Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Extensive Analysis of Deep Learning-based Deepfake Video Detection

D. Myvizhi[1], J. C. Miraclin Joyce Pamila[2]

[1]PG Scholar, Department of Computer Science and Engineering, Government College of Technology, Anna University, Chennai, India
[2]Professor, Department of Computer Science and Engineering, Government College of Technology, Anna University, Chennai, India

E-mail: [1]myvi.64453@gct.ac.in, [2]miraclin@gct.ac.in

Abstract

Deepfake is the practice of replacing an existing image or video with someone else's likeness. Currently, the spread of face-swapping deepfake strategies is increasing, producing a considerable range of naturalistic fake videos that cause danger to everyone. Due to the issue made by deepfake, recognizing between real and fake videos becomes a major issue. Deep learning is an efficient and useful approach for detecting deepfake videos and images. Research in recent years has focused on understanding how deepfake works and a number of deep learning-based approaches that are developed to do so. The aim of this study is to give an in-depth look at how several architectures work, along with the challenges encountered when detecting deepfake videos. It examines the existing deepfake detection methods using machine learning and deep learning approaches.

Keywords: Deepfake detection, face-swapping, deep learning, CNN, RNN

1. Introduction

As the use of social networks like Facebook, Instagram, Twitter, and YouTube, along with the availability of smartphones with exclusive cameras, has made video and image creation, sharing, and editing easier than ever before. The deepfake issue is one of the emerging problems in concurrent society. Deepfake has been extensively used to exchange celebrities' faces over pornographic videos and used to propagate inaccurate information and rumors. Face manipulation poses a threat to world security and poses a danger to the interaction between humans and biometric-based authentication and identification services. The manipulation of face frames, therefore, undermines trust in digital security applications.

DOI: https://doi.org/10.36548/jucct.2022.1.001
Received: 03.03.2022, received in revised form: 01.04.2022, accepted: 17.04.2022, published: 28.04.2022

Therefore, it become one of the major challenges to detect faces from images or videos. In the past, fake video synthesis methods were detected manually using artifacts and inconsistencies in the video synthesis process.

Deep learning plays a great role in deepfake detection. Deep learning is defined as the use of multiple layers hidden in the network. Deep learning, like artificial neural networks, uses a large number of bounded-size hidden layers to extract higher levels of information from raw input data. The number of hidden layers depends on the convolution of training data. If the dataset contains huge data to get the desired result, then there is a need for multiple hidden layers. A variety of methods are employed to detect the deepfakes, including the color of an eye, blinking of an eye, facial emotion, color of lips, movement of lips, voice, nose, etc.,

There are two types of video manipulation methods: those that use deep convolutional neural networks to identify visual divergence within frames and those that use deep recurrent neural networks to identify temporal divergence across frames. The creation of fake videos is widely increasing which becomes complex, and thus there increases the need of detecting deepfake. The detection model is framed in a manner that should give more accuracy and low loss. Deep learning techniques involve several architectures such as Convolutional neural network (CNN), Recurrent neural network (RNN), and Long short-term memory (LSTM).

CNN has shown great ability and scalability for image and video processing applications. It is mainly used for image processing, image segmentation, and for auto-correlated data. CNN has the unique capability of extracting features from images which can then be applied to many different applications. The architecture contains more layers such as input, output, and hidden layers for handling pixel data. The final CNN classification will generate better and more precise Deepfake detection models.

RNN architecture uses the previous layer output as input to the next layer. It has the ability to handle time-series information. It will retain memory for an elongated period. LSTM is a special kind of RNN as it contains several gates such as input, output, intermediate, and forget. LSTM contains memory cells for maintaining information that is performed in each gate. LSTM has the advantage of backpropagation.

This paper focused on the earlier approaches used for deepfake video detection, approaches, and challenges faced by the researchers. Section 2 includes a review of the video

detection process. Section 3 gives a comparative analysis of the existing methods. Finally, section 4 concludes this survey and gives the proposed idea to implement in the future.

2. Related Work

Aya Ismail et. al. [1] suggested a detection method for deepfake video using CNN architecture. For valid object identification, YOLO was employed. YOLOV2 is especially for selecting the face. It takes the input image probabilities and coordinates as bounding boxes. Several versions of YOLO are available. The network contains 164 layers for classifying the image features into specific categories. Each layer has the function to process the input and proceed to the next layer. Videos from Celeb-DF and FaceForensics++ are taken for the evaluation of the proposed model. The model is trained and tested using the merged dataset. Additionally, a gradient boosting classifier acts as a recognizer. The classifier performs well in classification so that it will reduce overfitting. The model produced an accuracy of 90.73%.

