Unfake

Desenvolvimento de uma Ferramenta para Detecção de Deepfakes de Áudio Utilizando Inteligência Artificial

Autores: Marcos Godinho Filho, Éric Carvalho Figueira

Orientador: Guilherme de Oliveira Macedo / Coorientadora: Andréia Cristina de Souza

Instituição: Colégio Técnico de Campinas - Unicamp

INTRODUÇÃO

A crescente evolução das ferramentas de **inteligência artificial (I.A.)** as torna cada vez mais eficientes e acessíveis globalmente. No entanto, algumas dessas tecnologias podem ser nocivas, caso usadas de forma mal-intencionada, e isso inclui as **deepfakes de áudio** . Elas são um tipo de mídia sintética que gera conteúdos realistas, tendo o potencial de **clonar a identidade** de um indivíduo, utilizando-a para a propagação de **notícias falsas**, deterioração de sua **reputação** e promoção de **fraudes** e violações de **segurança**. Assim, são necessárias maneiras de **detectar** se uma mídia é real ou foi sintetizada artificialmente.

OBJETIVOS

A pesquisa tem por objetivo desenvolver um **modelo classificatório de inteligência artificial** que seja capaz de identificar potenciais **áudios** produzidos com **deepfake em português**, sendo esse modelo disponibilizado em um **website**, permitindo aos usuários fazer upload de um áudio e receberem a probabilidade do conteúdo em questão ser real ou sintético.

METODOLOGIA

1. Deepfakes

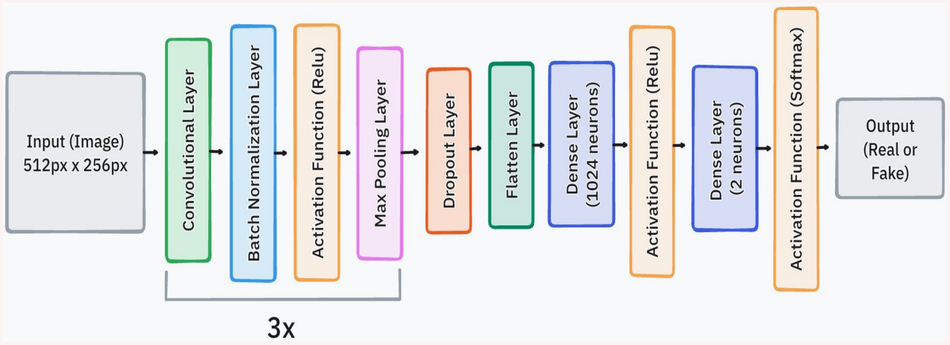
Uso de falas do **dataset do "CETUC"** em conjunto com transcrições para alimentar modelo **"XTTS"** e gerar **deepfakes**.

2. Espectrogramas

Uso da biblioteca "Librosa", em Python, para gerar **representações visuais** dos áudios contendo falas reais e sintéticas e produção de um dataset.

3. Rede neural

- Desenvolvimento de um **modelo classificatório** na arquitetura **CNN**, utilizando a biblioteca **"Tensorflow"**, da linguagem **Python**.
- **Treinamento** do mesmo utilizando o dataset produzido e avaliação do **desempenho**.



5. Website

Criação de uma **interface de usuário** para permitir o **envio** de áudios e obtenção das **previsões** do modelo através da API.

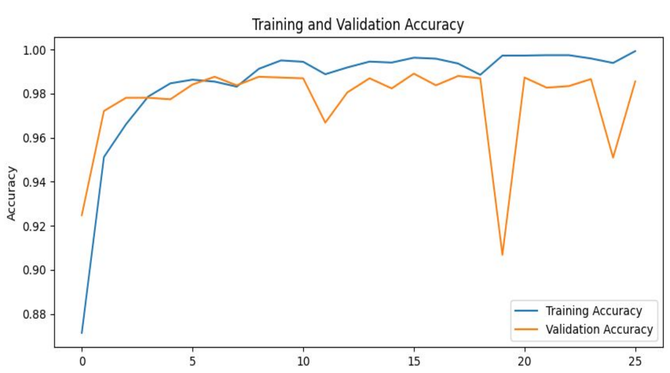
4. API

Criação de uma **API** em **Python** para receber **áudios**, convertê-los em **espectrogramas**, fornecê-los ao **modelo de I.A.** e retornar o **resultado** gerado pelo mesmo.

RESULTADOS

O projeto avançou de forma significativa em direção aos objetivos propostos. Foram geradas deepfakes de áudio utilizando o modelo **"XTTS"**, a partir dos áudios e transcrições da base de dados do **"CETUC"**. Além disso, mais de **180 mil espectrogramas CQT** foram produzidos a partir dos áudios, e ruídos foram adicionados aos mesmos, construindo-se um **dataset**.

Ademais, foi estruturado e treinado um modelo **CNN** para **detectar deepfakes**. O resultado do treinamento, utilizando uma parcela do dataset, pode ser visualizado ao lado.



Por fim, a **API** e o ***Website*** já foram **desenvolvidos e hospedados**. O **Dataset** contendo espectrogramas e o **Website** podem ser acessados, respectivamente, através dos QR Codes abaixo:

CONCLUSÕES

Conclui-se que o projeto desenvolvido pode auxiliar a **mitigar os efeitos nocivos** provocados pelas **deepfakes de áudio** nos setores **social, político e econômico** , bem como fomentar **pesquisas na área** pela produção de um dataset de deepfakes em português.

No entanto, analisando-se os resultados obtidos, nota-se que o **modelo CNN** está **enviesado** ao dataset utilizado, o que pode ser explicado por ter sido utilizado somente um modelo para geração das deepfakes, bem como uma quantidade de dados relativamente pequena. Assim, para melhores resultados futuros, é necessário um **dataset mais diverso e extenso.**

REFERÊNCIAS BIBLIOGRÁFICAS

ALMUTAIRI, Zaynab; ELGIBREEN, Hebah. A review of modern audio deepfake detection methods: Challenges and future directions. Algorithms, v. 15, n. 5, p. 155, 2022. Disponível em: <https://www.mdpi.com/1999-4893/15/5/155>. Acesso em: 15 ago. 2024.

CASANOVA, Edresson et al. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. arXiv e-prints, p. arXiv: 2406.04904, 2024. Disponível em: <https://arxiv.org/pdf/2406.04904>. Acesso em: 15 ago. 2024.

HONG, T. J. Uncovering the Real Voice: How to Detect and Verify Audio Deepfakes. Medium, 2023. Disponível em: <https://medium.com/htx-s-s-coe/uncovering-the-real-voice-how-to-detect-and-verify-audio-deepfakes-42e480d3f431#:~:text=A%20single%2dSpeaker%20module%20is>. Acesso em: 15 ago. 2024.

MUBARAK, Rami et al. A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. IEEE Access, 2023. Disponível em: <https://ieeexplore.ieee.org/document/10365143?denied=>. Acesso em: 15 ago. 2024.