International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Deepfake Video Forensics based on Transfer Learning

Rahul U, Ragul M, Raja Vignesh K, Tejeswinee K

Abstract: ***Deeplearning has been used to solve complex problems in various domains. As it advances, it also creates applications which become a major threat to our privacy, security and even to our Democracy. Such an application which is being developed recently is the "Deepfake". Deepfake models can create fake images and videos that humans cannot differentiate them from the genuine ones. Therefore, the counter application to automatically detect and analyze the digital visual media is necessary in today world. This paper details retraining the image classification models to apprehend the features from each deepfake video frames. After feeding different sets of deepfake clips of video fringes through a pretrained layer of bottleneck in the neural network is made for every video frame, already stated layer contains condense data for all images and exposes artificial manipulations in Deepfake videos. When checking Deepfake videos, this technique received more than 87 per cent accuracy. This technique has been tested on the Face Forensics dataset and obtained good accuracy in detection.***

Keywords: Deepfake Creation, Detection, MobileNet, Transfer Learning.

I. INTRODUCTION

Abundance of video clips is posted every day over video streaming web sites like YouTube and other social media. It is important to differentiate between genuine information and forgeries in this age of information explosion. In late years, Deepfake video clips have started popping up in social media. Found in the beginning, the different clasps were coursed in which the substance of famous on-screen characters like Keanu Reeves showed up however he didn't take part in it. Following, a user of Reddit published some clips with the face of different porn actresses traded with an acclaimed actress' face. Although these fake materials were quickly discovered and removed, the application used to build the forgery clips was still accessible to the general public. The latest Deepfake based application was also growing with increasing popularity.

Manuscript received on March 15, 2020.
Revised Manuscript received on March 24, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Rahul U*, Student, Department of Computer Science and Engineering, Rajalakshmi Engineering College (Anna University), Tamil Nadu, India.
Ragul M, Student, Department of Computer Science and Engineering, Rajalakshmi Engineering College (Anna University), Tamil Nadu, India.
Raja Vignesh K, Student, Department of Computer Science and Engineering, Rajalakshmi Engineering College (Anna University), Tamil Nadu, India.
Tejeswinee K, Department of Computer Science and Engineering, Rajalakshmi Engineering College (Anna University), Tamil Nadu, India.

Today anyone with reasonable computer skills and a good machine might be able to develop a Deepfake clip that looks like a film's special effect team makes. In early 2020 a video clip posted by Ladbible [1] shows a deepfake recreation of 'Back To The Future' With Tom Holland And Robert Downey Jr.

In fact, it engaged the same technology and altered actor Michael J Fox's face to be Tom Holland and Christopher Lloyd's face to Robert Downey Jr. Prior to these videos, thousands of video clips are created with enormous number of celebrities' facial images but without their approval. These Deepfake clips have captured a lot of public figures in the midst of bewilderment and dread. In the aforementioned paper, we characterize an advanced transfer learning established path that can catch the Deepfake video clips generated by either Deepfake applications or facial re-enactment.

The technique described in this paper is established on the common attributes of fake video clips that analyse face interpretation. Trivial clips, particularly Deepfake clips can only develop finite duration of face frames from a single clip of video material and as a result of the data and time, training phase is enormously time consuming. In order to develop good clips, the training dataset should be adjusted to the different angles and light conditions of the training data set. With inadequate training data, the training clips can leave even some face models to collapse at particular frames with distinct objects. So the aim of this paper is to catch these objects and decide if the video is being fabricated.

II. RELATED WORK

Re-enactment of synthesized facials has gained reputation in late years. Facebook has used a similar technology in their "Facebook Camera" to create emojis dynamically based on facial features. Following Face2Face's suggestion for an approach to recording real time facial shifts from one clip of video and re-enacting it on various RGB videos, any sound to lip join technique suggested by Suwajanakorn et al [2] has built a Recurrent Neural Network (RNN) capable of matching raw audio characteristics to certain mouths. Generative Adversarial Networks (GANs) on the generative models are very promising. A NVDIA researcher has advanced a style transfer system for producing photo realistic images of the human face. Akin conclusion can be obtained by modifying the encoded image attribute, an approach suggested by Lample et al [3], which suggests altering encoded image values in order to build new images. Lu et al [4] also proposed an attribute-guided face generator to spawn huge resolution facial image from lower resolution images. Shubham Sharma suggested a celebrity face generator [5], a celebrity face dataset is fed into this GAN-based generator and trained.

