Volume 9, Issue 05, May 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

A Review on Deep Fake Image Detection Technique

Yukta R. Alizad
MSC (Computer Science)
Indira College Malegaon

Mr. A.D.Patil
MSC (Computer Science)
Indira College Malegaon

Abstract-: ***Generative adversarial networks (GANs) are capable of synthesizing photo-realistic images from low-dimensional random noise. These artificially created images, often containing inappropriate content, can circulate on social media platforms, leading to significant repercussions. To combat this issue, an effective and efficient image forgery detector is imperative. However, conventional detectors struggle to identify fake images generated by GANs due to their manipulation of source images. In response, this paper proposes a deep learning-based approach for detecting fake images utilizing contrastive loss. Initially, multiple state-of-the-art GANs are employed to produce pairs of fake and real images. Subsequently, a modified DenseNet architecture is developed, structured as a two-stream network, to incorporate pairwise information. The proposed common fake feature network is then trained using pairwise learning to differentiate between features of fake and real images.***

Keywords: Deepfakes, Deep Learning; GAN; contrastive loss; deep learning; pairwise learning

I. INTRODUCTION

This paper explores a range of deep learning methodologies tailored for automatic classification and detection of deep fake images. Specifically, FaceForensics constitutes the primary dataset for training two neural networks, Xception and MobileNet, with pre-processed images. Each network yields four distinct models corresponding to major deepfake software platforms: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The evaluation of Modeland #039 ratings reveals remarkable accuracy in discerning between authentic and manipulated videos. However, this precision is notably sensitive and heavily reliant on the deep forgery platform utilized. To address this issue, we propose a voting mechanism that aggregates outputs from different models, offering a more consistent and effective solution (Deep fake Detection through deep learning).

As outlined in the Malicious Artificial Intelligence Report [11], it's imperative to acknowledge the potential duplicity in such endeavors, as misuse can skew research priorities and standards. Given the gravity of malicious attack vectors associated with deep fakes, we introduce a novel solution for detecting such videos.The key contributions of this work are as follows. Firstly, we advocate for a two-step CNN analysis to extract features at the frame level, followed by a time-aware RNN to capture temporal inconsistencies arising from facial alteration processes. Secondly, we evaluate the proposed method using a dataset of 600 videos, half of which comprise deep fakes sourced from various video hosts. Thirdly, we empirically demonstrate the efficacy of our approach in identifying suspect videos manipulated by deep fakes with an accuracy of 94%, outperforming random detection baselines in a balanced setup (deep fake detection using recurrent neural network).

As the 2020 US election nears, the proliferation of fake videos has garnered significant media attention. In an era plagued by fake news, concerns regarding the authenticity of online content continue to escalate. To mitigate these challenges, Facebook and Instagram implemented policies in January 2020 aimed at curbing the dissemination of misleading "deepfake" videos during elections [1]. However, the success of this approach hinges on the ability to accurately differentiate between genuine and manipulated videos, which constitutes the focal point of this paper ("Deepfake Detect through Deep Learning"). Technological advancements have unfortunately facilitated the misuse of fake technologies, particularly in generating explicit content featuring celebrities and politicians. This trend has exacerbated the spread of propaganda and misinformation, contributing to numerous societal issues [4]. (Reference: "DeepVision: Deep Fake Detection Using the Blink Pattern of the Human Eye")

II. LITERATURE SURVEY

In the paper [1]" Deepfake Video Detection Using Recurrent Neural Network", David Guera and Edward J Delp propose a temporalaware pipeline to automatically detect deepfake videos. In order to detect deepfake videos, firstly we need to have a clear knowledge of how it is created, which helps us to understand the weak points of deepfake generation so that by exploiting those weak points, deepfake detection can be done. In the approach discussed in this paper, framelevel scene inconsistency is the first feature that is exploited. If the encoder is not aware of the skin or other scene information, there will be boundary effects due to a seamed fusion between the new face and the rest of the frame which is another weak point. The third major weakness that is exploited here is the source of multiple anomalies and leads to a flickering phenomenon in the face region. This flickering is common to most of the fake videos. Even though this is hard to find with our naked eye, it can be easily captured by a pixellevel CNN feature extraction. Dataset used here contains 300 videos from the HOHA dataset. Preprocessing steps are clearly described in this paper. Here the proposed system is composed of a convolution LSTM structure for processing frame sequences. CNN for frame feature extraction and LSTM for temporal sequence analysis are the 2 essential components in a convolutional LSTM. For an unseen test sequence, set of features for each frame

are generated by CNN. After that features of multiple consecutive frames are concatenated and pass them to the LSTM for analysis which finally produces an estimated likelihood of the sequence being either a deepfake or nonmanipulated video. With less than 2 seconds this system could accurately predict if the fragment being analyzed comes from a deepfake video or not with an accuracy greater than 97 percentage.

