Proceedings of the International Conference on Automation, Computing and Renewable Systems (ICACRS 2022)
IEEE Xplore Part Number: CFP22CB5-ART: ISBN: 978-1-6654-6084-2

DeepFake Detection Through Key Video Frame Extraction using GAN

Lalitha S, Kavitha Sooda

Department of Computer Science & Engineering, B.M.S. College of Engineering, Bangalore

lalitha.scs20@bmsce.ac.in, kavithas.cse@bmsce.ac.in

2022 International Conference on Automation, Computing and Renewable Systems (ICACRS) | 978-1-6654-6084-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACRS55517.2022.10029095

***Abstract*— The emergence of deepfake videos in recent years has made image falsification a real threat. A deepfake video uses deep learning technology to substitute a person's face, emotion, or speech with the face, emotion, or speech of another person. Finding such deceptive deepfake videos on social media is the first step in preventing them. A robust neural network-based technique to identify false videos is presented in this paper. An important video frame extraction approach is used to speed up the process of finding deep fake videos. A model made up of a convolutional neural network (CNN) and a classifier network consisting of GAN technology is provided. Resnet, Resnext50 and LSTM were passed over in favor of the Confusion Matrix when deciding which structure to pair with the classifier while detecting the fake video. The model is a method for detecting visual artefacts. The subsequent classifier network uses the feature vectors from the CNN module as this is the input to categorize the video whether it is fake or real one. The dataset is considered from DeepFake Detection Challenge to get the best model. The key goal is to get high accuracy without using a lot of data to train the model. In comparison to earlier efforts, the key video frame extraction method dramatically decreases computations by achieving 97.2% accuracy using the Deepfake Detection Challenge dataset.**

Keywords— Convolutional Neural Network, Deepfake, LSTM, Resnet, Resnext50.

I. INTRODUCTION

A technology that may overlay facial images of a target person over videos of a source person to create videos of the target person acting or saying what the source person says or express is called a "deepfake," which comes from the words "deep learning" and creating a "fake" video. Deepfake algorithms can produce fake photos and videos that are so convincing that no one can distinguish them from the real ones. A rising sense of anxiety developed due to deep fakes, which allows the creation of proof for the scenarios that have never occurred, has started to emerge. Politicians and celebrities are the ones who are most impacted by this. Anyone can be seamlessly integrated into a photograph or video that they have no prior knowledge of using Deepfake. The speed at which systems can synthesis images and videos have increased recently as a result of widespread technological advancement. Deep learning has been successfully applied to tackle a variety of challenging problems, including big data analytics, computer vision, and human-level control. Deep learning advances have, however, also been exploited to create software that endangers national security, democracy, and individual privacy. One of the most recent deep learning-based technology based tool used to create deepfake video are Faceswap, Faceit, DeepFaceLab and DeepFakeCapsuleGAN. Deepfakes are the forms of entertainment produced by artificial intelligence that can also be lip-synced or puppeteered, according to a more general definition an Artificial Intelligence is used to detect another Artificial Intelligent with the help of GAN consisting of generator and a discriminator. Figure 1 displays a selection of phoney photos created by Style-GAN [1] that have a realistic appearance. Videos that have been altered so that the mouth movements match the soundtrack are known as lip-sync deepfakes [2]. Puppet-master deepfakes are videos of a target subject (puppet) animated to mimic the facial expressions, eye, and head movements of a separate subject (master) sat in front of the camera.

FIGURE 1. Style-GAN [1] images examples.

Our face represents who we are. People tend to remember people by their faces. Face modification, therefore becomes the technique that is most commonly used when an image and video forgery are involved. Face forgery in multimedia has dramatically expanded during the past 20 years. Among the

known works, Breglera [3] produced a video in 1997 using an image-based method. Garrido[4] suggested the idea of changing the actor's face without altering the actor's expression. It's particularly significant that theis[5] the real-time expression transfer work from 2015, which improved the lip syncing, to make people aware of how severe video counterfeiting is.

