Detector de DeepFake

Jamil Anderson Mansur
Henrique Kenzo Odaguiri
Universidade Federal de São Paulo
Sâo José dos Campos, São Paulo
jamil.mansur@unifesp.br
henrique.odaguiri@unifesp.br

Resumo-Este projeto tem o objetivo de treinar e testar a capacidade de uma inteligência artificial em detectar um vídeo criado artificialmente, popularmente chamados de "Deepfakes". Usando redes neurais, um conjunto de vídeos será fornecido para o treinamento do algoritmo. Entre os vídeos, existem rostos reais e outros gerados por computador. Sendo assim, ao tratar e treinar a rede, medidas avaliativas serão expostas para obter a conclusão e o quão apta a tecnologia é em detectar algo que ela mesma pode criar.

1.Introdução do problema

1.1. Contextualização do problema

Apesar do avanço da Inteligência Artificial estar contribuindo para a prosperidade global, indivíduos mal-intencionados a tem utilizado com o propósito de aplicar golpes e realizar outros crimes. Por exemplo, o uso de aplicativos de "deepfake" estão em alta e prejudicando a imagem de pessoas reais ao reproduzirem as suas imagens em situações comprometedoras. Por isso, é imprescindível o desenvolvimento de tecnologias que identifiquem tais imagens.

1.2. Motivação

Segundo a polícia de HongKong, um funcionário do setor financeiro de uma multinacional caiu numa fraude no valor de US$ 25 milhões no início de 2024 após os golpistas utilizarem do deepfake para se passarem pelo diretor financeiro.

Também em 2024, imagens pornográficas geradas por Inteligência Artificial da cantora norte-americana Taylor Swift tiveram milhões de visualizações antes de serem removidas das redes sociais. [1]

As imagens adulteradas também são usadas na política. Em 2019, a ex-presidente da Câmara dos Deputados dos Estados Unidos, Nancy Pelosi, foi vítima de um deepfake que se baseou em um vídeo autêntico para sugerir que a representante democrata tinha dificuldades na fala em um discurso.

Posto isto, é possível dizer que a detecção de deep fakes torna-se essencial para confirmar a veracidade de um vídeo ou foto, visto que a adulteração destes tem se tornado cada vez mais eficiente e podendo ser feito por ferramentas cada vez mais acessíveis.[2]

2.Conceitos Fundamentais

2.1. Inteligência artificial

A inteligência artificial é um campo da ciência da computação que se dedica ao estudo e ao desenvolvimento de máquinas e programas computacionais capazes de reproduzir o comportamento humano na tomada de decisões e na realização de tarefas, desde as mais simples até as mais complexas. Existe uma série de diferentes métodos por meio dos quais uma IA pode reproduzir o comportamento humano. Os dois principais são:

Machine learning: chamado de aprendizado de

máquina, é o processo que acontece de maneira automatizada. O reconhecimento e a reprodução de padrões são feitos pela IA com base na sua experiência prévia, adquirida pela utilização de algoritmos. Um dos principais exemplos são os mecanismos de pesquisa na internet.

Deep learning: subcampo do machine learning, utiliza-se de redes neurais (unidades conectadas em rede para a análise de bancos de dados e informações) para emular o cérebro humano

2.2. Deepfake

O deepfake é uma técnica que utiliza aplicações com inteligência artificial (IA) para alterar fotos ou vídeos. Com esta técnica, é possível trocar o rosto de uma pessoa na cena ou modificar o que a pessoa fala.[2]

2.3. Rede Neural Convolucional

As Redes neurais convolucionais tentam imitar o funcionamento de um cérebro humano criando vários "Perceptrons" que agem como neurônios realizando funções básicas de cálculo e logo em seguida une cada um deles entre si, formando uma complexa rede que compartilha seus resultados e a cada camada tenta aprimorar seus resultados com base no que foi obtido na camada anterior

3.Trabalhos Relacionados

"Deepfake Video Detection Using Recurrent Neural Networks" O estudo por David Güera e Edward J. Delp, publicado na 15ª Conferência Internacional IEEE sobre Vigilância Avançada de Vídeo e Sinal em 2018, propõe uma abordagem para detectar vídeos deepfake utilizando redes neurais recorrentes (RNNs).[3]

