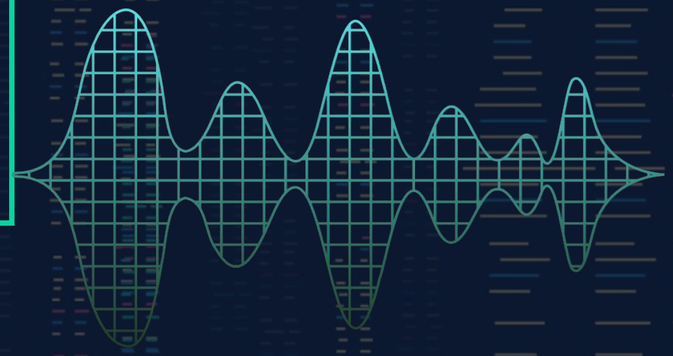
Detecção de Deep Fakes

João Victor do Nascimento Secate

Processamento digital de imagens



01. O QUE SÃO DEEP FAKES

- Deep Fake é uma tecnologia capaz de introduzir, por meio de IA, o rosto de qualquer pessoa em um outro corpo.
- Mesmo que muitos chamam de Deep fake, montagens feitas utilizando software de manipulação de imagem, como Photoshop, não são Deep Fakes.
- Deep Fakes são criadas utilizando **aprendizagem de máquinas**.
- Exemplo 1: Velozes e Furiosos 7 (2015) - Cena criada utilizando computação gráfica.
- Exemplo 2: The Mandalorian S02 (2020) - Cena criada utilizando Deep Fake.

02. COMO SÃO FEITAS

- Deep Fakes são criadas utilizando **aprendizagem de máquinas**, ou seja, utilizar dados e algoritmos para permitir que uma inteligencia artificial aprenda igual a um ser humano.
- Em Deep Fakes, uma **grande quantidade** de imagens são analisados por um algoritmo.
- Esse algoritmo analisa os rostos presentes na imagens, aprendendo sobre a **estrutura e detalhes** desse rosto.
- Com essa analise, é possível replicar um rosto analisado e o sobrepor em outro rosto também analisado.
- O algoritmo realiza esse processo **diversas vezes**, aprendendo como replicar um rosto da forma mais convincente.

03. O RISCO DAS DEEP FAKES

Deep Fakes muitas vezes são utilizadas para:

- Campanhas de desinformação
- Interferência em eleições
- Chantagem
- bullying e assédio
- Pornografia não consensual
- Boatos e notícias falsas
- Golpes e fraudes

