Aluno: Gustavo Lacerda

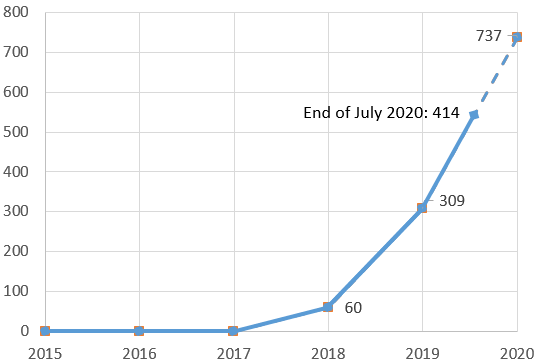
Orientador: Raimundo

**DEEPFAKES EXPOSED**

**1 HISTÓRIA E O PROBLEMA DAS DEEPFAKES**

Deep fake é uma tecnologia que usa Inteligência Artificial (IA) para para criação de imagens, vídeos e sons falsos utilizando dados reais. Essa tecnologiasurgiu, com esse nome, inicialmente no ano de 2017 através de um usuário do fórum *reddit* nomeado “deepfakes”. Através de um página neste fórum /r/deepfakes ([*https://www.reddit.com/r/deepfakes/*](https://www.reddit.com/r/deepfakes/)*)*, no qual usuários postavam suas deepfakes caseiras. A página foi banida futuramente pelos administradores do reddit por causa do conteúdo pornográfico criado e publicado, principalmente, usando rostos de atrizes famosas.

Na área acadêmica as deepfakes são bastante recorrentes na área de visão computacional, e existe relato de técnica de deepfake desde 1997. Esta tecnologia, portanto, ficou congelada por certo tempo até que, com o advento de melhores algoritmos de deep learning e poder de processamento, conseguiram ser mais acessíveis e têm tido aumento exponencial de popularidade e discussões.



**Imagem 1** - Número de publicações relacionadas à deepfake até julho de 2020 com projeção linear esperada até o final deste ano (Dados de: <https://app.dimensions.ai>)

De acordo com projeção linear feita com dados do site <https://app.dimensions.ai> no final de 2020 se previa cerca de 737 trabalhos relacionados como mostrado no gráfico acima. Porém, de acordo com o mesmo site utilizando o mesmo método de pesquisa, houveram 1.333 trabalhos relacionados à deepfake até o final do ano, um aumento de cerca de 80% em relação à projeção. Este aumento se deve ao fato da grande popularização das deepfakes, que viralizaram com memes, vídeos de políticos famosos e de artistas, além de que os computadores caseiros aumentaram seu poder de processamento ao ponto em que a maior parte das imagens geradas possam ser feitas de forma amadora, a transformando em *mainstream*.

Apesar da grande popularização das deepfakes ser relacionada a memes, há uma outra parte bastante nociva para a informação que circula na era da internet. Pois elas estão sendo massivamente utilizadas para criar *fake news de luxo* com rostos de políticos, além da criação de pornografia ilegal envolvendo figuras públicas.

**2 CRIAÇÃO DE DEEPFAKES**

**2.1 Métodos de Criação de DeepFakes**

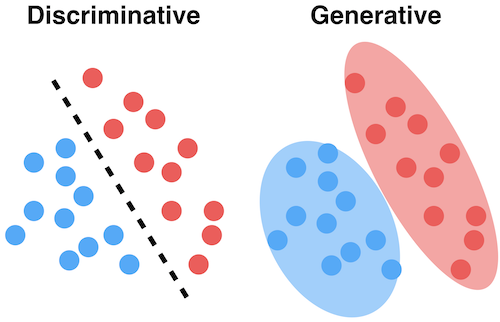
Os deepfakes utilizam um tipo de rede neural chamada autoencoder. Que consistem em um encoder, que reduz a imagem para um espaço de dimensão inferior, e um decoder que reconstrói a imagem. Nas deepfakes essa tecnologia é aplicada em uma imagem, no qual o encoder armazena as principais características desejadas de um rosto e posições corporais. Assim, essas informações podem ser decodificadas especialmente para um alvo, no qual, as características do alvo são impostas a partir da imagem original.

Um *upgrade* que ficou popular para esse tipo de arquitetura é a Rede Adversária Generativa (GAN - Generative Adversarial Networks). Esse modelo ficou bem famoso e é amplamente utilizado na geração de imagens.

As GANs são arquiteturas de redes neurais profundas compostas por duas redes colocadas uma contra a outra. Esta arquitetura é uma das mais recentes e foi introduzida em 2014 em um artigo produzido por pesquisadores da Universidade de Montreal. O potencial das GANs se deve ao fato de serem ótimas em imitar distribuições de dados, sendo capaz de criar universos semelhantes ao mundo real no que tange imagens, músicas, falas e literatura.