"Deepfakes Classification of Faces Using Convolutional Neural Networks" O estudo por Jatin Sharma, Sahil Sharma, Vijay Kumar, Hany S. Hussein e Hammam Alshazly, estuda e treina um algoritmo capaz de detectar deep fakes a partir do rosto das pessoas usando técnicas de transfer learning, data augmentation e usando redes pré-treinadas como VGG16 e ResNet50. [4]

"Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model" Por Santosh Kolagati, Thenuga Priyadharshini e V. Mary Anita Rajam criam um algoritmo que, usando a detecção de pontos faciais, são extraídos dados referentes a vários atributos faciais dos vídeos. Esses dados são passados para um perceptron multicamadas para aprender as diferenças entre vídeos reais e deepfakes. Simultaneamente, uma rede neural convolucional é utilizada para extrair características e treinar nos vídeos. Esses dois modelos são combinados para construir um detector de deepfake multi-input. [5]

4.Objetivo

Esse projeto tem como objetivo evidenciar a viabilidade de se desenvolver uma ferramenta com o uso da Inteligência Artificial para combater a ameaça gerada por indivíduos mal-intencionados que utilizam a mesma tecnologia. Isso se mostra de extrema importância para conter o medo e apreensão existente na sociedade acerca do progresso do ramo da Inteligência Artificial, pois demonstra que o desenvolvimento de tecnologias nocivas ocorre concorrentemente com desenvolvimento de contramedidas.

5.Metodologia Experimental

O primeiro passo de nosso projeto é escolher um conjunto de dados com uma diversidade de pessoas e técnicas de *deepfake* para evitar que o algoritmo se desenvolva com uma capacidade limitada no reconhecimento de uma varidade de faces adultaredas.

Durante alguns turnos, parte desse conjunto de dados será separado para treinar a rede neural e a outra parte usada para a validação das predições realizadas por ela. Para isso, diversas bibliotecas, como Pytorch e

scikit-learn serão usadas.

5.1 Dados

Uma coleção de vídeos, originária de um concurso chamado "Deepfake Detection Challenge", está sendo utilizada. Estes vídeos são de pessoas conversando, gesticulando e interagindo normalmente. [6]

O dataset completo e em formato compactado "*.zip*" tem cerca de 470GB e contém 401 vídeos. Devido a dimensão do dataset, ele é dividido em partes menores. Foi escolhida uma partição experimental chamada "train_sample_videos" que possui amostras variadas de todo o conjunto, fornecendo assim uma variedade de pessoas e técnicas de *deepfake*.

Este conjunto amostral possui um arquivo metadata.json que contém o nome de cada arquivo .mp4 e o rótulo (REAL/FAKE). Um vídeo pode ser "FAKE" se ocorrer uma troca de rosto, ou "REAL", caso não tenham sido feitas alterações.

5.2 Tratamento

Visto que o dataset já foi criado e disponibilizado com a finalidade de ser usado para treinar e testar algoritmos de detecção de *deepfake*, ele dispensa a necessidade de seleção e rotulagem dos vídeos. Além disso, a base já se encontra balanceada, ou seja, a proporção de vídeos reais e modificados é parecida.

Assim, foi realizada a extração dos vídeos da pasta compactada para um diretório chamado "temp_zip_content". Para cada vídeo, três frames aleatórios são lidos e, utilizando a biblioteca *openCV*, os rostos são detectados, recortados das imagens e salvos no diretório "temp_frames_dir". Ao total, foram gerados 1203 recortes.

Depois, as faces salvas passam por algumas transformações de cor, dimensão e orientação. Além disso, o tipo de dado é alterado para Tensor, para assim ter compatibilidade com o Pytorch.

5.3 Divisão dos dados

Definida a base de dados, utilizamos a técnica de validação cruzada que divide a base em duas partes, uma para a fase de treinamento e outra para a de teste, utilizando a biblioteca *scikit-learn*. As proporções das divisões dependem do número de divisões (folds) escolhido.

