PR #42933 完整报告

vllm-project/vllm

Reduce memory usage for granite_speech.

合并时间：2026-05-25 14:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42933

执行摘要

- 一句话：用 Einsum 替换 Sum 减少显存占用
- 推荐动作：值得合并：这是一个小巧而高效的显存优化，仅修改一行核心表达式，经维护者审核和测试验证。开发者可借此了解如何通过 Einsum 避免广播中间张量的显存爆炸。

功能与动机

PR body 明确指出原始 torch.sum 实现会存储完整的中间矩阵，在 ibm-granite/granite-speech-4.1-2b 模型上消耗超过 10GB 显存，导致 12G 和 16G 显存卡无法运行。作者 Yihuki 在评论中强调 "This blocks using granite_speech 4.1 for 12G and 16G card and is a very tiny synonymous change"。

实现拆解

1. 定位问题代码：在 vllm/model_executor/models/granite_speech.py 的 GraniteSpeechConformerBlockAttention.forward 方法中，计算相对位置嵌入时，原代码通过 query_states.unsqueeze(-2) * rel_pos_emb_expanded 创建形状为 (bsz, num_blocks, num_heads, context_size, context_size, head_dim) 的 6D 中间张量，然后沿最后一维求和（torch.sum(..., dim=-1)）。该中间张量在 context_size 较大时显存开销巨大。
2. 替换为 Einsum：使用 torch.einsum("bnhid,ijd->bnhij", query_states, rel_pos_emb) 直接计算最终形状 (bsz, num_blocks, num_heads, context_size, context_size) 的注意力分数，避免实例化完整 6D 张量，显存占用大幅降低。
3. 移除不再需要的扩展步骤：删除了 rel_pos_emb_expanded = rel_pos_emb.view([1, 1, 1] + list(rel_pos_emb.shape))，因为 Einsum 直接利用 rel_pos_emb 的原始形状完成运算。
4. 保持语义等价：乘以 self.scale 的逻辑与原来一致。改动仅局限于 pos_attn 计算部分，不影响后续的掩码、SDPA 和输出处理。

关键文件：

- vllm/model_executor/models/granite_speech.py（模块 模型执行；类别 source；类型 core-logic）：唯一变更文件，核心改动是将相对位置嵌入计算从 torch.sum 替换为 torch.einsum，消除大型中间张量，显存节省超 10GB。

关键符号：未识别

关键源码片段

vllm/model_executor/models/granite_speech.py

唯一变更文件，核心改动是将相对位置嵌入计算从 torch.sum 替换为 torch.einsum，消除大型中间张量，显存节省超 10GB。

```
# vllm/model_executor/models/granite_speech.py ( 修改后 )

# 计算相对位置嵌入 ( 修改前后对比 )
dist = attention_dists.to(hidden_states.device)
rel_pos_emb = self.rel_pos_emb(dist)
# 原实现：先扩展 rel_pos_emb 并创建 6D 中间张量，再求和（消耗大量显存）
# rel_pos_emb_expanded = rel_pos_emb.view([1, 1, 1] + list(rel_pos_emb.shape))
# pos_attn = torch.sum(query_states.unsqueeze(-2) * rel_pos_emb_expanded, dim=-1) * self.
scale
# 新实现：使用 einsum 直接计算，避免中间 6D 张量，显存占用大幅降低
pos_attn = (
    torch.einsum("bnhid,ijd->bnhij", query_states, rel_pos_emb) * self.scale
)
```

评论区精华

- 代码审查主要由 gemini-code-assist[bot] 自动执行，未提供具体反馈。
- 模型维护者 alex-jw-brooks 明确认可（"Looks good to me"）并批准，DarkLight1337 也批准合并，无争议。
- 暂无高价值评论线程

风险与影响

- 风险：风险极低：1）只改变了 vllm/model_executor/models/granite_speech.py 中单一符号（pos_attn 计算）的内联表达式；2）torch.einsum 在语义上与原始的 torch.sum 加逐元素乘法等价，数值精度一致；3）已有测试 tests/models/multimodal/generation/test_granite_speech.py 全部通过；4）改动量仅删除 3 行、新增 1 行，逻辑可直接审查。无性能、安全或兼容性风险。
- 影响：积极影响：使 ibm-granite/granite-speech-4.1-2b 模型能在 12GB 和 16GB 显存显卡（如 RTX 3080/4060）上运行，显著降低硬件门槛。范围：仅影响 Granite Speech 模型的该注意力模块，其他模型和行为不受影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR