DeepFake Detection using InceptionResNetV2 and LSTM

Priti Yadav[1], Ishani Jaswal[2], Jaiprakash Maravi[3], Vibhash Choudhary[4] and Gargi Khanna[5]

[1,2,3,4]*Student, National Institute of Technology, Hamirpur,H.P, India*
[5]*Associate Professor, National Institute of Technology, Hamirpur,H.P, India*

Abstract

"Seeing is believing" is simply not true anymore and has huge ramifications for many different aspects of our life. As technology is improving, it's becoming easier and easier to develop deepfakes. In fact, some of it is even possible at the palm of hand with app. It's not easy to detect deepfakes. It has become difficult for the human eye to detect deepfakes. But meanwhile some researchers are working on finding ways to recognize deepfake. Deepfakes are the media which are synthesized using the algorithms of AI. The algorithms of AI are made to learn the attributes of the target image and the source image. The target image is then superimposed on the source image. We aim in detection of video deepfakes using deep learning neural networks like LSTM and InceptionResNetV2. We succeeded to build deepfake detection model by using transfer learning where the pretrained InceptionResNetV2 CNN is used to extract features and for vector formation. The LSTM layer has been trained using the features and the resultant confusion matrix provides us the validation and testing accuracy. The respective model achieved 84.75 percent and 91.48 percent accuracy for 20 and 40 epochs respectively.

Keywords

Deepfake, InceptionResNetV2, Deep learning, Neural Network, LSTM, Generative adversarial network.

1. Introduction

In the last few years, digital technology has advanced to the point where we can drastically alter how anything seems online. One can think of defects as Photoshop videos but there's a lot more to it. Those defects actually mean use of artificial intelligence to teach machines how to react, read and mimic people's facial expressions and voices. This is done by providing a machine with actual photos videos and voice samples of the person. The system learns from the data provided and is then able to generate a completely fictional video of that person. This is done in two ways the first is where the deep fake is created using another actor on the place of first person, the machine learns how to encode or process the data from both videos and find similarities between the two then compresses this data and through a decoding process it swaps the information of both the videos. So while you see and hear the voice of a person the information you're actually receiving is that of the actor. The other way is using a generative adversarial network (GAN) [1].

International Conference on Emerging Technologies: AI, IoT, and CPS for Science Technology Applications, September 06–07, 2021, NITTTR Chandigarh, India

pritinith@gmail.com (P. Yadav); ishanijaswal8@gmail.com (I. Jaswal); Jai516843@gmail.com (J. Maravi); vibhash18.vc@gmail.com (V. Choudhary); gargi@nith.ac.in (G. Khanna)

In this deep learning algorithms are used to create a synthetic image out of noise which are then added to stream of real images. When this is processed multiple times over and over again it gets realistic faces of non-existent people. Basically, through deep learning we now have the power to create convincing videos of people seeing and doing pretty much anything we want. Deepfakes have been categorized to three types: face synthetics, face swap and face expression manipulation. In face synthetics, the most popular approach is to produce face synthetics is StyleGAN. The generator model is trained in such that it separates the high level features from other features. One method to detect these face synthetics is extracting manipulated region of face. This system gives binary output whether the image is true or false. Face swap is another kind of deepfake which is obtained after swapping the face of the target with real person. It uses image blending, face alignment, cropping and other technologies to swap the face, generating faceswap deepfake. To detect face swap mostly CNN and RNN are trained to recognize traces that are leftover during their generation. Facial attributes and expressions manipulation implies some modification in the attributes like the gender, color of the face, color of the skin or hair, the age of the person or making the person sad or happy [2].

2. Literature survey

In the recent times there is an explosive growth in the number of deepfakes. In the present times there are many softwares that facilitate the creation of these deepfakes. They are becoming a threat to privacy, democracy and trust. So, there is an increase in the demand for deepfake analysis. We are listing some of the approaches for deepfake detection.

A. Jadhav et.al, [3] have developed a web based platform with which a user can easily classify it as real or fake by uploading the video. The model used ResNeXt and LSTM. The method is user friendly and reliable. The approach was based on feature extraction from frame level features using ResNext and sequential processing using LSTM. The methodology included dividing the videos into frames and then cropping the frames across face. Some of the selected frames were combined to form face in video thereby they created a new dataset containing face cropped of all videos. Then they have calculated quite good accuracy using confusion matrix which is used for model evaluation.

Y. Li, S. Lyu, [4] have proposed a new method of comparison between generated face areas and their surrounding regions using Conventional Neural Networks. The method was based on observation whether images of limited resources can be generated by the DF algorithm.

U. A. Ciftci, I. Demir and L. Yin, [5] have aimed on feature extraction and then computing coherence and temporal consistence. The method extracted biological signals from facial regions from the fake and real video pair. An SVN and a CNN have been trained to find probabilities of authenticity.