As Redes Adversárias Generativas pertencem ao conjunto de modelos generativos, que possuem a capacidade de criar um novo conteúdo. Estes algoritmos diferem dos discriminativos, que tentam classificar os dados de entrada. Por exemplo, um algoritmo discriminativo, a partir da entrada do conteúdo de um e-mail tentaria classificá-lo como spam ou não. Algoritmos generativos também podem ser usados como classificadores, porém possuem abordagem diferente, eles estão mais preocupados em saber como se obtém certa coisa.

* Modelos discriminativos aprendem a fronteira entre classes.
* Modelos generativos modelam a distribuição de classes individuais.

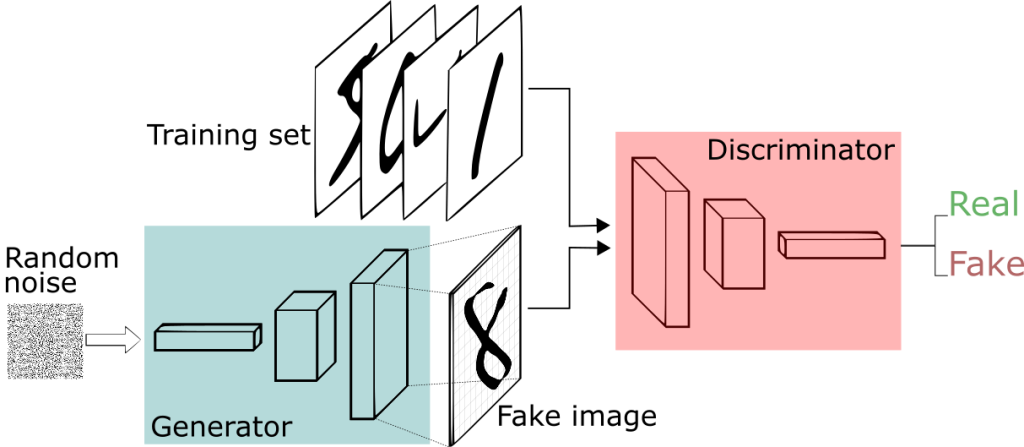


**Imagem 2** - Modelo Discriminativo vs Generativo (link: https://medium.com/@akankshamalhotra24/generative-classifiers-v-s-discriminative-classifiers-1045f499d8cc)

Nas GANs há duas redes, uma geradora e uma discriminadora, no qual a primeira gera novos dados e a segunda avalia se os dados gerados pertencem ao conjunto de dados de treinamento real.

As etapas de uma GAN são:

* O gerador considera números aleatórios e retorna uma imagem.
* Essa imagem gerada é inserida no discriminador ao lado de um fluxo de imagens tiradas do conjunto de dados real e verdadeiro.
* O discriminador obtém imagens reais e falsas e retorna probabilidades, um número entre 0 e 1, com 1 representando uma previsão de imagem autêntica e 0 representando previsão de imagens falsas (geradas pela rede generativa).



**Imagem 3** - Exemplo do funcionamento de uma arquitetura GAN

Com essa competição das redes para tentar diminuir seus erros em suas respectivas funções, isso acaba gerando um algoritmo bastante eficiente. Na medida que, enquanto a rede geradora tenta cada vez mais gerar uma imagem fidedigna para burlar a rede discriminadora, a rede discriminadora tenta melhorar sua detecção para detectar imagens falsas, gerando assim um loop de retroalimentação que melhora ambas as redes. Por esse motivo essa arquitetura é tão popularmente usada na criação de novos dados.

**2.1 Principais Ferramentas De criação de DeepFakes**

**Tabela 1 -** Principais Ferramentas de Criação de DeepFakes encontradas no github

| **Ferramenta** | **Link** | **Principais Características** |
| --- | --- | --- |
| Faceswap | https://github.com/deepfakes/faceswap | - Using two encoder-decoder pairs.  - Parameters of the encoder are shared. |
| Faceswap-GAN | https://github.com/shaoanlu/faceswap-GAN | Adversarial loss and perceptual loss (VGGface) are added to an auto-encoder architec-  ture. |
| Few-Shot Face  Translation GAN | https://github.com/shaoanlu/fewshot-face-translation-GAN | - Use a pre-trained face recognition model to extract latent embeddings for GAN  processing.  - Incorporate semantic priors obtained by modules from FUNIT and SPADE. |
| DeepFaceLab | https://github.com/iperov/DeepFaceLab | - Expand from the Faceswap method with new models, e.g. H64, H128, LIAEF128,  SAE.  - Support multiple face extraction modes, e.g. S3FD, MTCNN, dlib, or manual. |
| DFaker | https://github.com/dfaker/df | - DSSIM loss function is used to reconstruct face.  - Implemented based on Keras library. |
| DeepFake tf | https://github.com/StromWine/DeepFake\_tf | Similar to DFaker but implemented based on tensorflow. |
| Deepfakes web | https://deepfakesweb.com/ | Commercial website for face swapping using deep learning algorithms. |