5.4 Treinamento

Após tratar os dados, estes são lançados para treinamento em uma tarefa de classificação.

Redes neurais serão utilizadas com a arquitetura VGG16 usando um modelo já pré-treinado. VGG-16 é uma rede neural convolucional profunda. Ela é composta por cinco grupos de camadas de Convolução e de Max Pooling, seguidas por três camadas fully conected. Além disso, no final a acurácia e a função de perda CrossEntropyLoss foram calculadas.

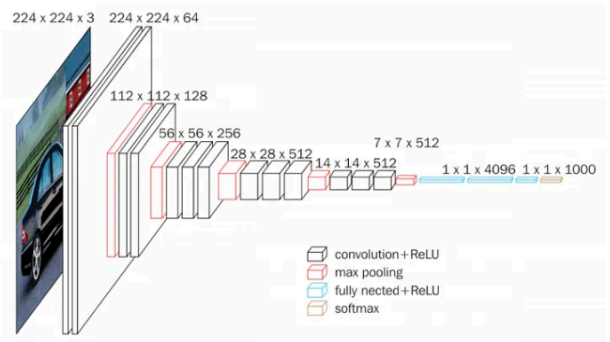

Fig 1: Ilustração:VGG16

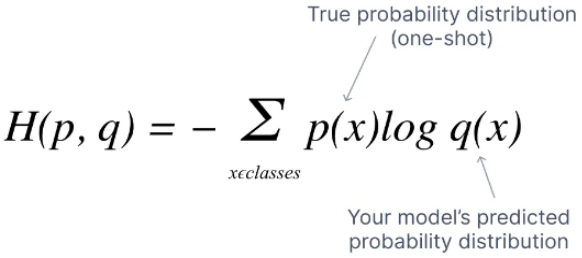

Fig 2: Funcionamento CrossEntropyLoss

$$Acurácia = \frac{Previsões\ Corretas}{Total\ de\ Previsões}$$

Fig 3: Funcionamento Acurácia

Ao todo 60 épocas foram utilizadas. Em cada uma dessas épocas, foi utilizado o algoritmo de otimização para gradiente descendente Adam (*Adaptative Moment Estimation*), que ajusta dinamicamente a taxa de aprendizado de cada parâmetro da rede.

5.5 Protocolo de validação

Cross Validation (CV) é uma técnica muito utilizada para avaliação de desempenho de modelos de aprendizado de máquina. O CV consiste em particionar os dados em conjuntos (partes), onde um conjunto é utilizado para treino e outro conjunto é utilizado para teste e avaliação do desempenho do modelo. A utilização do CV tem altas chances de detectar se o seu modelo está sobreajustado aos seus dados de treinamento, ou seja, sofrendo *overfitting*.

Para tornar nosso treinamento mais robusto e confiável, utilizamos o protocolo de validação cruzada "K-Fold". O K-fold consiste em dividir a base de dados de forma aleatória em K subconjuntos (em que K é definido previamente) com aproximadamente a mesma quantidade de amostras em cada um deles. A cada iteração, treino e teste, um conjunto formado por K-1 subconjuntos são utilizados para treinamento e o subconjunto restante será utilizado para teste gerando um resultado de métrica para avaliação (ex: acurácia). Esse processo garante que cada subconjunto será utilizado para teste em algum momento da avaliação do modelo.

Fig 4: Ilustração: K-Fold

5.6 Medidas avaliativas intermediárias

Para avaliar o desempenho do modelo usamos a função de perda para acompanhar seu comportamento conforme o passar das épocas. Seu mínimo observado foi em 30% na 22º época.

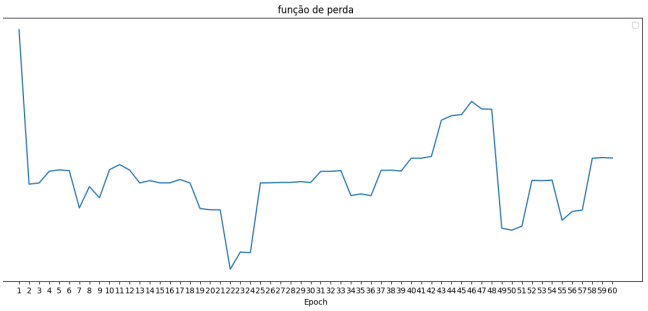

Fig 5: Gráfico Função de perda
(Valor x Época)

Além disso, nessa mesma época obtivemos a acurácia com um máximo de 90%.

