Gustavo Cunha Lacerda – Agosto 2021

GÜERA, David; DELP, Edward J. Deepfake video detection using recurrent neural networks. In: **2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)**. IEEE, 2018. p. 1-6.

O texto introduz que que a tentativa de trocar rostos em imagens é datada de 1865 e não é um conceito novo. Esse tipo de manipulação apenas cresceu de popularidade com a acessibilidade gerada pelos computadores atuais, que, com maior poder de processamento, permitiram a manipulação de rostos ao alcance de alguns cliques. Com essa facilidade e poder, o deepfake se tornou um problema para sociedade, pois é substancialmente utilizado em criação de pornografia e fake news.

Com o objetivo de rastrear essas deepfakes, principalmente em vídeos, esse trabalho propõe uma técnica, utilizando redes neurais para detectar inconsistências presentes na criação dos deepfakes.

O autor realiza uma análise da formação de deepfakes, no qual encoders são aplicados para redimensionamento, resgatando características de um rosto, para poder fazer, assim, a troca dessas características com outro rosto. No processo de troca dessas características nem sempre o rosto alvo e o original estão em iguais condições de luz e até formato de arquivo, o que dificulta os algoritmos de criação de deepfakes a gerarem uma imagem realista fidedigna. Essas falhas na criação podem ser alvos de algoritmos que tem objetivo de detectar as deepfakes.

Depois, no artigo, é proposto uma rede neural recorrente para detectar deepfakes. Esse sistema é composto por uma CNN para extração de características a cada frame do vídeo e uma LSTM para análise. Inspirado pelo IEEE Signal Processing Society Camera Model Identification Challenge, foi adotado o Inception V3 usando o modelo pré-treinado da ImageNet para extração das características que serviram como entrada para a rede LSTM. Já na rede LSTM foi usado os dados de entrada da CNN com 2 nós com a probabilidade de o vídeo ser parte do conjunto de deepfake ou do conjunto de vídeos reais.

Para analisar os resultados, o pesquisador utilizou 300 vídeos de diversos websites que armazenam vídeos e incorporou mais 300 vídeos aleatoriamente selecionados do banco de dados HOHA, totalizando 600 vídeos. Esses dados foram divididos randomicamente na proporção 70/15/15 para treino, validação e teste respectivamente. Além disso houve um pré-processamento de cada frame dos vídeos com esses parâmetros:

- Retirada da média do canal de cada canal
- Redimensionamento de cada frame para 299x299
- Quantidade de frames N = 20, 40, 80 para ter o controle de quanto foi necessário para ter uma detecção precisa
- Otimização feita para o Adam

Os resultados foram divididos em quantidade de frames e precisão por conjuntos de treino, validação e teste e são os seguintes.

Model	Training acc. (%)	Validation acc. (%)	Test acc. (%)
Conv-LSTM, 20 frames	99.5	96.9	96.7
Conv-LSTM, 40 frames	99.3	97.1	97.1
Conv-LSTM, 80 frames	99.7	97.2	97.1

Concluindo assim, que esse modelo usando redes convolucionais é capaz de detectar um vídeo gerado por deepfake com menos de 2 segundos e com precisão acima dso 97%.