Fonte: NGUYEN, Thanh Thi et al. Deep learning for deepfakes creation and detection: A survey. arXiv preprint arXiv:1909.11573, 2019.

(Aplicativos que aplicam as ferramentas de criação de deepfakes)

**3 TÉCNICAS DE DETECÇÃO DE DEEPFAKES**

Com a popularização e a problemática das deepfakes surgiu a necessidade de ter algum método para que se possa detectá-los. Esse campo de pesquisa também cresceu bastante e se dividiu em duas principais áreas: detecção de imagens fakes e detecção de vídeos fakes.

**Tabela 2 -** Principais algoritmos de detecção de deepfakes testados encontrados no site *paperswithcode.com*

| **Nome** | **Artigo** | **Código Fonte** | **Método** | **Base de Dados** | **Teste** |
| --- | --- | --- | --- | --- | --- |
| Video Face Manipulation Detection Through Ensemble of CNNs | [arxiv](https://arxiv.org/pdf/2004.07676v1.pdf) | [github](https://github.com/polimi-ispl/icpr2020dfdc) | O algoritmo utiliza uma rede neural convolucional treinada para detectar se um rosto é falso ou não em cada frame de um vídeo. | FaceForenciss++  DFDC | [colab](https://colab.research.google.com/drive/1Awvm3QBdrwMfQf3EMbgEjRaD5_9OR4HR?usp=sharing) |
| MesoNet: a Compact Facial Video Forgery Detection Network | [arxiv](https://arxiv.org/pdf/1809.00888.pdf) | [github](https://github.com/DariusAf/MesoNet) | O algoritmo utiliza uma rede neural convolucional simples treinada para detectar o vídeo é um deepfake. | [pcloud(download)](https://e.pcloud.link/publink/show?code=XZnsxkZkEAgI1OgQIJHLnNl9ErhV4vpHuV0) | [colab](https://colab.research.google.com/drive/1zMvrVbmNfMB4Ecph3I7CsnLgl2bnx3XX?usp=sharing) |
| Face X-ray for More General Face Forgery Detection | [arxiv](https://arxiv.org/pdf/1912.13458v2.pdf) | [github](https://github.com/AlgoHunt/Face-Xray) | O algoritmo utiliza filtros para decompor a imagem e verificar se ela é formada por duas fontes. | FaceForencis++ | [colab](https://colab.research.google.com/drive/1OZC7d2S6Rf0w67U0b6DwcJtneA8pgNe_?usp=sharing) |

\*\* A maioria dos algoritmos utiliza FaceForencis++ e DFDC.

\*\* A maioria dos algoritmos utiliza a biblioteca face\_recognition ou blazeface do python para extração de rostos de frames.

\*\* A maior parte dos algoritmos precisa de treino que impossibilitam sua execução

\*\* Pytorch é mais comum que TensorFlow

\*\* Os vídeos são separados em frames antes de serem analisados

**4 BASES DE DADOS COM DEEPFAKES**

| **Base de Dados** | **Tipo** |
| --- | --- |
| [Celeb-DF (v2)](https://drive.google.com/open?id=1iLx76wsbi9itnkxSqz9BVBl4ZvnbIazj) | Vídeos |
| [Celeb-DF (v1)](https://drive.google.com/open?id=10NGF38RgF8FZneKOuCOdRIsPzpC7_WDd) | Vídeos |
| [FaceForenciss++](https://github.com/ondyari/FaceForensics) | Vídeos/Imagens |
| [DeepFake DEtection Challenge (DFDC)](https://www.kaggle.com/c/deepfake-detection-challenge/data) | Vídeos |
| [MesoNet Dataset](https://github.com/DariusAf/MesoNet#:~:text=Download%20link%20for%20the%20dataset) | Imagens |

**REFERÊNCIAS**

Data Science Academy. **Deep Learning Book**, 2021. Disponível em: <https://www.deeplearningbook.com.br>. Acesso em: Ago. 2021

NGUYEN, Thanh Thi et al. Deep learning for deepfakes creation and detection: A survey. **arXiv preprint arXiv:1909.11573**, 2019.