It is named CelebA and the generator is capable of generating lower resolution face images after sixth training epoch. A morphing violation occurs by merging the facial features of two or more people to create a different synthesized face that will be blurred onto the authentic face.

This type of attack commonly gives gigantic goals and basic appearance in face demeanor. Clemens et al [6] suggested a Convolution Neural Network (CNN) approach tested GoogleNet, AlexNet and VGG19 achieving between 17% and 11% False Exclusion Rate (FER) and 21% Fake Acknowledgement Rate (FAR) at 17%. Guera et al [7] used a Convolution Neural Network (CNN) for the extraction of features and coupled it with a convolutionary sequence processing LSTM network. They checked video clips from Deepfake and videos total of 300 taken from HOHA dataset. The aforementioned system on measuring accuracy has hit 97 per cent.

III. METHODOLOGY

A. Introduction to Common Flaws and The DeepFake

For face swapping, Deepfake extracts images of face from both video source and target video material. Autoencoder is used by Deepfake to obtain appropriate face features from the source data. After enough training photos have been collected, the encoder codes these to a feature map with help of face features and combine the facial attribute of source with the facial attribute of the target for a "facial mask". The target face is then replaced by this mask by different transformations in each frame (i.e color transfer and the blurring erosion). The output of the synthesized video is then decreased. The latest implementation techniques associated with Deepfake is DeepFaceLab. This source library is available freely which also introduce its predecessors with several new features: multiple training models, a preview of progress training and a model of CPU training. DeepFaceLab is now one of the popularly used tools in the Deepfake culture.

To obtain face from videos, DeepFaceLab uses face detection by Multi-task Cascade Convolutional Neural Network (MTCNN). During extraction, MTCNN gives more false positives than DLIBCNN [8]. The DLIBCNN produces low number of scrambled aligned face when video frames become unstable. The manual extractor provided by DeepFaceLab captures missed faces that enables the user to obtain missed faces from a particular frame, enabling the complete manual extraction from the source film for the best result. Deepfake videos typically have short, low resolution and short lengths. We can therefore logically come to the conclusion that if more blurred images, particularly around the facial area, are presented in a short video, indicates a high risk of forgery video. Nonetheless, we will review the production of Deepfake videos for research purposes in order to comprehend the explanation for these shortcomings.

The three key factors when making Deepfake video clips: (1) the final character of the video; For example, the result of the falsification needs to be of relatively high resolution and less artifacts and less natural face expression that corresponds to the context of the video; (2) the velocity of video training. The training pace is primarily affected by VRAM size, deep net structure of the neural network and the training data size; (3) the video duration making. More time and data are needed to prepare longer video clips. Understanding these aspects will guide us to comprehend the Deepfake video creation phase better and to find points of weakness. So many deepfake clips of video developed by the Deepfake apps have certain defects, like artifacts are visible, distinct resolution of the face and a sudden change of color when the model is moving quickly. Such errors are triggered by the learning algorithm of Deepfake. If there is inadequate training dataset to cover angles or facial lighting by contrast most Deepfake application users don't have enough resources such as computing or time to develop a well-trained model. Hence creating an impeccable Deepfake clip of video is very challenging. A "predicted mask" will be generated throughout DeepFaceLab, to suit the true face region to be concealed, but this process will also produce noticeable artifacts on the mask's corners. DeepFaceLab uses the following three methods to remove artifacts: (1) Use Gaussian filters to cover the boundary zone. (2) expand the region which is masked. (eg enclose jaws and the font side of head) and (3) physically adjust the mask shape. Applying mask directly could create a fully covered face mask for the destination face; this procedure could also add double chain and blurred edges. Compared with other synthesized videos it causes most Deepfake videos, particularly around the face, to be generated with relatively low resolution. In addition to the resolution, skin tone and facially reflected ambient lighting are important factors for Deepfake video detection. Choosing the destination source clips of video is central to creating a foolproof mask for the destination face model. This is because the head movement of the source model will impact directly the size of the training data. Hence, having fewer head shifts while preserving consistent ambient lighting is important for a perfect deepfake. As an observation this paper comes up with using transfer learning to notice graphical deviation within each frame on some existing Convolutional Neural Networks CNN.