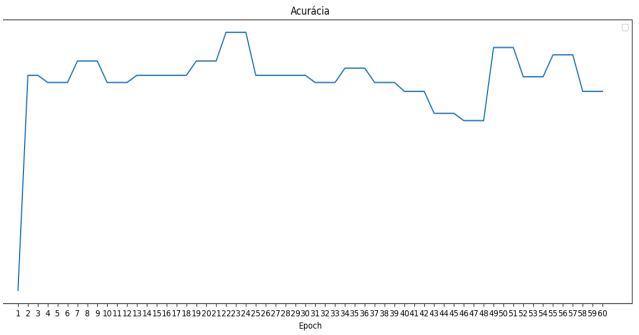

Fig 6: Gráfico de acurácia
(Valor x Época)

5.7 Early Stopping

Como observado, após a época 22, ocorre uma queda no desempenho do algoritmo, esta queda pode ser causada por diversos motivos, overfitting no conjunto de treino, confusão do algoritmo e até problemas na base de dados. Entretanto, em dado momento é possível observar que o algoritmo atinge medidas de acurácia de até 90%, sendo assim, neste instante os pesos estão muito bem calculados e aplicados, e com eles seria possível obter um algoritmo com grande capacidade de detecção. A partir deste momento é feito o Early Stopping, técnica que detecta as variações de desempenho do algoritmo e ao notar uma possível queda interrompe o treinamento e mantém os pesos até o melhor momento do treinamento.

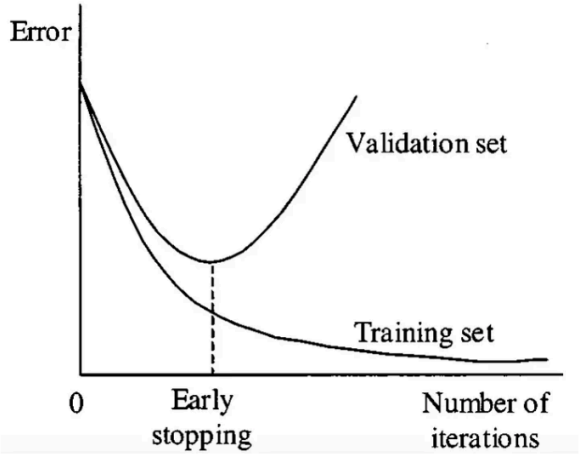

Fig 7: Ilustração early stop

5.8 Medidas avaliativas finais

Após toda a trativa de dados, treinamento, validação cruzada e aplicação do early stopping, concluímos com as seguintes métricas:

92% de acurácia na 22º época.

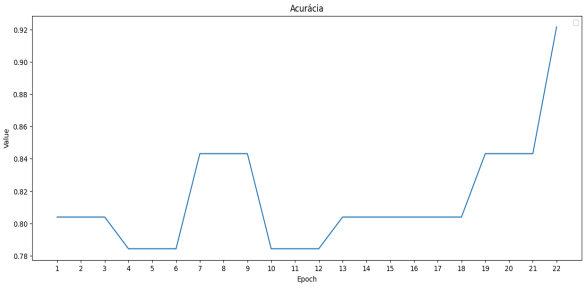

Fig 8: Gráfico de acurácia
(Valor x Época)

35% de perda na 22º época.