With deepfake technology, reality can be altered beyond what is tolerable. The truth is impacted by this disruptive technological shift. Some are meant to be humorous while others are not. They might pose a risk to democracy, the national security, and the identity of an individual [6, 7]. A deepfake video can compromise someone's privacy and damage them [8, 9]. Individuals start to have less faith in the news and the photos and videos the media presents to them. Political turmoil or violence may be resulted from this as well. In essence, DeepFake datasets are used to train the model, and its performance is then evaluated through trials. In this paper DeepFake detection methods for images and videos is applied. This paper will also go over datasets used to detect DeepFakes as well as the DeepFake production techniques. [10-13] focuses on the development and detection of DeepFake in images, audio, and videos.

The following are the key objectives:

- Describing the DeepFake model with the help of ResNext50 and LSTM methods.
- Training the model with the help of training dataset collected from the sources and validating the model
- Detection of the Deepfakes present in the loaded video with the help of the constructed model.

The remaining of the sections are organized as follows: Section 2 provides detail description of the related work. Section 3 contain the description of the sources of the datasets required to train the model and further would contain the high end model to describe the actual workflow of the Deepfake detection process followed by the results of accuracy obtained with the conclusion.

II. RELATED WORKS

This section tries to describe and evaluate the various deepfake detection methods that have been employed. There have been numerous attempts to identify deepfakes, which are deep learning-based in their foundation. These methods are effective at identifying video flaws or specific frames within the video.

There is another parallel sort of work that is common in addition to the visual artifact-based works. These are based on a video's temporal characteristics. It has been investigated to use a CNN and LSTM architecture integrated network. The input image sequence is used by the CNN module to extract features. Here, the feature vector is used by the long-short term memory to create a sequence detector. A completely connected layer then determines if the video is fake or real. In a different paper, a mixed network was utilized to categorize fake videos. Recurrent neural network (RNN) technology with a DenseNet structure have been integrated. Each video has a smart contract connection and a hierarchical relationship with the videos under it. If the original smart contract can be found, the footage is said to be pristine. The Inter Planetary File System (IPFS) peer-to-peer network stores data using unique hashes. This model says it can be expanded to include music or visuals. The ownership of a video has been established by identifying phoney video detection has also been done using spatiotemporal characteristics footage.

Other methods have successfully improved accuracy by spotting temporal discrepancies using CNN in combination with learning models like Recurrent Neural Network (RNN), Long Short-Term Memory Networks (LSTM), and Capsule Network[7] and have produced promising results on datasets containing videos created by FaceSwap and deepfake. Although there have been a number of advancements, more durable models are still required to recognize deepfakes of lesser quality, and sustaining significant accuracy in the face of ever-evolving deepfake generation techniques which continues to be difficult.

The purpose of this work is to discuss effective feature extraction and processing necessary to find discrepancies in the deepfake videos.

III. DATA ACQUISITION

Any machine learning or deep learning-based solution starts with data collecting. The deep learning-based deepfake detectors have been tested with various data set, and both have demonstrated outstanding performance on well-known benchmarks.

- **FaceForensics++(FF++)** designated by FF++ consists of 400 edited videos produced by four distinct deepfake generation algorithms, together with 100 original videos. Additionally, utilizing the H.264 codec and the C23 and C40 encodings, raw video contents are compressed using two quality settings. The information comes from 97 YouTube videos, all of which include a trackable, primarily frontal face without occlusions that allows automated tampering techniques to produce convincing forgeries. The data can be utilized for segmentation and classification of images and videos od binary mask supply. 1000 Deepfakes models are also offered in order to create and improve new data.
- **Celeb-DF(v2)** is another high-quality dataset that includes 5639 false videos and 590 authentic celebrity videos. We have chosen the test data from here consisting of 50% fake videos and 50% real videos, and the training and validation sets were divided into 80 and 20 percent of each, respectively.
- **DeepFake Detection Challenge (DFDC)** For the purpose of assisting in the training of detection algorithms, Kaggle offers a fairly huge face swap video dataset. The DFDC dataset is by far the largest face swap video dataset that is currently and publicly available. It contains over 100,000 total clips acquired from 3,426 paid actors and was developed using a variety of Deepfake, GAN-based, and non-learned methodologies. A Deepfake detection algorithm that has only been trained on the DFDC can generalize to real "in-the-wild"

Deepfake videos when analyzing potentially Deepfaked video. This the main dataset used in training the model in this paper.

