Future - Oriented Deepfake Detection on Human Face: A Systematic Review

Dr.D.R.Jiji Mol[1], Dr. Deepak Gupta[2]
[1]Postdoctoral Researcher, Lincoln University College, Malaysia.
[1]Assistant Professor, SRM Arts and Science College, Tamilnadu, India. EMail:dr.jiji@gmail.com
[2]Professor, Maharaja Agrasen Institute of Technology, India. EMail:drdeepakgupta.cse@gmail.com

Abstract:

Rapid technological advancements and easy access to the web have allowed users and communities to interact with each other on social platforms. Combined with the progress in Generative Artificial Intelligence (GenAI) systems, it has allowed the production of digital content that has a realistic flavor. Due to the advancements in Generative Adversarial Networks (GAN), one can create fake images, audio and video streams of individuals or use their audio and visual information to fit other environments. With recent advancements in deepfake technology, it is possible to generate convincing deepfakes in real-time, therefore, deepfakes are specifically employed to spread fake information and propaganda on social circles that tarnish the reputation of an individual or an organization. Recently, many surveys have focused on generating and detecting deepfake images, audio, and video streams. The study discusses existing deepfake models, detection techniques and the future directions.

Keywords: Deepfake, GAN, Artificial Intelligence, fake content.

1. Introduction

The term “deepfake” is referred to as an AI based technology synthesizing the media [1]. The recent generation of fake material, such as counterfeit images and videos, with the help of artificial approaches has become more common. Deepfake combines two literals, Deep and fake, related to fake material generated by Deep Neural Networks (DNN)s, [2].

The audio/video generated data via the deepfake technique is genuine and realistic (as we can see in Figure 1) and not easily recognizable by the human being. In deepfake content, mostly facial features are swapped from one person with the targeted person in video or images and fabricated material to provide the false impression that the target has said mimicked that someone else has [3].

Deepfakes are predicted to take current sources of misinformation and disinformation to new heights, making them vulnerable to manipulation by foreign governments, trolls, bots, conspiracy theorists, and hyper partisan media known as fake news 2.0 [4].

As artificial intelligence technology advances, the generalization performance of images, audio and video deepfake detection systems has improved rapidly and significantly. However, these systems still face substantial challenges when confronted with adversarial attacks. Deep learning is frequently used to discuss, analyze, and estimate the problem of detecting deepfake videos. Every video is categorized as either real or a deepfake. It is possible to create a contradiction of this type while creating and testing deepfake detection algorithms using videos that are neither authentic nor produced using deepfake generation methods. However, the problem becomes more complicated when the detection technique is applied in real-world situations.

2. Technology Behind Deepfake

a) CNN

Convolutional Neural Networks is one of the core building blocks behind the deepfake. It is a kind of deep learning architecture used in computer vision and robotics.

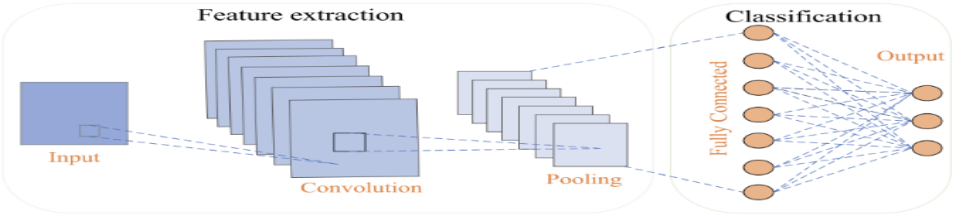

Figure 1: Architecture of CNN.

The basic architecture of CNN is shown in Figure 1. The CNN model consists of three layers: Convolutional, Pooling and Fully connected. In a convolutional model, an array of numbers called the kernel is applied across the inputs called tensor to construct the feature map. Two main steps of CNN training are forward propagation and backward propagation. In the forward propagation, the image moves through the network layer by layer. The kernel scans the image to detect the patterns like edges, shapes and textures. Based on these data, the network makes predictions. The gradient descent optimization technique updates the kernel and weight parameters during the back propagation to reduce the error.