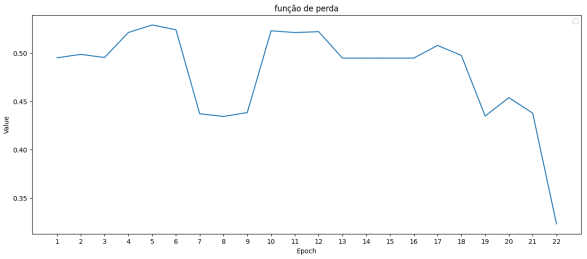

Fig 9: Gráfico de perda
(Valor x Época)

Como citado anteriormente no tópico 5.2, a base de dados já está balanceada na relação de dados rotulados com Falso e Verdadeiro, sendo assim as medidas expostas representam o resultado final.

6. Conclusão

Com base no estudo feito, é possível observar que a evolução das inteligências artificiais, em especial as redes neurais, acabaram por criar uma "faca de dois gumes". Ao mesmo tempo que diversas tecnologias surgiram para facilitar e salvar vidas, outras estão sendo usadas para criar confusão e tirar vantagem sobre grupos de pessoas.

No caso da criação de *deepfakes*, as redes neurais são a causa e a solução, mesmo sendo as responsáveis por criar as falsificações estas foram capazes de, com uma acurácia de 92%, detectar o que é real e o que foi criado artificialmente.

Estes dados acabam por mostrar que a inteligência artificial é capaz até mesmo de corrigir os problemas que ela mesmo causa, fator que impacta positivamente a sua aceitação em uma sociedade que discute o quão longe

devemos ir com o desenvolvimento desta tecnologia devido ao medo que suas capacidades de desenvolvimento causam.

O código fonte comentado e com as instruções necessárias para a compreensão do usuário está disponível neste link: <https://github.com/henriqueodaguiri/Deepfake-Detector>.

6.1 Trabalhos futuros

A continuidade do trabalho se dá pela evolução das redes neurais e consequentemente das técnicas de *deepfake* mais realistas e difíceis de serem detectadas, exigindo a constante renovação dos dados de treinamento e aperfeiçoamento das técnicas de detecção. Além disso, os vídeos adulterados não se limitam às modificações das faces, mas também de vozes. Sendo assim, para criar um algoritmo que tenha plena capacidade de detectar cenas criadas artificialmente, este deve ser capaz de saber não só quando ocorre alteração na imagem, mas também se a voz é verdadeira e ou é emulada. No trabalho apresentado focamos apenas na detecção da parte visual. A detecção de áudio implica em outras técnicas de análise.

Referências

[1] Chen, Heater; Magramo, Kathleen. Golpistas usam deepfake de diretor financeiro e roubam US$ 25 milhões. **CNN Brasil**, 2024. Disponível em: <https://www.cnnbrasil.com.br/economia/negocios/golpistas-usam-deepfake-de-diretor-financeiro-e-roubam-us-25-milhoes/>. Acesso em 26 de maio de 2024.

[2] O que é deepfake e como ele é usado para distorcer a realidade. **g1**, 2024. Disponível em: <https://g1.globo.com/tecnologia/noticia/2024/02/28/o-que-e-deepfake-e-como-ele-e-usado para-distorcer-realidade.ghtml>. Acesso em 26 de maio de 2024.

[3] Guera, David; Delp, Edward J. Deepfake Video Detection Using Recurrent Neural Networks. **IEEE Xplore**, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8639163>. Acesso em 26 de maio de 2024.

[4] Sharma, Jatin; Sharma, Sahil; Kumar, Vijay; Hussein, Hany S.; Alshazly, Hammam. Deepfakes Classification of Faces Using Convolutional Neural Networks. **ResearchGate**, 2022. Disponível em: <https://www.researchgate.net/profile/Hammam-Alshazly/publication/361526845_Deepfakes_Classification_of_Faces_Using_Convolutional_Neural_Networks/links/62e3cc373c0ea8788765d8d4/Deepfakes-Classification-of-Faces-Using-Convolutional-Neural-Networks.pdf>. Acesso em 26 de maio de 2024.

[5] Kolagati, Santosh; Priyadharshini, Thenuga; Rajam, V. Mary Anita. Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model. **ScienceDirect**, 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2667096821000471>. Acesso em 26 de maio de 2025.

[6] Deepfake Detection Challenge. **Kaggle**, 2024. Disponível em: <https://www.kaggle.com/competitions/deepfake-detection-challenge/data>