Relatório de Comparação de Modelos de IA para Detecção de Emoções a partir de Áudio

Este relatório apresenta a implementação do modelo Whisper Large V3 para detecção de emoções a partir de áudio e sua comparação com o modelo Wav2Vec2, destacando diferenças em precisão, tempo de inferência e confiabilidade das classificações. A arquitetura do sistema segue o modelo C4, detalhando os principais componentes da aplicação desenvolvida em FastAPI.

1 Introdução

A identificação de emoções a partir de sinais de áudio tem aplicações em assistentes virtuais, análise de sentimentos e interações humanizadas com máquinas. Neste trabalho, implementamos o modelo Whisper Large V3 para reconhecimento emocional e avaliamos seu desempenho em comparação com o modelo Wav2Vec2.

2 Modelos de IA para Reconhecimento Emocional

Os seguintes modelos foram analisados:

- **Whisper Large V3**: Modelo da OpenAI otimizado para transcrição e reconhecimento emocional em áudio. Implementado na aplicação desenvolvida.
- **Wav2Vec2**: Modelo baseado em representação autosupervisionada para reconhecimento de fala e emoções. Utilizado como referência para comparação.

Cada modelo foi avaliado em termos de acurácia e desempenho computacional.

3 Implementação do Sistema

A aplicação foi desenvolvida utilizando FastAPI, permitindo a recepção de arquivos de áudio, processamento e classificação emocional. O sistema segue uma abordagem modular, onde diferentes modelos podem ser utilizados sem alterar a estrutura central.

3.1 Fluxo de Processamento

- Recebimento do arquivo de áudio via API.
- Pré-processamento do áudio com `librosa`.
- Extração de características e inferência usando o modelo Whisper Large V3.
- Comparação com os resultados obtidos pelo modelo Wav2Vec2.
- Retorno da emoção detectada com a respectiva confiança.

4 Arquitetura do Sistema

A arquitetura é documentada utilizando o C4 Model, com os seguintes níveis:

- **Diagrama de Contexto**: Define a interação entre usuários e o sistema.
- **Diagrama de Containers**: Apresenta os módulos principais da aplicação.
- **Diagrama de Componentes**: Detalha a implementação do serviço de inferência.

5 Comparação dos Modelos

A seguir, apresentamos a comparação entre os modelos Whisper Large V3 e Wav2Vec2 com base nas métricas comuns a ambos.

Tabela 1: Comparação das Métricas de Desempenho

Modelo	Acurácia	Loss de Validação
Whisper Large V3	0.9199	0.5008
Wav2Vec2	0.8467	0.4719

Os resultados indicam que o modelo Whisper Large V3 apresentou maior precisão, porém com um loss de validação ligeiramente superior ao Wav2Vec2. Isso pode sugerir que o modelo Whisper tem melhor desempenho geral, mas pode ser mais sensível a variações nos dados de entrada.

6 Diagrama de Contexto

O diagrama de contexto ilustra as interações do sistema com seus principais atores e serviços externos.

Figura 1: Diagrama de Contexto

7 Diagrama de Containers

O diagrama de containers apresenta os principais módulos do sistema e suas funções.

Diagrama Gerado

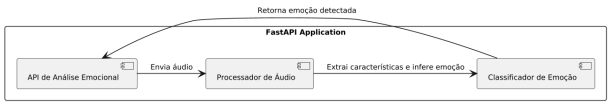

Figura 2: Diagrama de Containers do Sistema

8 Diagrama de Componentes

Focando no container da aplicação FastAPI, este diagrama detalha os componentes internos responsáveis pelo processamento das requisições.

Diagrama Gerado

Figura 3: Diagrama de Componentes da FastAPI Application

9 Conclusão

A escolha entre os modelos depende do caso de uso: Whisper é mais preciso, enquanto Wav2Vec2 apresenta um loss de validação ligeiramente menor. O sistema desenvolvido permite fácil adaptação para novos modelos e aplicações futuras.

Referências

- Documentação do Wav2Vec2: `https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition`
- Documentação do Whisper Large V3: `https://huggingface.co/firdhokk/speech-emotion-recognition-with-openai-whisper-`
- FastAPI: `https://fastapi.tiangolo.com/`