PR #47308 完整报告

vllm-project/vllm

[ModelRunner V2] Warmup cross-attn properly in encoder-decoder case

合并时间：2026-07-02 03:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47308

执行摘要

- 一句话：修复 encoder-decoder 模型预热时 cross-attn 零 key 问题
- 推荐动作：值得精读，特别是理解 encoder-decoder 模型在预热中的特殊处理方式。设计简洁，对理解 vLLM 的 warmup 机制和 cross-attention 集成有参考价值。

功能与动机

PR body 明确指出 'E.g. for whisper. Existing warmup was resulting in zero-key attention op.'，即现有预热流程对 encoder-decoder 模型生成了零 key attention，需要修复。

实现拆解

1. 新增导入(vllm/v1/worker/gpu/warmup.py)：添加 MultiModalFeatureSpec、PlaceholderRange 和 CrossAttentionSpec 的导入，为 dummy encoder 输入和跨注意力 KV 缓存计算提供支持。
2. 生成 dummy encoder 输入：在 warmup_kernels 函数中，通过检查 model_runner.is_encoder_decoder 和 max_encoder_len 属性，构造一个 warmup_mm_features 列表，包含一个 MultiModalFeatureSpec 对象，其 data=None，仅设置占位符 PlaceholderRange。该 dummy 特征会注册到 encoder 缓存中，但不会真正调度 encoder。
3. 修正 cross-attention 块计数：在 _warmup_block_count 内部函数中，当 spec 为 CrossAttentionSpec 时，将 num_tokens 替换为 max_encoder_len，确保 cross-attention KV 缓存分配足够的块。
4. 传递 dummy 特征：在构建 NewRequestData 时，将 warmup_mm_features 作为 mm_features 参数传入 Request 对象，使预热请求携带 encoder 输入。

关键文件：

- vllm/v1/worker/gpu/warmup.py（模块 预热；类别 source；类型 core-logic；符号 warmup_kernels, _warmup_block_count）：在 warmup_kernels 中添加 encoder-decoder 模型的 dummy encoder 输入处理，并修正 cross-attention KV 缓存块计数。

关键符号：warmup_kernels, _warmup_block_count

关键源码片段

vllm/v1/worker/gpu/warmup.py

在 warmup_kernels 中添加 encoder-decoder 模型的 dummy encoder 输入处理，并修正 cross-attention KV 缓存块计数。

```
# vllm/v1/worker/gpu/warmup.py 中的关键修改片段

@torch.inference_mode()
def warmup_kernels(...):
    ...
    kv_cache_groups = model_runner.kv_cache_config.kv_cache_groups
    num_kv_cache_groups = len(kv_cache_groups)

    # 新增：Encoder-decoder 模型的预热处理
    # 给每个请求一个 dummy encoder 输入，使 cross-attention
    # 能在非空 key 序列上预热，避免零 key attention 操作。
    # 该 dummy mm_feature 会注册到 encoder 缓存中，但仅读取
    # encoder 长度，不会实际调度 encoder。
    max_encoder_len = getattr(model_runner.model_state, "max_encoder_len", 0)
    warmup_mm_features: list[MultiModalFeatureSpec] = []
    if model_runner.is_encoder_decoder and max_encoder_len:
        warmup_mm_features = [
            MultiModalFeatureSpec(
                data=None,
                modality="",
                identifier="_warmup_encoder",
                mm_position=PlaceholderRange(offset=0, length=max_encoder_len),
            )
        ]

    # 修正：计算每个 KV cache group 的块数时，对 cross-attention spec
    # 使用 encoder 长度而不是 token 长度。
    def _warmup_block_count(num_tokens: int, spec: Any) -> int:
        if isinstance(spec, CrossAttentionSpec):
            num_tokens = max_encoder_len
        num_blocks = cdiv(num_tokens, spec.block_size)
        if isinstance(spec, MambaSpec) and spec.mamba_cache_mode == "align":
            num_blocks += spec.num_speculative_blocks
        return num_blocks

    ...
    # 在构建 NewRequestData 时，传递 warmup_mm_features 给 Request
    new_reqs = [
        NewRequestData.from_request(
            Request(
                req_ids[i],
                prompt_token_ids,
                sampling_params,
                pooling_params,
```

```
                mm_features=warmup_mm_features, # 新增参数
            ),
            block_ids=tuple(_alloc_blocks(n) for n in prefill_block_counts),
            prefill_token_ids=prompt_token_ids,
        )
    ]
```

评论区精华

无 review 评论。PR 由 WoosukKwon 审核并批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：仅限于预热路径，不影响正常推理；只在 is_encoder_decoder 为 True 且 max_encoder_len > 0 时生效；dummy 输入不包含实际数据，仅用于占位。
- 影响：影响范围：仅涉及 encoder-decoder 模型（如 Whisper）在 V1 引擎上的预热阶段。修复后 cross-attention 可正确预热，避免零 key 操作，提升性能与正确性。
- 风险标记：核心路径变更

关联脉络

- PR #47029 [Bugfix] Prevent padding placeholders from reaching embeddings: 同为 v1 encoder-decoder 相关 bugfix，涉及模型运行器中的特殊输入处理。