USO DE DEEP FAKES

96%

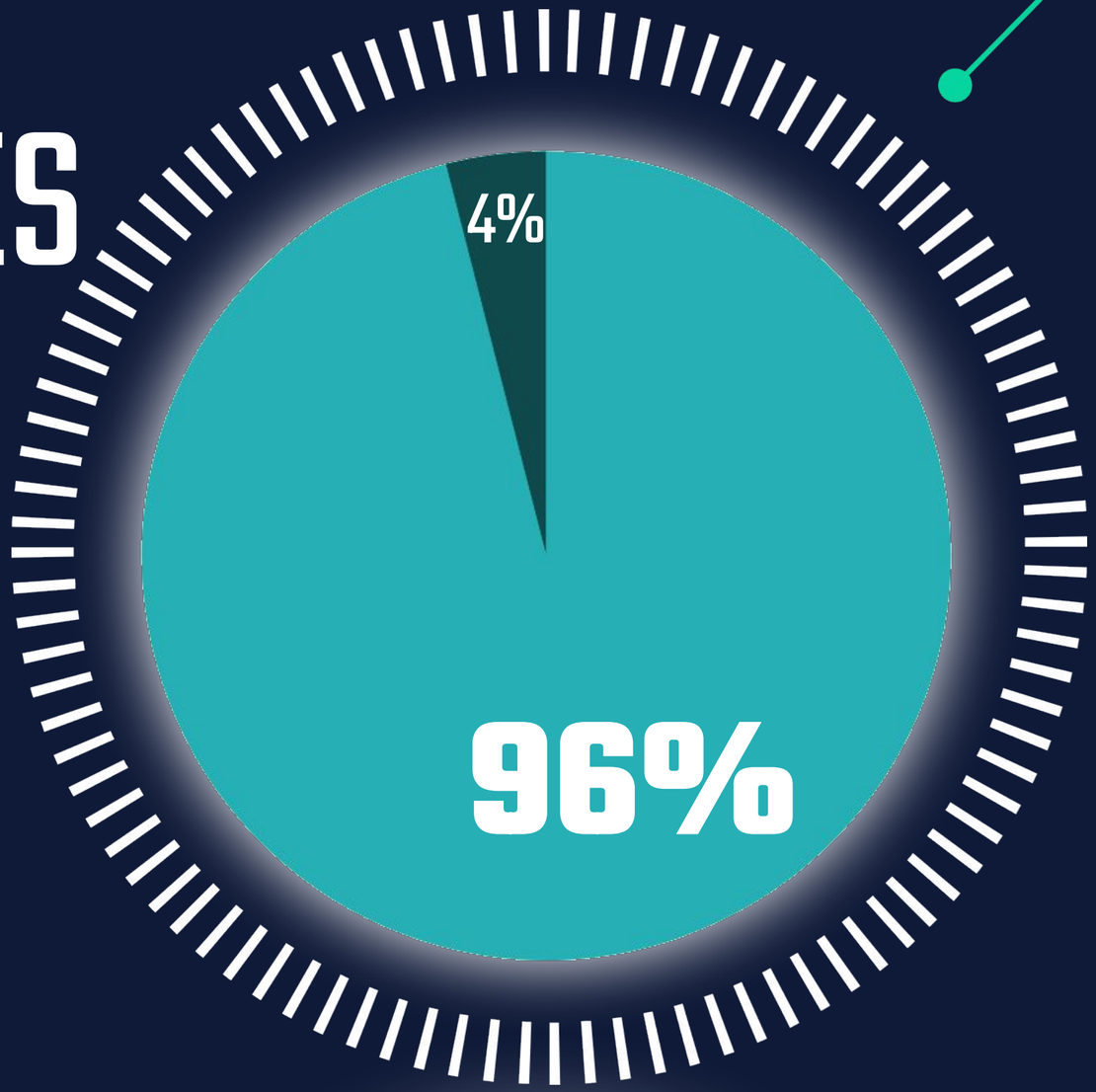
Usado em pornografia

Videos pornograficos, tendo como alvo mulheres, especialmente celebridades.

4%

Demais usos

Noticias falsas, imagens e videos virais, efeitos visuais, golpes e fraudes.



04. DETECÇÃO DE DEEP FAKES

- Como qualquer outra tecnologia, Deep Fakes possuem falhas, algumas podem ser diferenciadas visualmente.
- Exemplos como rasgos nas imagens, borramento irregular, inconsistência nos olhos e bocas, expressões faciais anormais e vídeos com baixa qualidade.
- Com o avanço das Deep Fakes, diversos métodos foram implementados para automaticamente diferenciar imagens falsas e verdadeiras

