

PR #48258 完整报告

vllm-project/vllm

[ROCm][CI] Transformers: pass only one of input_ids/inputs_embeds

合并时间: 2026-07-14 01:28

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48258>

执行摘要

- 一句话: 修复 Transformers 后端 CUDA Graph 捕获时 input_ids 与 inputs_embeds 冲突
- 推荐动作: 该 PR 值得阅读, 尤其适合理解 vLLM 中 CUDA Graph 捕获路径与 Transformers 模型之间的交互。关键设计决策: 选择在模型层 (Base.forward) 进行防御性检查, 而非修改 CUDA Graph 捕获逻辑, 从而避免引入更广泛的变动。这是 4 行代码修复经典契约违规的范例。

功能与动机

vLLM 的 Transformers 建模后端在 CUDA Graph 捕获 / 预热期间, 会同时传入虚拟的 input_ids 和预计算的 inputs_embeds, 违反了 HuggingFace Transformers 模型要求“必须指定且仅指定一个”的契约。例如 Gemma3 会抛出 ValueError: You must specify exactly one of input_ids or inputs_embeds, 导致 ROCm CI 测试 test_can_initialize_large_subset[TransformersMultiModalForSequenceClassification] 失败。

实现拆解

1. 在 vllm/model_executor/models/transformers/base.py 的 forward 方法中, 在调用 self.model() 之前新增一个检查分支 (第 634-636 行): 如果 input_ids 和 inputs_embeds 同时不为 None, 则将 input_ids 置为 None。
2. 优先级决策: 保留 inputs_embeds 是正确选择, 因为当预计算了嵌入时, 它们可能包含融合了多模态图像和文本的嵌入, 是权威输入, 而 input_ids 在此路径中是虚拟的。
3. 该分支仅在两者同时非 None 时触发; 所有仅传入单个参数的路径不受影响, 因此其他 Transformers 后端模型无变化。
4. 没有新增测试文件, 但依赖已有 CI 测试验证。测试命令: `pytest -x -s "tests/models/test_initialization.py::test_can_initialize_large_subset[TransformersMultiModalForSequenceClassification]"`, 修改后测试通过。

关键文件:

- vllm/model_executor/models/transformers/base.py (模块 模型执行器; 类别 source; 类型 data-contract): 核心修复文件: 在 Base.forward 方法中增加检查, 确保同时提供 input_ids 和 inputs_embeds 时, 将 input_ids 置为 None 以遵守 HF 契约。仅有 4 行新增代码, 但解决了 CUDA Graph 捕获路径导致的模型初始化失败问题。

关键符号: Base.forward

评论区精华

核心讨论：reviewer hmellor 最初质疑该问题是否为 ROCm 特有（CUDA CI 中未失败），并怀疑是否由 ROCm runner 中同时传入两者所致。提交者 stefankoncarevic 通过端到端追踪确认，问题根源在于 CUDA Graph 捕获路径（vllm/v1/worker/gpu/cudagraph_utils.py 的 `create_forward_fn`）同时构建了 `input_ids` 和 `inputs_embeds`。hmellor 随后指出移除 runner 中的 `input_ids` 会导致 runner 仅支持 `embeds`，因此支持将修复移回 `Base.forward`，从而保持所有路径正确。最终达成一致：在模型层修复是最佳位置。

- 问题根因分析：runner 还是模型层？(correctness): 共识：修复放在 `Base.forward`（模型层）是最佳选择，因为修改 runner 会破坏其兼容性（仅支持 `embeds`）。

风险与影响

- 风险：风险极低。改动仅 4 行新增代码，且只影响同时提供 `input_ids` 和 `inputs_embeds` 的路径（主要是 CUDA Graph 捕获）。正常推理路径（仅传一个参数）完全不受影响。所有 Transformers 后端模型在此之前若同时接收到两者本就会崩溃（如 Gemma3），因此新增检查不会引入回归。
- 影响：直接影响：修复了 ROCm（可能也包括其他平台）上使用 Transformers 后端的多模态模型（如 Gemma3）在 CUDA Graph 捕获时的初始化失败。间接影响：该修复确保了所有 Transformers 后端模型在 CUDA Graph 捕获场景下无论是否同时传入两者都能正常工作，提升了系统的鲁棒性。由于是在模型层处理，不影响 runner 或其他组件，影响范围集中且正面。
- 风险标记：核心路径变更，缺少测试覆盖（仅依赖现有 CI）

关联脉络

- 暂无明显关联 PR