IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Review Paper on Deepfake Video Detection using Neural Networks

Arya Shah[1], Ashwin Thakur[2], Atharva Kale[3], Harsh Bothara[4], Prof. D. C. Pardeshi[5]
Students, Department of Artificial Intelligence and Machine Learning[1,2,3,4]
Professor, Department of Artificial Intelligence and Machine Learning[5]
All India Shri Shivaji Memorial Society Polytechnic, Pune, Maharashtra, India

Abstract: *With the increasing computational power, the creation of indistinguishable human synthesized videos, known as deepfakes, has become remarkably easy. These realistic face-swapped deepfakes have raised concerns as they can be utilized for malicious purposes such as causing political unrest, fabricating terrorism events, spreading revenge porn, and blackmailing individuals. In this research, we present a novel deep learning-based method that effectively distinguishes AI-generated fake videos from real ones. Our approach focuses specifically on detecting replacement and reenactment deepfakes. We harness the power of Artificial Intelligence (AI) to combat the challenges posed by AI itself. The core of our system lies in a ResNext Convolutional Neural Network, which extracts frame-level features. These features are then used to train a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) that classifies videos as either manipulated (deepfake) or authentic (real).To ensure real-time applicability and enhance the model's performance on real-world data, we evaluate our method using a large and balanced dataset.*

Keywords: Deepfake Video Detection, convolutional Neural network (CNN), recurrent neural network (RNN), Long short term memory (LSTM).

I. INTRODUCTION

The widespread use of social media has been greatly facilitated by the advancement in smartphone cameras and the availability of reliable internet connections, enabling easy creation and sharing of digital videos. Deep learning has gained immense power due to increased processing capabilities, surpassing what was once considered unimaginable. However, this progress has also brought about new challenges. The emergence of deep generative adversarial models capable of modifying audio and video samples has led to the creation of "DeepFake" videos. These videos are often spread through social media platforms, leading to issues such as spamming and the dissemination of false information. The existence of such DeepFake videos can have detrimental effects, causing intimidation and deception. Therefore, it is crucial to develop technologies that can effectively detect and identify these fakes, thereby preventing their spread online.

Our proposed technique is based on the same underlying process employed by GANs to generate DeepFake videos. We focus on a specific feature of these videos, where face photos synthesized by the DF algorithm undergo an affine warping process to align with the facial features of the source person while maintaining a fixed size. This warping introduces discernible artifacts in the resulting DeepFake video due to resolution discrepancies between the warped face area and the surrounding context. Furthermore, we capture the temporal inconsistencies between frames introduced by GAN during the video reconstruction process by splitting the video into frames, extracting features, and utilizing a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM). To streamline our approach, we directly train the ResNext CNN model to replicate the resolution inconsistencies observed in affine face wrappings.

To address this problem, we present a novel deep learning-based technique that successfully distinguishes between AI generated fake videos (DF Videos) and genuine ones. Understanding the workings of the Generative Adversarial Network (GAN) is pivotal in identifying DeepFake videos. GANs use input videos and images of a target person to replace their faces with those of another person (the "source"). Deep adversarial neural networks are trained on face photos and target videos to automatically map faces and facial expressions, forming the foundation of DeepFake creation. With appropriate post-processing, these generated videos can achieve a high level of realism. The GAN

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

replaces the input image in each frame by dividing the video into frames and then reconstructing the video using techniques like autoencoders.

II. RELATED WORK

Different types of deepfake detection methods are available today and each method has its own advantages and disadvantages. This paper tries to evaluate such methods from different papers and points out how these methods can be combined and modified in a new project in order to get more accurate results.

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

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

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 In- terpreting 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 accu- racy 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 some- thing 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 analysing 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. CONCLUSION

We present a novel solution that utilizes a neural network architecture for the classification of videos into deepfakes or real, providing a comprehensive measure of confidence in the model's predictions. coders. Our approach focuses on detecting deepfakes at the frame level, utilizing a ResNext Convolutional Neural Network (CNN), and extends to video classification using Recurrent Neural Network (RNN) in conjunction with Long Short-Term Memory (LSTM). By leveraging these techniques, our proposed method demonstrates the capability to identify whether a video is a deepfake or real, based on the parameters outlined in the associated research paper. We are confident that our method will yield a high level of accuracy when applied to real-time data. The combination of frame-level detection and video classification, along with the integration of deep learning components, positions our approach as a robust solution for discerning between authentic and manipulated videos. This has implications for enhancing the accuracy and reliability

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

of real-time video content analysis, contributing to the broader field of video forensics and authentication. This paper will be helpful to beginners for understanding the basic concepts of deepfake video detection.

REFERENCES

[1] Yuezun Li, Siwei Lyu, "ExposingDF Videos By Detecting Face Warping Artifacts," in arXiv:1811.00656v3.
[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu "Exposing AI Created Fake Videos by Detecting Eye Blinking" in arxiv.
[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen " Using capsule networks to detect forged images and videos".
[4] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng Xu "Deep Video Portraits" in arXiv:1901.02212v2.
[5] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin "Detection of Synthetic Portrait Videos using Biological Signals" in arXiv:1901.02212v2.
[6] Luisa Verdoliva. Media forensics and deepfakes: an overview. arXiv preprint arXiv:2001.06564, 2020.
[7] Martyn Jolly. Fake photographs: making truths in photogra phy. 2003.
[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.
[9] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.
[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
[11] An Overview of ResNet and its Variants : https://towardsdatascience.com/an-overview-of-resnet- and-its-variants-5281e2f56035
[12] Long Short-Term Memory: From Zero to Hero with Pytorch: https://blog.floydhub.com/long-short-term- memory-from-zero-to-hero-with-pytorch/
[13] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_mod els_tutorial.html
[14] https://discuss.pytorch.org/t/confused-about-the-image- preprocessing-in-classification/3965
[15] https://www.kaggle.com/c/deepfake-detection- challenge/data
[16] https://github.com/ondyari/FaceForensics
[15] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.
[16] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to- image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5967–5976, July 2017. Honolulu, HI.
[17] R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," in CVPRW. IEEE, 2017.
[18]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. 1–6.
[19]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.
[20]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.
[21]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.
[22]M. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," 04 2020, pp. 053–058.