IV. IMPLEMENTATION METHODOLOGY

The work in this paper initially requires the Graphic Processing Unit(GPU), as rendering graphics is the main function of a graphics card. The computer cannot show any information or any video or image related data without a graphics card. To put it simply, the GPU uses information from the CPU and software to decide where to arrange pixels on the screen. This paper makes use of Nvidia GeForce RTX 30 Series because of it ray-traced graphics and cutting-edge AI capabilities with new RT Cores, Tensor Cores, and streaming multiprocessors. Cuda 11.3 version dramatically speeds up computing applications by harnessing the power of GPUs. General-purpose computing on GPUs is a method that uses CUDA, a parallel computing platform and application programming interface, to enable software to leverage specific kinds of graphics processing units for general-purpose processing. The python library called pytorch is used which is the benchmark, shows that the performance of PyTorch is the best, which can be attributed to the fact that these tools offload most of the computation to the same version of the cuDNN and cuBLAS libraries. Developed primarily by Meta AI, PyTorch is an open source machine learning framework that is based on the Torch library and is used for tasks like computer vision and natural language processing. The training, validation, and testing subcomponents make up the three parts of the Deepfake detection process. The main element of the suggested model is the training component. It is the location for model helped in learning. Designing and creating more accurate DL models, to suit a certain issue domain into the model takes a lot of work.

This paper makes use of Convolution Neural Network(CNN), where the model utilises the CNN to extract and integrate a video's properties, as well as to detect video falsity. Generative Adversarial Network(GAN), which is to detect the Deepfake video. The GAN consists of two main components such as generator and discriminator. Initial training of the generator and discriminator networks takes into account fewer layers, and as a result, the generator produces false images from the video, that are extracted from the video frames cropped with the help of ResNeXt 50 with a resolution of 4x4. This work focuses on training the GAN along with CNN and extracting the discriminator as a specialised module to detect Deepfakes. Examining how the performance of the discriminator varies with various setups and training techniques, testing a variety of discriminator architectures using a variety of datasets. Finally, a model using ensemble approaches was proposed to increase the effectiveness of a set of GAN discriminators. The FF++ and DFDC dataset is appropriate for the research because it consists of two separate compressed level videos. The level of video compression in three distinct ways for DFDC are altered.

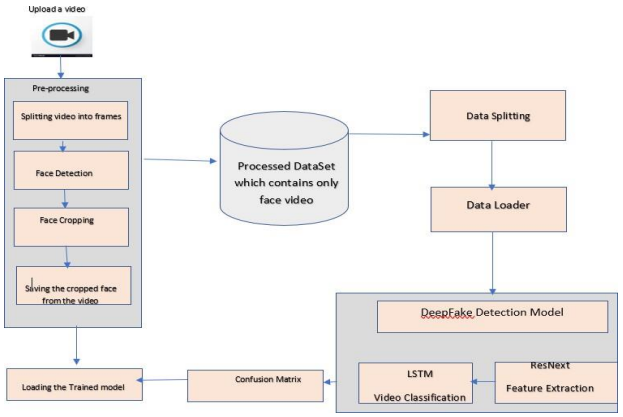

FIGURE 2. Proposed Model with detailed representation.

