A Survey on Deep fake Detection Methods

Darshan V Prasad[1], Harsha M[2], N Navneeth Krishna[3], Sanjay T.C[4], Dr.Kiran Y C[5]
[1 2 3 4] *Student, Global Academy of Technology, Bengaluru*
[5] *Professor and HOD,Department of Information Science and Engineering, Global Academy of Technology, Bengaluru*

Abstract-**Deep learning algorithms have recently expanded their applications beyond big data analytics to include intrusion detection systems. Artificial intelligence and image processing advances are changing and challenging how people interact with digital images and video, and because of their intrinsically contentious character and the reach of contemporary society, they are intended to propagate harmful content and disinformation to millions of people.**

A picture may say a thousand words, but what if the photograph has been fabricated? The term "fake news" has recently gained popularity, yet with today's photo manipulation techniques, even the most vigilant eyes can be tricked.One of these areas is the use of several software such as faceapp and fakeapp to create modified media files known as deepfake. From massive data analysis to human biometric systems, deep learning algorithms are used.

Due to their user-friendly characteristics, these applications are growing more popular with the general public and are employed in a range of fields including digital fraud, cybercrime, politics, and even military actions.As a result, it's critical to build detection technologies that can detect and remove this form of forgery, as well as to take a new step forward in video and audio forensics. In this article, we discuss different detection and production methodologies currently used in deepfake research.

INTRODUCTION

Deepfakes are synthetic media created with software that portray humans talking or doing things they don't actually do. During the early stages of video and audio manipulation, several approaches in the field of image processing were developed.

Anyone with even a single ounce of Photoshop skill can alter images to modify their contents, interpretation, and perhaps everything. However, this type of fabrication is also being widely researched in recent years, and commercial tools that can identify and explain it are also available. However, when compared to modern deep learning generation and detection systems involving auto encoders and deep convolution networks, the degree of accuracy has been quite poor.

In most cases, these learning phases necessitated a huge number of photos or videos in order to train the model to make photorealistic imitation duplicates. Typically, this dataset for free usage is created from publicly available photographs or videos of politicians and celebrities, which is why these people are the first targets of deepfake. The most well-known deepfake video was the one that resulted from Barack Obama's speech, which went viral on the internet and had an influence on political election in the United States. As a result, it has been classified as a national security danger since Deepfake is being used to create movies of various political figures for the purpose of political manipulation. This was even utilised to create a fictitious satellite picture of the Earth with an item that does not exist. This was done to deceive the military analysts and even lead a squad to cross a bridge in the middle of a conflict. Latest technological advancements have enabled the creation of deepfake films using only a single picture, posing a major threat to civilization. Along with the drawbacks, these deepfakes have several positives, particularly in sphere of media creation, where they may reproduce films of individuals who've lost their voices or update episodes without having to redo them. The groundwork for the creation of Deepfake in the latest scenario employing Deep Learning's Generative Adversarial Network (GAN).

There are many GAN methods that is used to create deepfake images. Some of the GAN methods are StyleGAN, stackGAN, W-GAN, SR-GAN etc. The research in this field has lately released a number of deepfake datasets to aid other researchers in creating detection techniques for these deepfakes in an attempt to improve the quantity of data accessible.

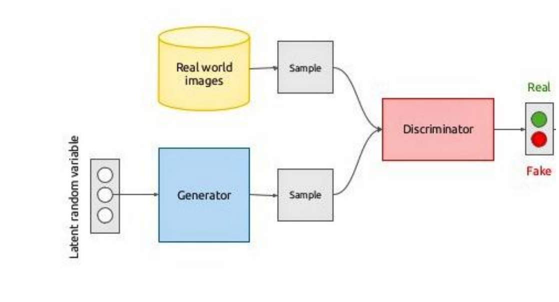

Fig-1 GAN System Model

As illustrated in fig 1, the deepfake video is generally made by employing two GAN networks which are based on the AI system. The first is known as the Generator, while the second is known as the Discriminator. Essentially, the generator is used to create deepfake videos, while the discriminator decides whether or not the video is fake. Each time the discriminator correctly recognises a video as fake, it provides a hint to the generator on how not to make the next Deepfake video.The generator and discriminator will combine to build a Generative Adversarial Network. The advancement of these deep learning techniques, which are employed in the execution of most of the online Deepfake generators, as well as their ease of use, has increased their popularity among both professionals and amateurs. Deepfake was created and generated using a variety of deep learning methods, including convolutional neural networks, recurrent neural networks, long short-term memory, and even a blend of these approaches. Finding the credibility of these digital evidences is a major issue for media forensics experts and investigators. To encourage greater research and development in the detection and prevention of Deepfake. Facebook and Microsoft have announced a Deepfake detection challenge. Google also sponsored a similar event by releasing a dataset (Google Net) for research purposes. More such benchmark datasets, on the other hand, aid in boosting performance and evolving methodologies. Nevertheless, it would be impossible to create a large dataset for each new deepfake creation method in order to train deep neural networks.As new sophisticated picture forging methodologies are released every day, distinguishing modified from authentic photographs is getting increasingly challenging. When trained on a specific counterfeiting method, naive classification systems based on Convolutional Neural Networks (CNNs) perform well at identifying changes in images and videos. On instances from unobserved manipulation techniques, however, their performance suffers noticeably

