PR #35727 完整报告

vllm-project/vllm

[model] support FireRedASR2

合并时间：2026-03-04 11:41

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/35727

执行摘要

本 PR 为 vLLM 新增了对 FireRedASR2 语音识别模型的支持，通过新增模型实现和音频处理器，扩展了多模态功能。在 review 过程中，解决了设备硬编码和批处理问题，但音频文件大小限制可能带来稳定性风险，建议用户预处理大文件。

功能与动机

根据 PR body，作者需要添加对 FireRedASR2 模型的支持，以使用户能够通过 vLLM 服务端进行音频转录。作者提供了使用示例，如 `vllm serve allendou/FireRedASR2-LLM-vllm -tp=1 --dtype=float32`，并提到用户可以从阿里云 PAI 购买相关服务，体现了商业化应用需求。

实现拆解

实现主要分为以下模块：

- 核心模型实现：在 vllm/model_executor/models/fireredasr2.py 中，新增 FireRedASR2ForConditionalGeneration 类，继承自 Qwen2ForCausalLM，实现 SupportsMultiModal 和 SupportsTranscription 接口，处理音频输入和转录逻辑。关键代码块包括 RelPosMultiHeadAttention 注意力和 Swish 激活函数。
- 音频处理器：在 vllm/transformers_utils/processors/fireredasr2_processor.py 中，新增 FireRedASR2FeatureExtractor 和处理器，使用 kaldi-native-fbank 库提取 Mel 特征，并处理音频 token 替换，支持批处理。
- 注册与文档：更新模型注册表（registry.py）、测试注册、文档（supported_models.md）和依赖文件（requirements/common.txt），确保模型可被正确识别和使用。

评论区精华

Review 中主要讨论了以下关键点：

- 设备硬编码：gemini-code-assist[bot] 指出 .cuda() 调用硬编码设备，作者 fixed 以提升设备无关性。
- 批处理逻辑错误：gemini-code-assist[bot] 指出处理器中批处理逻辑错误，作者 fixed 以确保多音频文件正确处理。
- 提示硬编码：gemini-code-assist[bot] 建议使用用户提供的 prompt，但作者回复模型必须包含中文，可能限制了多语言支持。

风险与影响

- 技术风险：初始版本存在设备硬编码风险，已修复；批处理逻辑错误可能影响准确性；音频文件超过 50MB 可能导致服务挂起，需用户预处理。新增依赖 kaldi-native-fbank 增加部署复杂度。
- 影响分析：用户可直接使用新模型进行音频转录，扩展了 vLLM 应用场景；系统需管理新依赖；团队需维护约 1183 行新代码，长期可能增加技术债务。

关联脉络

从近期历史 PR 分析，PR 36803（测试 Nemotron-3-Super）和 PR 37932（多模态嵌入处理）与本 PR 相关，表明 vLLM 正持续扩展模型库和多模态功能。结合 Issue 评论中用户报告的音频文件问题，未来可能需要优化大文件处理逻辑或添加更多测试覆盖。