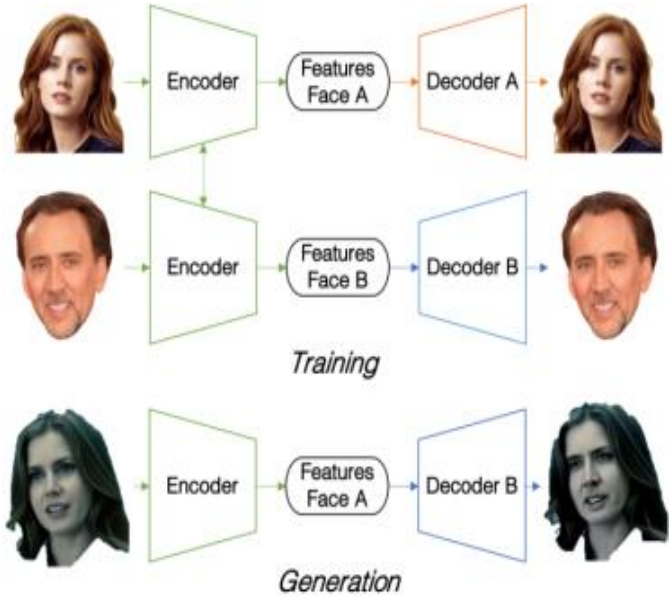

Fig. 1 what makes deep fakes conceivable is figuring out how to compel both inert appearances to be encoded on similar highlights. This is illuminated by having two systems having the equivalent encoder, yet utilizing two distinct decoders (top). At the point when we need to do another faceswap, we encode the info confront and disentangle it utilizing the objective face decoder (below).

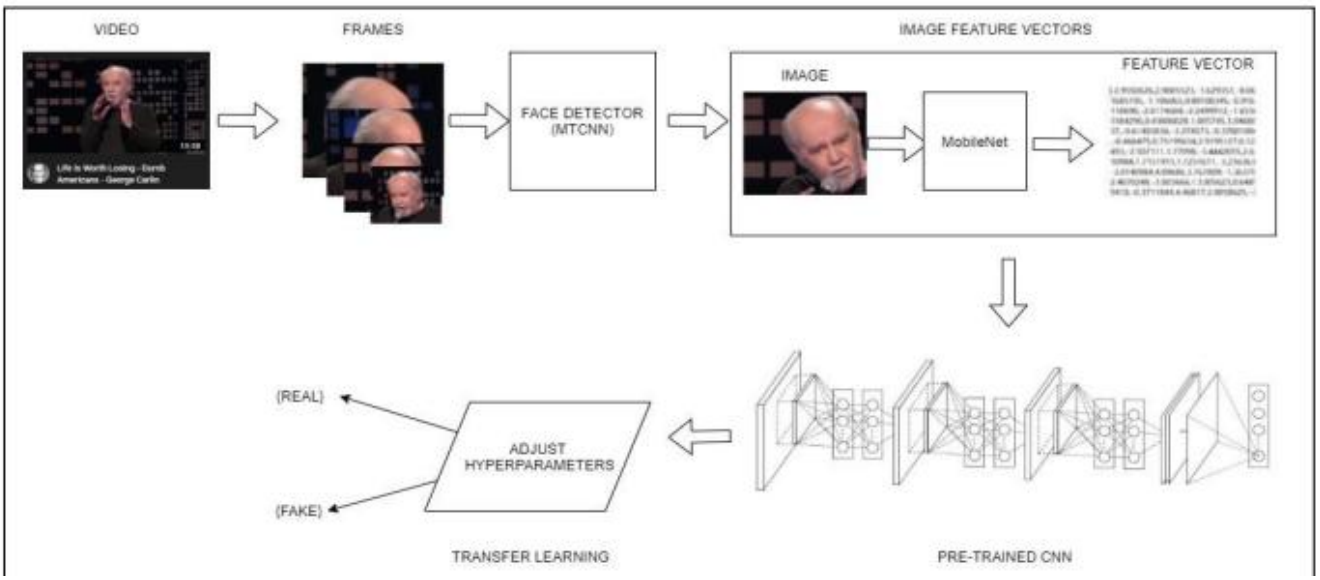

Fig. 2 the Deepfake videos are converted to frames and fed into face detector (MTCNN) to obtain the image feature vectors by using MobileNet. These Feature vectors are given as input to the pretrained MobileNet and then perform Transfer Learning in Pretrained MobileNet Neural Network to classify the Videos as Fake or Real

B. Transfer Learning

Training performance is influenced by two factors in the implementation of transfer learning on the existing model:(1) The original bias and weight of the pre-trained neural network. (2) New dataset to retrain the neural network. Focus therefore on which method should assess the effectiveness of the model which is re-trained and its accuracy in classification [9]. The vigorous re-training of the entire neural network always will not contribute to the most successful network. On the one side, this method is ideal if the retraining data exceeds the pretrained data, or if the data which is used for re-training has little connection with pre-trained data. With this setting, the hyperparameters of previously trained models are just an "initialiser". In order to achieve higher precision, the training algorithm has to be hyper-tuned intensively and the overall training time and the chance to overfit greatly increases.