Frame extraction is only done in compressed videos because social media uses compressed data. Decompression has not been performed. The complete framework of the proposed method is shown in Fig. 2. The proposed model makes use of the resnext50 as it is one of the best network which repeats the building block that aggregates a set of transformations with the same topology and also it has a power of parallel stacking of layer rather than sequential that would happen in resnet. Resnext follows split transform merge technology. Resnext50 used in the proposed model has 50 layers, in which each layer has 32 nodes having the 4 dimensional model, which is capable of learning 25.0 × 106 number of parameters as shown in the Figure 3. The output of the resnext model after the pooling layer which is the feature extraction that is done here. The feature vector is fed into the sequential layer, further it is then passed to the LSTM model having the GAN technology incorporated to classify and determine the real and fake video. The LSTM layer has 2048 latent dimensions, 2048 hidden layers, and 0.4 dropout probability. The LSTM's additional output is then processed at the linear layer, adaptive average pooling layer, and softmax, which ultimately generates as real or fake. With the aid of the confusion matrix, which is mentioned in Table 1, the accuracy of the result is obtained.

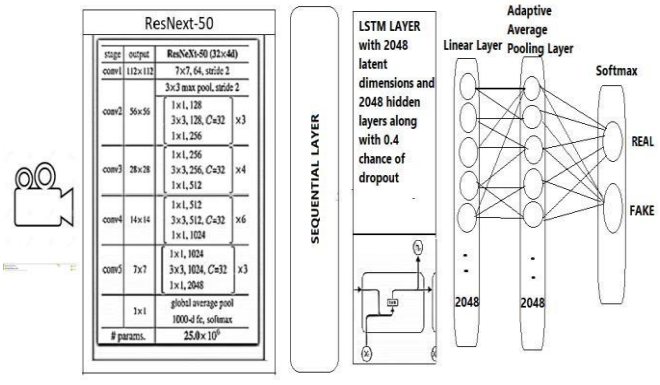

FIGURE 3. Resnext50 & LSTM Pre-trained model.

According to previous knowledge, single-dimensional scaling is not as difficult as it could be; yet, in order to identify an acceptable scaling scale to improve network efficiency, it is also required to take the resource occupation of the network and its learning rate into account. A formula will be used to express the three-dimensional comprehensive scaling problem in the EfficientNet. This refer to the full convolutional network as the function mapping of its convolutional layer is as follows:

$$Q_i = F_i(X_i) \quad (1)$$

With Qi being the output tensor and F_i being the convolution layers, Xi being the input tensor. The Batch dimension in the tensor is left out for ease of presentation, allowing the whole convolutional network Q, which consists of k convolutional layers. A formula will be used to express the three-dimensional comprehensive scaling problem in the ResNext. The complete convolutional network, which is referred to as Y, can be thought of as having the following function mapping in its n convolutional layers:

$$Y = F_k \odot \cdots \odot F_2 \odot F_1 X_1 \quad (2)$$

The work is based on binary classification, and we create the confusion matrix as shown in Table 1 to visualise the model's performance. Accuracy, as stated in Eq. (2), is the first performance indicator we can extrapolate from the confusion matrix. Accuracy is equal to T P plus T N plus F P plus F N. A table called a confusion matrix is used to describe how well a classification system performs. The output of a classification algorithm is shown and summarised in a confusion matrix. In Table 1, a matrix of confusion is displayed.

TABLE 1. Definition of TP, TN, FP and FN- Confusion Matrix.

True Positive (TP) Reality : FAKE (1) Model Predicted : FAKE (1)	False Negative (FN) Reality : FAKE (1) Model Predicted : REAL (0)
False Positive (FP) Reality : REAL (0) Model Predicted : FAKE (1)	True Negative (TN) Reality : REAL (0) Model Predicted : REAL (0)