Author	Year of Publication	Methodology Used	Findings
Deressa Wodajo et al. [1]	2021	GAN, Convolutional Neural Network (CNN), Convolutional Vision Transformer (CViT) , Feedforward Neural Net- work (FFNN)	DeepFake Detection Challenge Dataset (DFDC)-91.5, AUC value of 0.91, loss value of 0.32
Mohammed Hasanet al.[2]	2020	Haar wavelet transform	Accuracy of 90%
Luca Guarnera et al.[3]	2020	Expectation – Maximization (EM) algorithm	Accuracy of 90%
Yushaa Shafqt Malik et al.[4]	2020	Using a CNN model around the face and compare the differences	95% accuracy when implemented on Xception Net architecture
Christoph Busch et al.[5]	2020	The UBM model used is the TensorFlow implementation of PixelCNN++.	DF- 99.30% F2F- 98.25% FS- 99.11% NT- 98.46%
OliverGiudice et al.[6]	2020	Expectation Maximization (EM) algorithm.	ATTGAN= 92.67% GDWCT = 88.40% STARGAN= 93.17% STYLEGAN=99.65% STYLEGAN2=99.8%
Sangyup Lee et al. [7]	2020	Detecting with consecutive video frames using a combination of both CNN and RNN.	Using a small dataset, the model was able to achieve a total of 96% score
Ricard Durall et al.[8]	2020	Frequency domain analysis i)Discrete Fourier transform ii)Azimuthal average and Classifier Algorithm, iii)Support vector machines	HQ-80%, LQ-91%

Omkar Salpekar et al.[9]	2020	CFFN (Common Fake Features Network)	This method yielded a training accuracy of 94% and validation accuracy of 91%.
Davide Cozzolino Justus Thies et al.[10]	2019	Forensic Transfer	Slower growth and a smaller gain is observed, with Xception- Net almost closing the gap at 100 shots, both with accuracy exceeding 90%.
Samuli Laine et al.[11]	2019	redesigned GAN FFHQ (HQ dataset) WGAN-GP	
Kritaphat Songsriet al.[12]	2019	Back Bone Architecture, Combining Classification and Localization branches, Adding facial landmarks, Model	Raw - 96.58%, HQ - 94.85%, LQ - 89.33%
Andreas Roessler et al.[13]	2019	XceptionNet and Support Vector Machine (SVM) classifier	DF - 96.36%, F2F - 86.86%, FS - 90.29%, NT - 80.67%, REAL - 52.40%
Han Zhanget al. [14]	2019	multi-stage StackGAN	9.55 +/- 0.11 (inception scoring) with color consistency

LITERATURE SUMMARY

Challenges

Open source face swapping software and applications result in a large number of Deepfake video clips that have a larger impact on social network. Identifying and screening such video clips content has become a challenging topic. The absence of quality Deepfake and actual video datasets which can be used for training datasets for research purposes is the major difficulty in the creation of a Deepfake detection algorithm and the datasets available will include only these two options: deepfake or real. Another important challenge is that the dataset is difficult to train on normal machines. High-end machines with excellent computation power is required.

CONCLUSION

Deepfakes are digitally edited clips of individuals doing or saying things they don't actually do.Visual inspection is insufficient to draw a conclusion on the validity, and available technologies to detect if the film has been tampered with are likewise insufficient. Because the visual appearance of Deepfakes will ultimately be so enormous that judging truthfulness only on the basis of visual validation would be challenging.The approach to solution here is to use of ingenious technologies to develop a methodology that can detect the deception in Deepfakes. As many deepFake technique can create images with finite resolutions and sizes, which must then be pixilated and tailored to suit the image that must be bartered with the original. The subsequent picture on the ROI creates unique artefacts in the deepfake image that may be managed to capture by analysing for variations the ROI and the remaining portion of the picture. The technique based on image quality metrics (IQM) and support vector machine (SVM) in CNN network may be employed for the classification of the video as real or false statistical correlations between the feature space for the extraction of the feature vectors related to the input.Future research might lead to the development of a system that might automatically detect deepfakes that employ an audio-visual method for detecting inconsistencies in facial movements and verbal in speech. Best method for deepfake detection of images and video will be identified.

REFERENCES

[1] Wodajo, Deressa and Solomon Atnafu. “Deepfake Video Detection Using Convolutional Vision Transformer.” (2021).

[2] M. A. Younus and T. M. Hasan, "Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform," 2020 International Conference on Computer Science and Software Engineering (CSASE), 2020.

[3] L. Guarnera, O. Giudice and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," in IEEE Access, vol. 8.

[4] Y. S. Malik, N. Sabahat and M. O. Moazzam, "Image Animations on Driving Videos with DeepFakes and Detecting DeepFakes Generated Animations," 2020 IEEE 23rd

International Multitopic Conference (INMIC), 2020.
[5] A. Khodabakhsh and C. Busch, "A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling," 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), (2020).
[6] L. Guarnera, O. Giudice and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (2020).
[7] Tariq, Shahroz & Lee, Sangyup & Woo, Simon. A Convolutional LSTM basedResidual Network for Deepfake Video Detection (2020).
[8] Durall López, Ricard & Keuper, Margret & Pfreundt, Franz-Josef & Keuper, Janis. (2019). Unmasking DeepFakes with simple Features (2020).
[9] Salpekar, Omkar. "DeepFake Image Detection." (2020).
[10] Cozzolino, Davide, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. "Forensic transfer: Weakly-supervised domain adaptation for forgery detection." (2019).
[11] T. Karras, S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence. (2019)
[12] Songsri-in, Kritaphat & Zafeiriou, Stefanos. Complement Face Forensic Detection and Localization with Facial Landmarks (2019).
[13] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
[14] H. Zhang et al., "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence. (2019).