Marwa Elpeltagy et. al.[2] employed a face detector for detecting the faces. As YOLO is used to accurately identify the region on video. It will detect up to 22%. The EfficientNet model is designed to perform image classification. The model aimed to attain high accuracy and produce high efficiency through scaling. The vector is generated as an output. Along with EfficientNet, a Bidirectional LSTM structure is used that makes the information flow forward and backward. Therefore, the overfitting issues are decreased. It is examined on Celeb-DF and FaceForensics++. Using evaluation metrics of the confusion matrix, the model obtained an accuracy of 89.38%.

Mitra et. al. [3] presented a model based on CNN architecture. A comparative study was done on Xception, Inception, and Residual network. The demonstration was made on datasets such as Face Forensics++ and Deepfake detection challenge. Xception network contains 71 layers, InceptionV3 network contains 48 layers, and Residual network contains 50 layers. All these models load the pretrained version of image nets. As a residual network contains many layers, it can switch the connections between layers. These models are then combined with a classification network that contains a fully connected layer, a dropout layer with the ‘relu’ function, and a softmax layer. Among the three, the Xception with the classification model network produced a better result. Various compression levels were made

on the model. Therefore, it produced 96% accuracy at level 23 and 93% accuracy at level 40, and so on.

Guera D et. al. [4] proposed a system that involves both CNN and RNN. The function of CNN is to extract the features from videos and the function of RNN is to check whether the video is altered or not. The progress was carried on the dataset that was collected from multiple websites. The system accurately predicted whether a fragment of the video was taken from the deepfake or not with an accuracy of 97%. The model produced a better result for the video length of 2 seconds.

Yadav D et. al. [5] discussed the creation, disadvantages, and detection of deepfake techniques. Generative adversarial neural networks(GAN) contain generators and discriminators. The function of the generator is to create the deepfake images in the dataset. The image discriminator finds the fake images. The dataset should be high-resolution with matching faces, dress code, body language, and color for creating fake. The major disadvantage includes forgery, misuse, and threatening others. Deepfake was detected using 2 methods. The first is based on the eye that blinks. A person usually blinks within 10 seconds. Therefore blinking plays a major role in finding deepfake. The other method is with MesoNet structure. With this structure, it is able to find the difference accurately. MesoNet along with LSTM provided a high accuracy within the range of 95 to 98.

Irene Amerini et. al. [6] developed an optical flow method to identify the dissimilarities between each frame in the video. The method compares the current frame with the next frame to observe the difference that occurs between those frames. The work was done using FaceForensics++. CNN models such as VGG16 and ResNet50 models are tested. Both the models contain as much of layers for processing the input frame to binary classification. The final layer is the classification layer that will produce a binary output as real or fake. VGG16 produces an accuracy of 81.61% and ResNet produces an accuracy of 75.46%. Among both models, it was concluded that VGG16 is better than ResNet50.

Nguyen et. al. [7] made the study the techniques to create and detect deepfake in the videos. For creating fake videos using deep learning, the autoencoder was used. Two encoders and decoders are useful for creating deepfake. The temporal and visual characteristics help in detecting fake videos. Two classifiers such as deep and shallow are used to find the video either real or fake.

3. Comparative Analysis

Table 1 highlights the techniques, dataset, accuracy, and observation made for detecting Deepfake.

Table 1. Comparative Analysis based on the accuracy

Paper	Techniques used	Dataset used	Accuracy	Observation
[2]	YOLO (face detector) EfficientNet-B5 Bi-LSTM	CelebDF-FaceForensics++ (c23) merged	89.38%	Face detector YOLO is used for extracting faces. EfficientNet along with Bidirectional LSTM is used to extract features. The model will extract the spatial and temporal features for detecting deepfakes.
[1]	YOLOV3 (face detector), CNN(Inspection ResNetV2) XGBoost Classifier	CelebDF-FaceForensics++ (c23) merged	90.73%	YOLO is the object detector that will extract the faces from the videos. The spatial features on the surfaces are extracted by Inception ResNetV2. The model will detect the spatial information to find the originality of the video.
[7]	CNN LSTM Deep, Shallow classifier	Face2Face	92.5% (deep) 83.2% (shallow)	By using several layers of LSTM, accuracy may be improved. The model is compared with deep and shallow classifiers. LSTM along with the deep classifier gives better accuracy.
[8]	Spatial Transformer Network (STN) Multi Recurrent Networks	FaceForensics++	94.35%	STN is combined with RNN to classify the sequences. STN will train the video fastly. The model is not trained for larger datasets as