Average Pooling Layer: Through the process of down sampling, this layer is in charge of lowering the dimensions of the data. For illustration, suppose a 2x2 average pooling layer is used with a 4x4 input matrix. The average of those values is essentially taken for each 2x2 block in the input matrix, which minimizes the dimensions. Network parameters are trained to their ideal value when the loss function L is optimized after multiple epochs. The learnable parameters, weights, and biases of the generator are adjusted after each iteration in accordance with the discriminator's recommendations. By backpropagating through the gradients of the discriminator's output in relation to the created image, the network changes the learnable parameters. The discriminator essentially instructs the generator on how to adjust each pixel to make the image appear more realistic. Consider that the discriminator believes an image produced by the generator has a 0.29 (29%) likelihood of being a real image. The generator's task is to modify its learnable parameters so that, for example, after computing backpropagation, the probability rises to 30%. Finally the video is detected as fake or real and accuracy of the result is calculated with the help of Confusion Matrix. In this paper accuracy of 97.2% is obtained for the detection of the DeepFake videos.

V. RESULTS

In the initial experiment, 200 videos of FF++ unseen data were tested on the model, and the ResNext50 feature extractor was used since it crops out the face level features from the retrieved frames.

Once the features are cropped and validated against the train and test data, the LSTM model classifies the videos uploaded for experiment are either fake or real. 97.25% accuracy as shown in Table 2. for epoch 13, for the videos with compression level c = 23 was achieved, the accuracy keeps increasing as the number of epochs increases and becomes constant at some point where all the features are extracted from the video frames and compared with the test and train model.

TABLE 2. Accuracy results for 200 video dataset

Epoch	Loss	Accuracy
1	0.564482	87.569488
2	0.551042	89.125489
3	0.524949	90.214587
4	0.491312	90.894215
5	0.474652	91.356895
6	0.440665	91.998756
7	0.431125	92.896235
8	0.421456	93.125865
9	0.412364	94.992561
10	0.402536	95.255658
11	0.395626	96.123595
12	0.385695	97.221312
13	0.381456	97.251252

Figure. 5 shows the result of the confusion matrix that was obtained for the dataset of 200 videos.

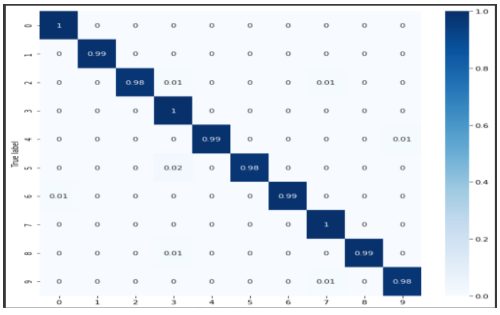

Figure 5. Results of Confusion Matrix

The graphs are plotted for Accuracy Figure 6 and Loss Figure 7 against the number epochs carried out for the total dataset uploaded in the model.

