project/vllm

fix: keep DeepSeek V4 RoPE cache on inv_freq device

合并时间：2026-06-05 06:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43926

执行摘要

- 一句话：修复 DeepSeek V4 RoPE 缓存在 meta device 上的构造失败
- 推荐动作：该 PR 是经典的一行 bugfix，虽小但修复了深层 device 一致性问题。值得精读，尤其是理解 torch.device 在 meta device 场景下的传播模式。

功能与动机

修复 DeepSeek V4 模型在 meta device 下构造时的崩溃。作者提供了可复现脚本，并在 PR body 中给出了完整 traceback：RuntimeError: Tensor on device meta is not on the expected device cpu!。该问题影响使用 meta device 进行模型初始化的场景（如分布式加载、模型预构建等）。

实现拆解

在 vllm/model_executor/layers/rotary_embedding/deepseek_scaling_rope.py 的 DeepseekV4ScalingRotaryEmbedding._compute_cos_sin_cache() 方法中，将创建位置张量 t 时的 device 参数从 current_platform.device_type 改为 inv_freq.device。这样当 inv_freq 位于 meta device 时，t 也会创建在 meta device 上，从而避免 torch.einsum 出现 device 不匹配错误。

关键文件：

- vllm/model_executor/layers/rotary_embedding/deepseek_scaling_rope.py（模块 嵌入层；类别 source；类型 data-contract；符号 DeepseekV4ScalingRotaryEmbedding._compute_cos_sin_cache）：修改了 DeepseekV4ScalingRotaryEmbedding._compute_cos_sin_cache() 中的 device 参数，确保位置张量 t 与 inv_freq 位于同一设备，修复 meta device 下构造崩溃。

关键符号：DeepseekV4ScalingRotaryEmbedding._compute_cos_sin_cache

关键源码片段

vllm/model_executor/layers/rotary_embedding/deepseek_scaling_rope.py

修改了 DeepseekV4ScalingRotaryEmbedding._compute_cos_sin_cache() 中的 device 参数，确保位置张量 t 与 inv_freq 位于同一设备，修复 meta device 下构造崩溃。

```
# file: vllm/model_executor/layers/rotary_embedding/deepseek_scaling_rope.py
```

```
class DeepseekV4ScalingRotaryEmbedding(DeepseekScalingRotaryEmbedding):
    # ... (docstring omitted)

    def _compute_cos_sin_cache(self) -> torch.Tensor:
        inv_freq = self._compute_inv_freq(self.scaling_factor) # 可能位于 meta device
        t = torch.arange(
            self.max_position_embeddings * self.scaling_factor,
            device=inv_freq.device, # 修复前为 current_platform.device_type，导致 device 不匹配
            dtype=torch.float32,
        )
        freqs = torch.einsum("i,j -> ij", t, inv_freq) # 此时 device 一致
        cos = freqs.cos() * self.mscale
        sin = freqs.sin() * self.mscale
        cache = torch.cat((cos, sin), dim=-1)
        return cache
```

评论区精华

评审人 AndreasKaratzas 最初建议添加回归测试，但作者认为测试代价低且不必要，最终双方同意只保留一行注释。后续 zyongye 要求删除注释，作者照做，最终 zyongye 批准该 PR。

- 是否添加回归测试 (testing): 不添加测试，仅保留注释。
- 评论内容清理 (style): 删除注释，仅保持代码变更。

风险与影响

- 风险：变更仅一行代码，将 device=current_platform.device_type 改为 device=inv_freq.device，逻辑等价且更灵活。因 inv_freq 通常位于 CPU 或 meta device，不会引入性能退化。极低风险。
- 影响：仅影响 DeepSeek V4 模型在 meta device 上的 RoPE 缓存构造。修复后，用户可以在 meta device 下成功初始化该模型，后续加载真实权重时正常工作。不影响其他模型或运行时路径。
- 风险标记：暂无

关联脉络

- PR #43827 [DSv4] Adding TRTLLM gen attention kernel: 同属于 DeepSeek V4 模型支持系列，该 PR 为新特性，本 PR 修复了其中的 device 一致性问题。