In the paper [2] "Effective and Fast Deepfake detection method based on Haarwavelet Transform" by Mohammed Akram Younus and Taha Mohammed Hasan describes another method to detect deepfake videos by haar wavelet transform. The method described here take the advantage of the fact that during deepfake video generation, deepfake algorithm could only generate fake faces with specific size and resolution. In order to match and fit the arrangement of the source's face on original videos, a further blur function must be added to the synthesized faces.This transformation causes exclusive blur inconsistency between the generated face and its background outcome deepfake videos. The method detects such inconsis-tency by comparing the blurred synthesized areas ROI and the surrounding context with a dedicated Haar Wavelet transform function. The two main advantage of this Haar Wavelet transform function is that it first distinguishes different kinds of edges and the retrieves sharpness from the blurred image. It is very effective and fast since the uniform background of the faces in the images will have no effect and it does not need to reconstruct the blur matrix function. To estimate the blur extend, two methods such as direct and indirect can be used. Direct method can measure the blur function extent by testing some distinctive features in an image. Eg: edge feature. The indirect method depends on the blur reconstruction function when the H matrix is unknown (H matrix is blur's estimation and blur identification). Dirac structure, Step structure, and Roof structure are the different types of edges present in an image.A blur extends is identified by taking the sharpness of roof structure and G step structure into account. The sharpness of the edge is indicated by the parameter (0¡¡ /2), if is larger means the edge is sharper. By comparing the blur extent of the ROI with the blur extend of the rest of the image, we can determine if the images(frames of video) have tampered or not. UADFV dataset which contains 49 unmanipulated and 49 manipulated videos is used here. Videos are divided into frames and from each frame, the face region is extracted and deepfake detection algorithm using haar wavelet transform is applied. This algorithm is clearly described in this paper. This proposed model contains an accuracy of 90.5 percentage

In the paper [3]," OC Fake Dect: Classifying Deepfakes using OneClass Variational Autoencoder" by Hasam Khalid and Simon S. Woo, the proposed model needs only real images for training. As new methods for deepfake video creation are increasing today due to technology advancement, for a model to detect such videos, datasets containing fake videos are very scarce for training. It affects the model's accuracy. But in the model proposed in this paper needs only real videos for training so that it can overcome data scarcity limitation. FaceForensic ++ is the dataset used here. It contains real images and 5 sets of fake images: FaceSwap dataset, Face2Face dataset (F2F), Deepfake dataset(DF), Neural Tex-tures dataset (NT), Deepfake detection Dataset(DFD), After collecting the video datasets, they are converted into frames and face detection and alignment is done using MTCNN. One class variational encoder is used here. It consists of an encoder and a decoder. At the encoder side, image is given as input, and scaling is done using convolutional layer and mean and variance is calculated and the result is given as input into decoder and the RMSE value is calculated which is low for real image and high for fake images. Two methods are discussed in this paper: OCFakeDect1 and OCFakeDect2. In OCFakeDect1 from input and output image itself, reconstruction score is computed directly and in OCFakeDect2 contains additional encoder structure which computes reconstruction score from input and output latent information. Eventhough it has 97.5 percentage accuracy, better performance is only on NT and DFD datasets.

In the paper [4] "Deep Fake Source Detection via Interpreting Residuals with Biological Signals", Umur Aybars Ciftci, Ilke Demir and Lijun Yin presented a deep fake source detection technique via interpreting residuals with biological signals. To their knowledge it is the first method to apply biological signals for the task of deep fake source detection. In addition to this they had experimentally validated this method through various ablation studies their experiments had achieved 93.39accuracy on FaceForensics++ dataset on source detection from four deep fake generators and real videos. Other than this they had demonstrated the adaptability of the approach to new generative models, keeping the accuracy unchanged. After studying biological signal analysis on deepfake videos, it is found that ground truth PPG data along side original and manipulated videos enabled new direction in research on deepfake analysis and detection.