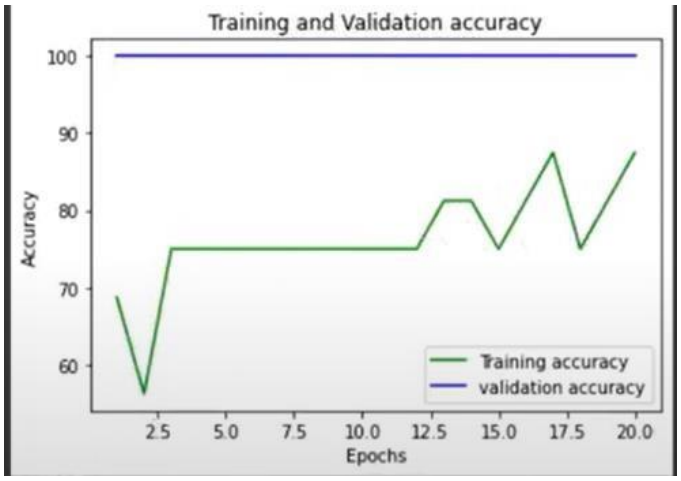

Figure 6. Results of graphs plotted for accuracy of the model

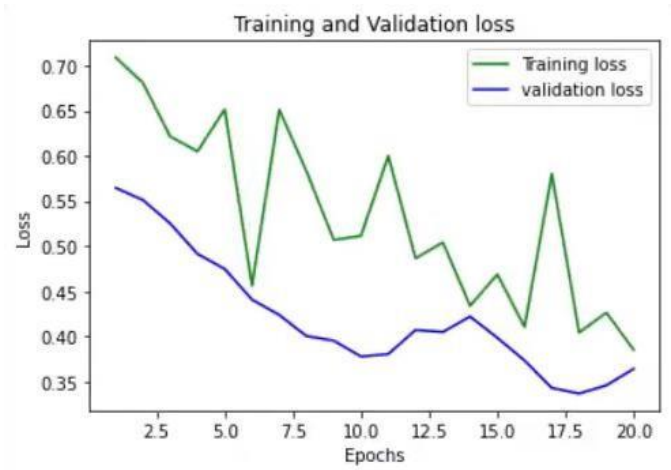

Figure 7. Results of graphs plotted for Loss against each epoch

Because the model was trained using only a portion of the DFDC dataset, its accuracy decreased when merged with the FF++ dataset. Given the size of the DFDC dataset, it is hypothesized that training the model with the whole dataset would have improved accuracy.

VI. CONCLUSION AND FUTURE WORK

The evaluation of the FF++ and DFDC datasets was successful. It is expected that the proposed method can be implemented at an edge device with the necessary adjustments because it considerably minimizes the computations. The field of embedded deep learning is expanding. Today's cloud-dependent deep learning area is being driven by significant demand for numerous application domains. The computation is significantly decreased by the method's ability to identify bogus videos by identifying crucial video frames. Deploying a video detection model at an edge device is thus a positive step.

However, the structure of deep neural networks is too vast to fit on edge devices due to memory constraints. So, a major endeavour for future work would be to reduce run-time memory and model size.

RFERENCES

[1] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4401-4410. 2019.
[2] Prajwal, K. R., Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. "A lip sync expert is all you need for speech to lip generation in the wild." In Proceedings of the 28th ACM International Conference on Multimedia, pp. 484-492. 2020.
[3] Bregler, Christoph, Michele Covell, and Malcolm Slaney. "Video rewrite: Driving visual speech with audio." In Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pp. 353-360. 1997.
[4] Garrido, Pablo, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. "Automatic face reenactment." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4217-4224. 2014.
[5] Thies, Justus, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. "Real-time expression transfer for facial reenactment." ACM Trans. Graph. 34, no. 6 (2015): 183-1.
[6] Chesney, Bobby, and Danielle Citron. "Deep fakes: A looming challenge for privacy, democracy, and national security." Calif. L. Rev. 107 (2019): 1753.
[7] Korshunov, Pavel, and Sébastien Marcel. "Vulnerability assessment and detection of deepfake videos." In 2019 International Conference on Biometrics (ICB), pp. 1-6. IEEE, 2019.
[8] Gerstner, Erik. "Face/off:" DeepFake" face swaps and privacy laws." Def. Counsel J. 87 (2020).
[9] Reid, Shannon. "The deepfake dilemma: Reconciling privacy and first amendment protections." U. Pa. J. Const. L. 23 (2021): 209.
[10] D'Amiano, L., Cozzolino, D., Poggi, G., Verdoliva, L.: A PatchMatch-based dense-field algorithm for video copymove detection and localization. IEEE Transactions on Circuits and Systems for Video Technology 29(3), 669–682 (2019).
[11] Ding, X., Yang, G., Li, R., Zhang, L., Li, Y., Sun, X.: Identification of motion-compensated frame rate up-conversion based on residual signals. IEEE Transactions on Circuits and Systems for Video Technology 28(7), 1497–1512 (2018).
[12] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge dataset. arXiv 2006.07397 (2020)
[13] Ganiyusufoglu, I., Ngo, L.M., Savov, N., Karaoglu, S., Gevers, T.: Spatio-temporal features for generalized detection of deepfake videos (2020).