MÉTODOS POPULARES DE DETECÇÃO DE DEEP FAKES

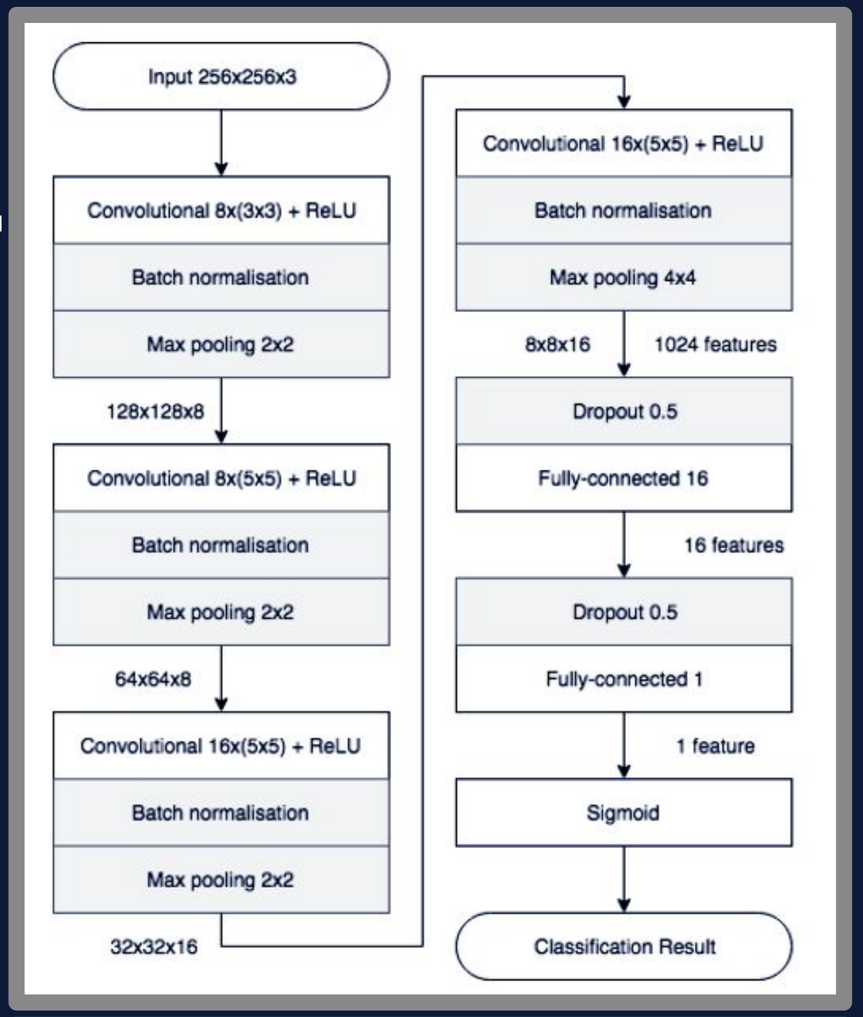
Metodo	Modelo	Performance afirmada
MesoNet (Afchar et al., 2018)	CNN	98%
Guera et al. (Güera and Delp, 2018)	CNN+LSTM	97.1%
FakeCatcher (Ciftci and Demir, 2019)	Traditional operator+CNN	96% (Dataset - FaceForensics++) 91.07% (Dados privados)
XceptionNet (Rossler et al., 2019)	XceptionNet	Qualidade original: 99.26% Alta qualidade: 95.73% Baixa qualidade: 81.00%

05. MESONET : Meso-4

- Utiliza redes neurais profundas

Arquitetura da rede Meso-4:

- Camadas iniciais: 4 camadas de convolução e pooling.
- Camadas finais: rede densa com 1 camada oculta.
- Técnicas de otimização:
- Ativação ReLU e Normalização por Lotes nas convoluções.
- Dropout nas camadas totalmente conectadas.
- Eficiência: apenas 27.977 parâmetros treináveis.
- Baseada em redes eficientes de classificação de imagens.



Detecção de Deep Fake com Meso-4

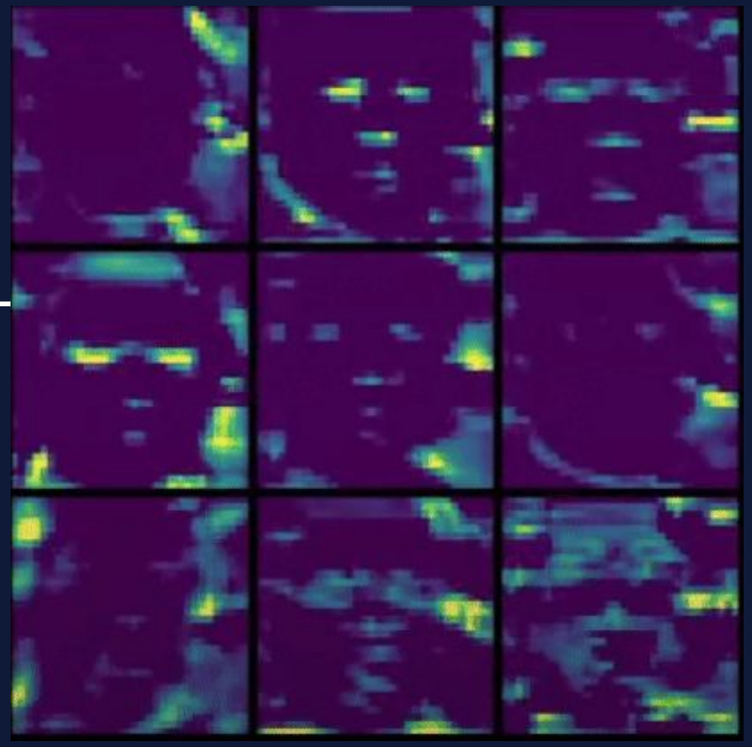
Passo 1: Alinhamento e extração de rosto