In the next stage of their work, . With ground truth PPG, they planned to create a new dataset with certain distribution variation as well as source variations. It is worth noting that these work looks for generator signatures in deep fakes, while the prevailing work reported by Ciftci et al. [23] looks for signatures in real videos. For detecting signatures on both real and fake videos, a holistic system combining these two perceptives can be developed. They posed this idea for their immediate future work.

In the paper [5] " Digital Forensics and Analysis of Deep-fake Videos" by Mousa Tayseer Jafar, Muhammed Ababneh, Muhammad Al-Zoube, Ammar Elhassan proposed a method detect deepfakes using mouth features.Nowadays deepfake videos can have an adverse effect on a society and these videos can challenge a person's integrity. Deepfake is a video that has been constructed to make a person appear to say or do something that they never said or did. Therefore there shows the increase in demand to detect methods to identify deepfakes. In this proposed model mouth features is used to detect deepfake video. A deepfake detection model with mouth features(DFT-MF),using deep learning approach to detect deepfake videos by isolating analyzing and verifying lip/mouth movement is designed and implemented here. Here, dataset contains the combination of fake and real videos. Some preprocessing is done prior to performing analysis. Then the mouth area is been cropped from a face. There will be fixed coordinates for face. Working on a typical image frame facial landmark detector is used to estimate the location of 68 (X,Y)coordinates. In next step all face containing closed mouth is excluded and face with only open mouth is been tracked having teeth with reasonable clarity. CNN is used to classify videos into fake or real based on a threshold number of fake frames based on calculating three variable word per sentence, speech rate and frame rate. If the number of fake frames is greater than 50 the video is been classified as fake or else as real.

III. System Architure

The methodology involves a two-step learning process depicted in Figure 1, combining the Common Fake Feature (CFF) network based on pairwise learning with classifier training. By integrating supervised learning into fake face image detection, challenges associated with collecting training samples from various GANs and the necessity of retraining the detector for new GAN-generated fake face images are

addressed. To tackle these challenges, fake and real images are paired, leveraging pairwise information to construct the contrastive loss for learning the discriminative Common Fake Feature (CFF) using the proposed CFFN. Once the discriminative CFF is acquired, the classification network utilizes it to determine the authenticity of the image. Further details regarding the proposed method are elaborated in subsequent sections.