D. Guera and J. Delp, [6] have used recognition pipeline to automatically detect deepfakes. They have proposed a two step analysis. During first stage, features are extracted at frame level using CNN. The second stage consists of RNN which will capture erratic frames introduced due to face swapping process. The dataset contains 600 videos collected from various online sources was analysed. The accuracy achieved by their model is 94 percent.

Y. Li, MC. Chang and S. Lyu, [7] have intoduced a new system of exposing deepfakes based on the eyeblinking which are generated using neural networks. The paper focused on analysing the eyeblinking in the video as it is a natural signal and it cannot be presented well in the synthesized media. In the method the videos have been first preprocessed to locate face area in each frame, then a Long Term Recurrent Convolution Network(LRCN) find out temporal incongruity.

3. Proposed work

The basic architecture to produce deepfake is encoder - decoder architecture, where the encoder acquires the features of the target and the source face and the task of the decoder is to get encoding features of the target face and then generate fake video [8]. Using high level processing, the quality of the video is enhanced and the left overs are removed but still few traces are left which are not visible by naked eye. These leftover traces are the key features of our detection model [9]. The proposed model comprises of InceptionResnetV2 for feature extraction. These extracted features are used to train a recurrent neural network which is made to analyse if the video has been put through manipulation or not. Only a small portion of the video is manipulated which means the deepfakes are shorter in time, therefore, the video is split into small frames and these frames are given as an input to detection model [10].

3.1. Dataset and Preprocessing

The Dataset has been collected from deepfake detection challenge dataset available on Kaggle, FaceForenscis and Celeb-deepfakeforensics [11]. It contains around 6458 videos. These videos also include real videos which were further manipulated by paid actors and then created into deepfake video by using different deepfake generator methods. 70 percent dataset has been used for training and 30 percent for testing the system. We also fed the machine with labels of the video files fed to the system during training period. The point where the original video has been converted as a deepfake video is captured as a frame and then analyzed during preprocessing. An average of 147 frames are extracted during preprocessing of a video. Due to less computation power, we used limited number of frames for training the model. After preprocessing frames are send further for training and testing in small batches.

3.2. Modelling

Model for this system conducts an image categorization analysis on each frame extracted from the video. We used a pretrained CNN model named InceptinResNetV2 [12] and RNN along with LSTM. We also need to define Loss function, Optimizer and other Hyper-parameters required for the training procedure. Depending on the state of training model, the learning rate should be adjusted to minimize the loss value.

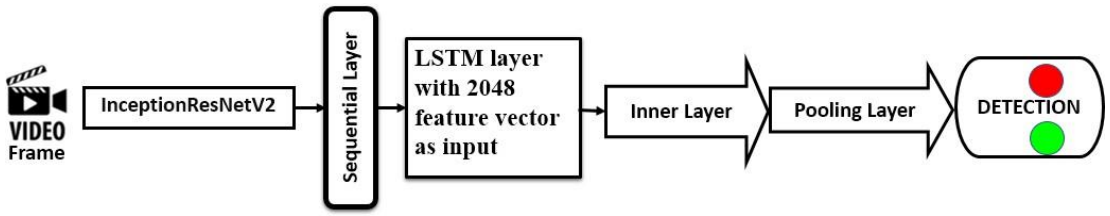

Figure 1: Model architecture.

3.3. InceptionResnetV2 for feature visualization and classification

InceptionResNetV2 is a combination of Inception and ResNet family having 164 layers for object detection and feature extraction from an image. Only the last layer is added to analyze the outcome. To train the InceptionResNetV2 CNN model, we simulate the resolution inconsistency in affine face wrappings directly during manipulation of the video. Using trained model helps to reduce size and training difficulty[13]. InceptionResNetV2extracts the features from each frame during the preprocessing 2048-time dimensional feature vectors are considered after last pooling and then LSTM is the next sequential layer. As CNN does not consider temporary discontinuity, it only looks into facial extractions and detection, we consider LSTM for sequence processing.

3.4. LSTM for sequential processing

Long Short Term Memory (LSTM) is a variety of Recurrent Neural Network (RNN) and it has feedforward connections. They are a special version of RNN that solves the issues of shorter memory. LSTM eradicated the vanishing gradient problem in RNN and they are designed in such a way that they learn long term dependencies of data and process the data sequentially. The output of CNN network acts as an input to 2048 LSTM layer [6]. LSTM processes the frames sequentially and then compare the features of the frame at different time [3]. By comparing the frame, it depicts whether the video is deepfake or not. After training, any video can be passed to the model for prediction.