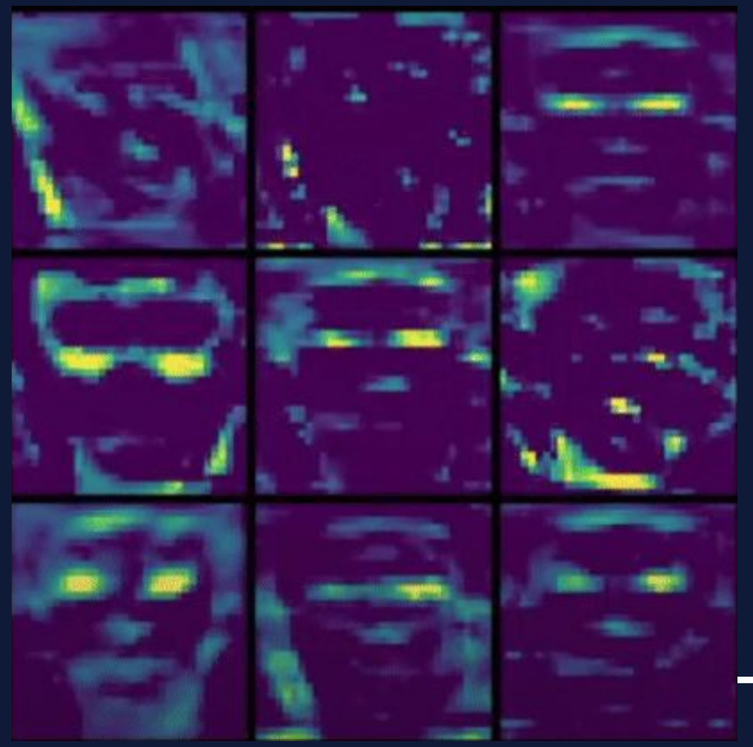
Passo 2: Previsão de quadro usando redes neurais profundas

Verdadeiro

Falso

Podemos observar como olhos são essenciais para identificar uma Deep Fake





Passo 3: Agregações de predições

Verdadeiro

Falso

MESONET : Meso-4

Pontos positivos

- Facil de implementar
- Não utiliza muitos recursos da máquina
- Open-Source

Pontos negativos

- Outras redes possuem uma média de precisão maior entre diferentes datasets, principalmente datasets mais recentes
- Falsos Positivos tendem a serem rostos femininos

MesoNet - Uso em outras bases de dados

Base	Acerto
DFD	85.02%
DF-TIMIT LQ	91.18%
DF-TIMIT HQ	83.71%
FF++(Deepfake) LQ	87.75%
FF++(Deepfake) HQ	97.04%
Wild-Deepfake	64.47%

Referencias

AFCHAR, Darius et al. MesoNet: a Compact Facial Video Forgery Detection Network. 4 set. 2018. Disponível em: https://arxiv.org/abs/1809.00888v1. Acesso em: 26 nov. 2024.

AJDER, Henry et al. The State of Deepfakes: Landscape, Threats, and Impact. Set. 2019. Disponível em: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. Acesso em: 1 dez. 2024.

FURIOUS 7 | Wētā FX. Disponível em: https://www.wetafx.co.nz/films/filmography/furious-7. Acesso em: 28 nov. 2024.

GAMBÍN, Ángel Fernández et al. Deepfakes: current and future trends. Artificial Intelligence Review, v. 57, n. 3, 19 fev. 2024. Disponível em: https://doi.org/10.1007/s10462-023-10679-x. Acesso em: 1 dez. 2024.

PROFESSIONAL VFX Artists Explain how to Spot Fake Videos. 27 nov. 2022. 1 vídeo (13 min 15 s). Publicado pelo canal Corridor Crew. Disponível em: https://www.youtube.com/watch?v=KqIW1TczgsA. Acesso em: 29 nov. 2024.

RANA, Md Shohel et al. Deepfake Detection: A Systematic Literature Review. IEEE Access, v. 10, p. 25494-25513, 2022. Disponível em: https://doi.org/10.1109/access.2022.3154404. Acesso em: 1 dez. 2024.

THOMAS, By Daniel. Deepfakes: A threat to democracy or just a bit of fun? 23 jan. 2020. Disponível em: https://www.bbc.com/news/business-51204954. Acesso em: 1 dez. 2024.

ZI, Bojia et al. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. 2020. Disponível em: https://arxiv.org/html/2101.01456v2. Acesso em: 1 dez. 2024.