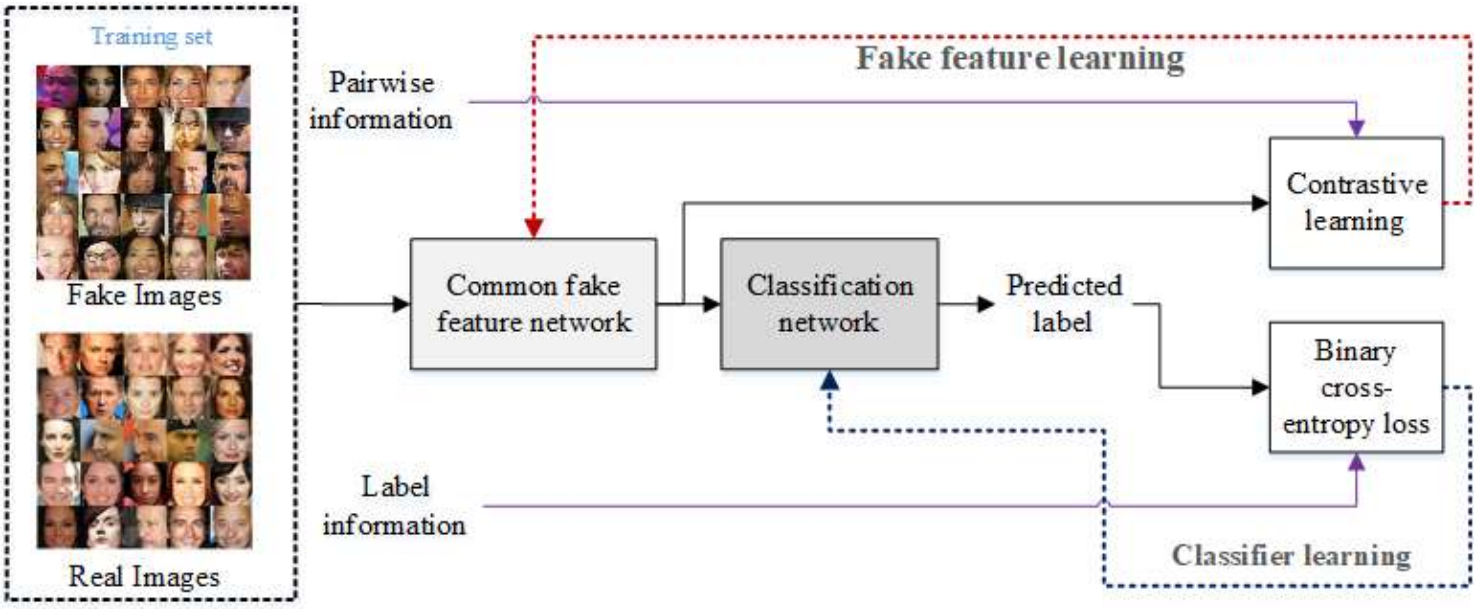

Figure 1. System architecture

CONCLUSION

Data scarcity of diverse deepfakes makes it difficult to train robust models. Continuously evolving deepfake creation techniques necessitate the development of adaptive detection models. Interpretability and explain ability of model decisions are crucial for building trust and accountability. In this paper, we have presented a brief review of some papers which describes different methods to detect deepfake videos and images. Also some of the methods can be modified or combined in our new project in order to get more accurate results than prevailing methods.

REFERENCE

1. D. Guera¨ and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp.
2. M. A. Younus and T. M. Hasan, “Effective and fast deepfake detection method based on haar wavelet transform,” in 2020 International Confer-ence on Computer Science and Software Engineering (CSASE), 2020, pp. 186–190.
3. H. Khalid and S. S. Woo, “Oc-fakedect: Classifying deepfakes using one-class variational autoencoder,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 2794–2803.
4. U. Ciftci, I. Demir, and L. Yin, “How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals,” 08 2020.
5. M. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, “Forensics and analysis of deepfake videos,” 04 2020, pp. 053–058.
6. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the International Conference on Learning Representations,Vancouver, BC, Canada, 30 April–3 May 2018.
7. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis.In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA,6–9 May 2019.
8. Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy,22–29 October 2017; pp. 2242–2251. [CrossRef]
9. Hsu, C.; Lee, C.; Zhuang, Y. Learning to detect fake face images in the Wild. In Proceedings of the 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 6–8 December 2018;pp. 388–391. [CrossRef]
10. Chang, H.T.; Hsu, C.C.; Yeh, C.H.; Shen, D.F. Image authentication with tampering localization based on watermark embedding in wavelet domain. Opt. Eng. 2009, 48, 057002.
11. Marra, F.; Gragnaniello, D.; Cozzolino, D.; Verdoliva, L. Detection of GAN-Generated Fake Images over Social Networks. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, Miami, FL, USA, 10–12 April 2018, pp; 384–389. [CrossRef]
12. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017;pp. 1610–02357.
13. Dang, L.M.; Hassan, S.I.; Im, S.; Moon, H. Face image manipulation detection based on a convolutional neural network. Expert Syst. Appl. 2019, 129, 156–168. [CrossRef]
14. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.

15. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv 2015, arXiv:1511.06434.
16. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
17. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved training of wasserstein gans. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5767–5777.
18. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Smolley, S.P. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017;pp. 2813–2821.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA,21–26 July 2017; pp. 4700–4708.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Liu,W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector.In European Conference on Computer Vision; Springer: Cham, Switzerland, 2016; pp. 21–37.
22. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego,CA, USA, 20–25 June 2005; Volume 1, pp. 539–546.
23. LeCun, Y.; Boser, B.E.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.E.; Jackel, L.D. Handwritten digit recognition with a back-propagation network. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 26–29 November 1990; pp. 396–404.
24. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
25. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1139–1147.
26. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 685–694.
27. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-Attention generative adversarial networks. In Proceedings of the 36th International Conference on Machine Learning; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR: Long Beach, CA, USA, 2019; Volume 97, pp. 7354–7363.