The pooling layer helps the network make it simpler and more efficient. It reduces the size of the feature maps by downsampling operations. It is also called translation invariance, which decreases the number of subsequent learnable parameters. Finally the network reaches the fully connected layers. These layers act like a final decision maker of CNN. It takes all the extracted features to predict the result.

b) RNN

A Recurrent Neural Network is a type of neural network designed to work with sequential data like audio, video and text frames. It is useful in deepfake to detect time-based patterns. It takes the output of the previous step as an input to the next step. All the inputs and outputs in neural networks are independent of each other.

The significant aspect of RNN is its hidden layer. The hidden layer remembers certain information on the sequence. The memory in RNN stores the information about the calculations. Long short-term memory (LSTM) handles long term dependencies in sequential data on RNN. Figure 2 shows the basic architecture of RNN.

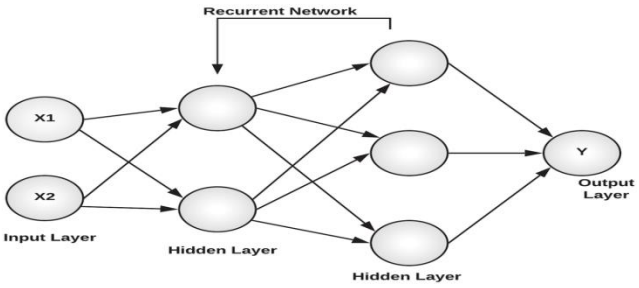

Figure 2: Architecture of RNN

c) GAN

The Generative Adversarial Network (GAN) is a back bone for generating high quality deepfake contents. It consists of two important components: Generator (G) and Discriminator (D). The basic structure of GAN is shown in Figure 3. The generator G captures the data distribution and Discriminator D estimates the probability sampling from the training data.

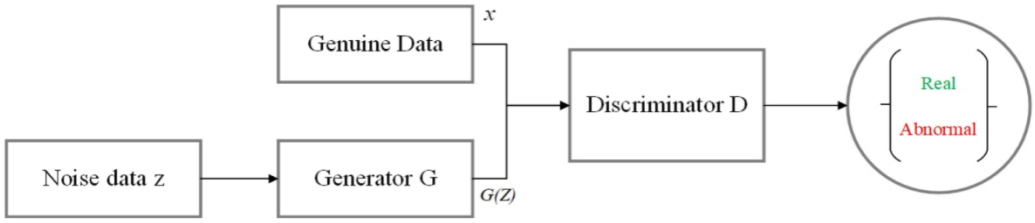

Figure 3: Architecture of GAN

3. Deepfake creation tools

The areas of digital face images and video manipulation are of leading interest because they use the power of GANs, which are capable of producing very realistic results. However, GANs still have challenges in establishing disentangled and controllable syntheses, particularly in the high-resolution domain.Several spoof videos created using GAN-based face-swapping techniques have been uploaded to YouTube and other video sites.

The AI-based generation of DeepFakes has a wide range of applications in the computer vision and graphics industries, including human face synthesis and stunning uploads to YouTube and other scenery productions[5]. Increasing numbers of face-swapping, face synthesis, face reenactment and attribute manipulation-based applications are becoming popular. Basically, face manipulation methods can be divided into five types [6]: entire face synthesis, identity swap, attributes manipulation, expression swap and miscellaneous.

a) Entire Face Synthesis

The entire face synthesis method generates new faces which do not exist. It is created by using GAN technology. Examples for this approach are Style-GAN [7], Style-GAN2 [8] and StyleGAN2-Ada [9]. These approaches produce realistic face images with high resolution. These kinds of manipulations are useful for 3D modelling, gaming and various business applications.

b) Identity Swap

Identity swap method is the most popular face replacement method from one person to another. The swapping can be done by using two types such as Graphics based approach or Deep learning technique based approach. This manipulation method is useful for the entertainment industry.

c) Attribute Manipulation

Attribute manipulation method is also known as face editing or face retouching. It changes the different aspects of a face such as gender, age, skin tone and hair type [10]. This method is most commonly used in virtual environments such as makeup, hairstyle and spectacles.

d) Expression Swap

Expression swap method is also called face reenactment. It swaps or modifies facial expressions of a person. The familiar expression swap techniques are Face2Face [11] and Neural textures [12]. These techniques swap the facial expressions of one person to another in a video.

e) Miscellaneous

Apart from these methods, deepfake can also be created by face morphing, face deidentification and audio-to-video (A2V) and text-to-video (T2V) [13].