On the other hand, the network only focuses on the creation of a classification layer using training parameters from a pretrained model: Fully Connected nodes and SoftMax activation. Although the exactness of this procedure is limited by the pre-trained network parameter setting, the average time for training can be significantly reduced. MobileNet is the cutting-edge neural network which has been trained on the ImageNet dataset. ImageNet contains over 15 million images from various objects, as well as artifacts inside the frame with the ImageNet data image function vector.

C. **Creating Image Feature Vectors**

The goal of transfer learning is to replenish the final layer using new labels of classification. The bias and weight of these newly formed layers are determined by the background propagation results. The layer value of the bottleneck, also known as the "Image Feature Vector," is the layer of classification just previous to the final output layer [10]. The set of values obtained from the bottleneck layer could be used to differentiate the new label by classification layer. During training, each picture will be again used, so to prevent recalculations on the same images, the values in bottleneck will be saved in the files for future use. Every image will be given over the network during the process of retraining to extract the image feature vectors. Since the file containing bottleneck values that contain the feature maps of each pictures when building new layer which is fully connected and SoftMax, the use of files known as bottleneck is important.

IV. EXPERIMENTS

The neural network adopted for retraining is MobileNet. The image data pass through the pre-processing phase before starting the reconstruction procedure. The images contain two categories - real label picture and fake images. It is very difficult to distinguish the distinction between many of these pictures through the low resolution of naked eye. The aim of this analysis is to appeal the conclusiveness of the proposed method and evaluate the neural network's efficiency.

A. **Dataset**

The dataset includes three image sets from the FaceForensics++ facial re-enactment dataset [11]. Every set contains various face images. Such pictures are cropped to diverse sizes and quality from multiple videos. Set 1 is a small subset.

For example, this group includes 500 extracted face images from a one video media. The purpose of this group is to check the capability of the neural network when it has been trained on small image sets. In these three sets the images are segregated into ratio of 75/25. Images are placed on various files in a separate folder with a name ‘real’ and ‘fake’. The algorithm chooses these images at random and further splits them into sets of preparation evaluation and checking Set 2 has 2500 pictures and last Set C has 15000 images.

Each divided dataset consisting of obtained face set from video created by Deepfake, Deepface, DFake and authentic video will be used for retraining model. Each frame in the video will be removed by single frame per second as PNG file. In this step, frames containing many faces will be erased. The derived frame will then be fed into the face recognition processor for facial extraction. In some frames, DLIB's facial extractor will periodically grab non facial region. Hence, trimming the dataset is useful prior to training the algorithm. Dataset information is shown in Table 1.

TABLE I. **Dataset Information**

	Size	Cropped facial image size	Number of videos	Real to fake ratios
Set 1	500	506*506 to 420*420	1	1:1
Set 2	2500	320*320 to 430*430	5	1:1
Set 3	15000	160*160 to 750*750	30	1:1

B. Training Parameters

Training phase is managed by several parameters The speed of gradient descent is controlled by Learning rate. Setting More value will improve the direction of learning, at the expense of losing the best accuracy; however, making this too low value leads to the local minimum being sticked to or can prolong the period of training.Thus it is necessary for the preparation to choose the parameters carefully.

The training steps are set by the limit of the total training. Another important aspect of a model training is the appropriate value of the training step. Increasing simply the number of training steps improves training performance; however, the rate of improvement can gradually decline as training progresses and accuracy of the training can even decrease as a result of over-adaptation.

There are four distinct alpha values which could be taken for MobileNet V1 model from α={0.25 0.5 0.75 1} and from resolutions {128 160 192 224} to adjust network efficiency[20]. The proposed method sets resolution=128 and alpha= 0.5 for this experiment to maintain a balance between time for training and correctness. The size of MobileNet V1 model created image feature vector is 1001.

C. Retraining

The re-training process takes 2500 steps and five images at each stage from the training set are selected randomly. The bottleneck file of the images selected is then fed to the layer which classifies. In fact, the test results are transmitted back to the classified layer and change the bias of layers and weight, thus improving the prediction's precision. The FaceForensics Lab data set is split into three sizes: 500 images, 2500 images and 15,000 photos.