	Bi-Directional RNN			well as long video clips.
[5]	CNN (MesoNet) RNN (LSTM)	Face2Face, Reddit user deepfakes	95%-98%	GAN is based on the min-max method. Meso-4 and MesoInception-4 architectures are used to detect facial video forgery. The model gives high accuracy and they can separate between image properties. LSTM will detect the changes in the frames easily.
[4]	CNN RNN (LSTM)	HOHA (600 videos)	97%	For training, validation, and test videos, the author extracts the contiguous sub-sequence of the fixed frame length for input. The model gives better accuracy in a short time.

4. Conclusion

This paper presents a survey on deepfake video detection. As deepfake video creation techniques develop on an ongoing basis, more effort is needed to improve existing detection methods. A variety of techniques with different features are used to categorize videos as real or fake using Machine Learning and Deep Learning techniques. Of the various technologies used, CNN with LSTM seems to give a better result. The future work is to use different face detectors for detecting face regions from video.

References

[1] Ismail, Aya, Marwa Elpeltagy, Mervat S Zaki, and Kamal Eldahshan. "A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost." Sensors 21, no. 16 (2021): 5413.

[2] Ismail, Aya, Marwa Elpeltagy, Mervat Zaki, and Kamal A. ElDahshan. "Deepfake video detection: YOLO-Face convolution recurrent approach." PeerJ Computer Science 7 (2021): e730.

[3] Mitra, Alakananda, Saraju P. Mohanty, Peter Corcoran, and Elias Kougianos. "A novel machine learning based method for deepfake video detection in social media." In 2020 IEEE International Symposium on Smart Electronic Systems (iSES)(Formerly iNiS), pp. 91-96. IEEE, 2020.

[4] Güera, David, and Edward J. Delp. "Deepfake video detection using recurrent neural networks." In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), pp. 1-6. IEEE, 2018.

[5] Yadav, Digvijay, and Sakina Salmani. "Deepfake: A survey on facial forgery technique using generative adversarial network." In 2019 International conference on intelligent computing and control systems (ICCS), pp. 852-857. IEEE, 2019.

[6] Amerini, Irene, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. "Deepfake video detection through optical flow based cnn." In Proceedings of the IEEE/CVF international conference on computer vision workshops, pp. 0-0. 2019.

[7] Nguyen, Thanh Thi, Quoc Viet Hung Nguyen, Cuong M. Nguyen, Dung Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. "Deep learning for deepfakes creation and detection: A survey." arXiv preprint arXiv:1909.11573 (2019).

[8] Sabir, Ekraam, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. "Recurrent convolutional strategies for face manipulation detection in videos." Interfaces (GUI) 3, no. 1 (2019): 80-87.

[9] Hopfield, John J. "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the national academy of sciences 79, no. 8 (1982): 2554-2558.

[10] Kwok, Andrei OJ, and Sharon GM Koh. "Deepfake: a social construction of technology perspective." Current Issues in Tourism 24, no. 13 (2021): 1798-1802.

[11] Wodajo, Deressa, and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer." arXiv preprint arXiv:2102.11126 (2021).

[12] Ranjan, Pranjal, Sarvesh Patil, and Faruk Kazi. "Improved generalizability of deep-fakes detection using transfer learning based CNN framework." In 2020 3rd international conference on information and computer technologies (ICICT), pp. 86-90. IEEE, 2020.

[13] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11. 2019.

Author's biography

D. Myvizhi received the B.Tech degree in Information Technology from Kongu Engineering College, Perundurai, Erode. She is currently pursuing an M.E degree in Computer Science and Engineering at Government College of Technology, Coimbatore. Her research interests include Machine learning, Image processing, and Deep Learning.

J. C. Miraclin Joyce Pamila is a Professor & Head of the Department of Computer Science and Engineering, Government College of Technology, Coimbatore. She teaches and guides students at both undergraduate and postgraduate levels. She is a life member of ISTE and currently takes up research in the area of Machine Learning, Data Analytics, and Natural Language Processing.