4. Deepfake detection Techniques

The deepfake detection techniques improve the confidentiality and integrity of multimedia contents.

a) Facial anti-spoofing technique

A variety of anti-spoofing techniques are used in the detection of deepfake. It uses facial biometrics such as eye blinking, fear, disgust, happiness, sadness and anger. Fine-tuning CNN model and data augmentation are used to identify the facial expression [14],[15]. By combining the techniques called preprocessing, feature extraction and classification the original face is identified [16].

b) Inherent statistical information

In fact, statistical data that relates to the source multimedia image, which shows the high level of uniformity among different images are used. As a result, inherent statistical characteristics of the image are used in order to identify any kind of fabricated areas in the image.

c) Pixel-level anomalies

Pixel-level anomalies are tiny inconsistencies in images or videos, which are invisible to human eyes but can be detected by AI models and forensics methods. In general, during the creation of the deepfake entire images or videos may not perfectly blend with the background naturally. Detection algorithm uses pixel patterns, color distributions and frequency signals to detect the deepfakes. The Photo Response Non-Uniformity (PRNU) is a famous facial retouching and face morphing detection technique [17].

d) Deep Neural Network

In the detection part of deepfake images and videos, the convolutional neural network (CNN) is playing a vital role, which is a part of DNN. It analyses the facial texture, lighting inconsistencies, blending boundaries and pixel irregularities. Some models also examine the temporal values such as facial features over time. If there are any changes or inconsistencies in the frames, the DNN model results it as fake. In multimodal detection, the DNN processes both audio and video together, whether the lip movements are really aligned with the speech patterns. The ResNet[18], DenseNet[19], FDFtNet [20], and SSTNet [21] are the best DNN detection models.

e) Artifact Analysis

Deepfakes frequently produce artifacts that are difficult to recognize by humans but can be easily identified by machine and forensic analysis. There are two two types of artifacts: Spatial and temporal artifacts. Irregularities in the background and GAN fingerprints are the spatial artifacts and behavior of a person, physiological signal; video frame synchronization and coherence are the temporal artifacts. By analyzing these parameters the algorithms easily find out the deepfakes.

5. Challenges in deepfake detection

Although deepfake detection technology has improved in recent years, still there are some challenges that need to be addressed due to the evolving nature of its generation techniques and practical constraints. The primary challenge is generalizability, as the models trained on a particular deepfake tools or datasets fail to recognize such differences when alternate type of manipulations are involved. Detection models are susceptible to subsequent processing of images such as compression, cropping, resizing or blurring. These processes are often prevalent on social platforms which destroy small details and impair performance. The huge compression results extensive data loss making it difficult to detect. Video detection is complicated further by temporal inconsistencies.

High quality, large and diverse datasets are required to train the models, yet few are available. It is not suitable for real-world scenarios. Most of the neural networks with deep learning are black box in nature. There is no reason behind their decisions and inappropriate when it comes to forensics where reasoning a decision must exist. Due to the resource limitation in smartphone environment, real-time

detection is difficult to achieve the high accuracy, intensive computation and quick response. Some other difficulties also faced by deepfake detection techniques are adversarial attacks, ethical issues, unlabeled data and multimodal content.

6. Future Directions

As deepfake technology becomes more advanced, the methods used to detect must also be improved. The deepfake detection must focus on simple, smart and reliable systems. Instead of training the system to detect a specific type of deepfakes, we can develop generalized models. Moreover the scalable models are required to address the high rate of diffusion of deepfakes on social media and real time inference to avoid the irreversible damage. Future studies need to focus on training the model on diverse and multi-generational datasets, which enhance generalizability.

Temporal aggregation should be investigated as it provides inter-frame consistency. Resilience can be further promoted by multi-modal systems that combine the use of audio-visual cues. In future, we can also explore biometric and physiological signals in the model. Combining reinforcement learning with game theory helps to detect anti-forensic attacks. There is also an advanced detection technique required for real-time deepfake detection, especially in live streaming and social media platforms. This must be fast and accurate. We can also integrate explainable AI methods to highlight manipulated regions for forensic evidence.