V. EXPERIMENT RESULTS

Table 2 gives an outline of the outcomes of our studies. The data set was split into 70% for training set and 15% for validation set and 15% for test set. During the experiment, the training steps at 2500 and learning rate at 0.005 was fixed to reflect the similarity.

VI. CONCLUSION

From this work, it is a suggested approach to detect the Deepfake generated videos. This approach relies on the restrictions that latest DeepFake applications can generate only short videos with low resolution, and the resolution along the facemask edge during conversion will further decrease by various settings. Deep neural networks, which are pre-trained, can catch these distinctive artifacts. The experimental test results show methods used in this paper is effective. All Deepface videos and Deepfake videos are effectively identified by the retrained model. This strategy has averaged 86.9 % accuracy and significantly reduced training times. The detection approach should also develop accordingly as the technology relating to the Generative Adversarial Network (GAN) continues to evolve. By carrying out rigorous research at the different levels, it is targeted to enhance the reliability of tests. Transfer learning is also best achieved by training in a pre-built, neural network on Deepfake detection, which includes specific training data for Deepfake classification.

	Training accuracy	Validation accuracy	Test accuracy	Training cross entropy	Training time (mins)	Test sample size
Mob_500	100%	97%	86.8%	0.0123	2:40	93
Mob_2500	100%	97%	84.8%	0.0346	6:15	525
Mob_15000	85%	91%	88.8%	0.5467	40:32	2484

TABLE 2. OVERALL PERFORMANCE AND PARAMETERS AT STEP =2500, LEARNING RATE=0.005

REFERENCES

1. https://www.ladbible.com/entertainment/tv-and-film-deepfake-recreate-back-to-the-future-with-robert-downey-jr-20200216?source=facebook.
2. S., Seitz , S., Kemelmacher-Shlizerman, I Suwajanakorn, : Synthesizing Obama: learning lip sync from audio. ACM Trans. Graph. 36(4), 1–13 (2017)

3. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: manipulating images by sliding attributes, arXiv:1706.00409 [cs] (2017)
4. Lu,Y., Tai, Y.-W., Tang, C.-K.: Attribute-guided face generation using conditional CycleGAN, arXiv:1705.09966 [cs, stat] (2017)
5. Shubham Sharma: Celebrity Face Generation using GANs (2018). https://medium.com/coinmonks/celebrity-face-generation-using-gans-tensorflow-implementation-eaa2001eef86
6. Eisert, P ., Seibold, C., Samek, W., Hilsmann, A.: Detection of face morphing attacks by deeplearning. In: Kraetzer, C., Shi, Y.-Q., Dittmann, J., Kim, H.J. (eds.) Digital Forensics and Watermarking, vol. 10431. Springer International Publishing, pp. 107–120 (2017)
7. Delp, E.J .,Guera, D.,: DeepFake video detection using recurrent neural networks. In:Proceedings 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, pp. 1–6 (2018)
8. Kazemi, V..: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition.(2014).https://doi.org/10.1109/cvpr.2014.241
9. Li, C., Balaban, S.: Transfer Learning with TensorFlow Tutorial: Image Classification (2018). https://lambdalabs.com/blog/transfer-learning-with-tensorflow-tutorial-imageclassification-example/
10. TensorFlow Hub. How to Retrain an Image Classifier for NewCategories(2019):#bottlenecks http://www.tensorflow.org/hub/tutorials/image_retraining
11. Rössler, A.,Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics: a large-scale video dataset for forgery detection in human faces, arXiv:1803.09179 [cs] (2018)
12. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications, arXiv:1704.04861 [cs] (2017)
13. Rahul U,Ragul M,Deepfake Forensics Using Recurrent Neural Networks", International Journal of Emerging Technologies and Innovative .http://www.jetir.org/papers/JETIR1908651.pdf

AUTHORS PROFILE

Rahul U is a student in the Department of Computer Science and Engineering at Rajalakshmi Engineering College(Anna University). His areas of specialization includes Machine Learning, Deep Learning, IoT . He is also a Pega CSSA.

Ragul M is a student in the Department of Computer Science and Engineering at Rajalakshmi Engineering College(Anna University). His areas of specialization includes Algorithm Analysis. He is also a Pega CSSA.

Raja Vignesh K is a student in the Department of Computer Science and Engineering at Rajalakshmi Engineering College(Anna University). His areas of specialization includes Data Analytics.

Tejeswinee K has completed her bachelor's degree in Anand Institute of Higher Technology and her master's at Sri Sivasubramaniya Nadar College of Engineering. Her areas of interest are Machine Learning, Deep Learning and Bioinformatics.