7. Conclusion

Deepfake technology has rapidly evolved due to the advancements in technology like CNN, RNN, GAN, LSTM. This article focused on various methods used for the creation of synthetic images, audio and video. Moreover, the deepfake detection techniques were also discussed. Various spatial, temporal, and multimodal detection approaches are analyzing facial artifacts, motion inconsistencies, audio-visual mismatches, and physiological signals. However, the detection process remains challenging due to rapidly improving generation methods, limited and evolving datasets, adversarial attacks, and the need for real-time analysis. This article also provides valuable insights into the challenges and opportunities, as well as the trends and directions for further exploration in the field of deepfake generation and detection.

References

1. A.M.V. Lalla and N.Y.Z. Harned, Artificial Intelligence: Deepfakes in the Entertainment Industry. Available online: https://www.wipo.int/wipo_magazine/en/2022 /02/article_0003.html, 2024
2. R. Gil, J. Virgili-Gomà, J.-M. López-Gil, and R. García, ‘‘Deepfakes: Evolution and trends,’’ Soft Comput., vol. 27, no. 16, pp. 11295–11318, Aug. 2023.
3. M. Albahar and J. Almalki, ‘‘DeepFakes: Threats and countermeasures systematic review,’’ J. Theor. Appl. Inf. Technol., vol. 97, no. 22, pp. 3242–3250, 2019.
4. Z. Akhtar, ‘‘Deepfakes generation and detection: A short survey,’’ J. Imag., vol. 9, no. 1, p. 18, Jan. 2023.
5. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ‘‘Generative adversarial nets,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 27, pp. 1–9, 2014.
6. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, ”Deepfakes and beyond: A survey of face manipulation and fake detection,’’ Inf. Fusion, vol. 64, pp. 131–148, Dec. 2020.
7. T. Karras, S. Laine, and T. Aila, ‘‘A style-based generator architecture for generative adversarial networks,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4401–4410, Jun. 2019.
8. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, ‘‘Analyzing and improving the image quality of StyleGAN,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8110–8119, Jun. 2020.

9. T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, ‘‘Training generative adversarial networks with limited data,’’ arXiv:2006.06676, 2020.
10. Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, ‘‘Star-GAN: Unified generative adversarial networks for multi-domain image-to-image translation,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 8789–8797, Jun. 2018.
11. J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, ‘‘Face2Face: Real-time face capture and reenactment of RGB videos,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2387–2395, Jun. 2016.
12. J. Thies, M. Zollhöfer, and M. Nießner, ‘‘Deferred neural rendering: image synthesis using neural textures,’’ ACM Trans. Graph., vol. 38, no. 4, pp. 1–12, 2019.
13. S. Agarwal, H. Farid, O. Fried, and M. Agrawala, ‘‘Detecting deep-fake videos from phoneme-viseme mismatches,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 660–661, Jun. 2020.
14. S. Umer, R. K. Rout, C. Pero, and M. Nappi, ‘‘Facial expression recognition with trade-offs between data augmentation and deep learning features,’’ J. Ambient Intell. Humanized Comput., vol. 13, pp. 721–735, Jan. 2021.
15. S. Hossain, S. Umer, V. Asari, and R. K. Rout, ‘‘A unified framework of deep learning- based facial expression recognition system for diversified applications,’’ Appl. Sci., vol. 11, no. 19, p. 9174, Oct. 2021.
16. S. Umer, B. C. Dhara, and B. Chanda, ‘‘Face recognition using fusion of feature learning techniques,’’ Measurement, vol. 146, pp. 43–54, Nov. 2019.
17. M. Koopman, A. M. Rodriguez, and Z. Geradts, ‘‘Detection of deepfake video manipulation,’’ in Proc. 20th Irish Mach. Vis. image Process. Conf. (IMVIP), pp.133–136, Aug. 2018.
18. K. He, X. Zhang, S. Ren, and J. Sun, ‘‘Deep residual learning for image recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778, Jun. 2016.
19. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ‘‘Densely connected convolutional networks,’’ in Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR), pp. 4700–4708, Jul. 2017.
20. H. Jeon, Y. Bang, and S. S. Woo, ‘‘FDFtNet: Facing off fake images using fake detection fine-tuning network,’’ in Proc. IFIP Int. Conf. ICT Syst. Secur. Privacy Protection. Cham, Switzerland: Springer, pp. 416–430, 2020.
21. X. Wu, Z. Xie, Y. Gao, and Y. Xiao, ‘‘SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features,’’ in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 2952–2956, May 2020.