4. Implementation

Python is the best known language for machine learning application, that's why we have used python to load the dataset and for face extraction. The dataset contains video files and these files are labelled as fake or real videos in a different Cvv file. After this the code matches the dataset with labelling file and find out if there is any missing file. After confirming the exact number of unique videos. Images are extracted from video and stored in the form of frames. At this time OpenCV is used for image recognition and interpretation. The captured frames are sent to the model for pre-processing. After the pre-processing, Inception-ResNetV2 comes to action as a transfer learning block. Inception-ResNetV2 removes the loss layer, and substitutes it with an output layer that detects the deepfake loss and is called deepfake detection loss output layer which has been already defined during the preprocessing[12].
The fine tuning of the network limits variants from either the data that has been identified in the dataset or the performance forecasted. This model has been compiled for 20 epochs and 40 epochs to master the training dataset. Sigmoid activation function has been used in the model which is useful for neural networks. This function maps required data from the graph to a value between 0 and 1. Further evaluation is performed based on the confusion matrix formed.

5. Detection Model and Results

The designed model was tested for 20 epoch and 40 epoch due to run time limitation and achieved 84.75 percent and 91.48 percent accuracy respectively. The resultant graphs obtained after implementation claims the truth that the validation and testing accuracy increase with increase in number of epoch. Resultant confusion matrix helps to evaluate the testing accuracy of the system.

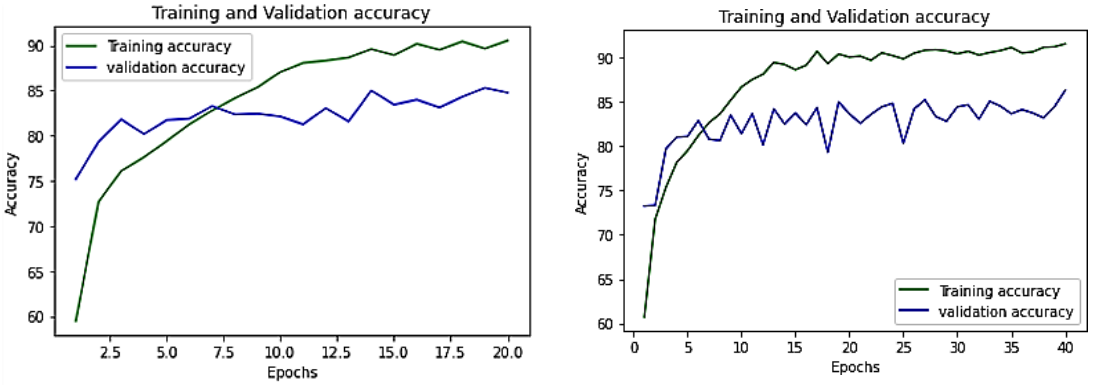

Figure 2: Graphs of Training and Validation accuracy for 20 and 40 Epoch

6. Conclusion and future scope

The faith of the masses has started to disintegrate due to the Deepfakes as the streaming content no longer seems to be authentic and real. In our paper we presented an approach that can automatically detect deep fake based on deep learning concept. In deepfakes the target face appears briefly in a video so the model divides user video into frames and these frames were further preprocessed using InceptionResNetV2 and LSTM. The method provided good level accuracy and reliability. The proposed methodology is capable of analyzing any video using convolutional LSTM system and also helps in detecting deepfake face which has been manipulated therefore preventing individuals from defaming. We can also hold experiments with more number of epochs and Learning Rate to get higher accuracy. In future one can extent this work by exploring more architectures that will help in implementing new detection techniques to detect deepfakes.

Acknowledgments

We take this opportunity to convey our appreciation to our supervisor, Dr(Mrs)Gargi Khanna, Associate Professor, Dept of ECE, NIT Hamirpur, who guided us throughout this project. It was team work so special apprentice to all the team members for their cooperation, hardwork and dedication towards the project.

References

[1] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, A. Mallya, Generative adversarial networks for image and video synthesis: Algorithms and applications, Proceedings of the IEEE 109 (2021) 839–862. doi:10.1109/JPROC.2021.3049196.

[2] Bouarara, H. Ahmed, Recurrent neural network (rnn) to analyse mental behaviour in social media, 2021. doi:10.4018/IJSSCI.2021070101.

[3] A. Jadhav, A. Patange, H. Patil, J. Patel, M. Mahajan, Deep residual learning for image recognition, International Journal for Scientific Research and Development 8 (2020) 1016–1019.

[4] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, 2019. arXiv:1811.00656.

[5] U. A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1. doi:10.1109/TPAMI.2020.3009287.

[6] D. Güera, 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. doi:10.1109/AVSS.2018.8639163.

[7] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: Exposing ai created fake videos by detecting eye blinking, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7. doi:10.1109/WIFS.2018.8630787.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014. arXiv:1406.2661.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[10] R. Raghavendra, K. B. Raja, S. Venkatesh, C. Busch, Transferable deep-cnn features for detecting digital and print-scanned morphed face images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1822–1830. doi:10.1109/CVPRW.2017.228.

[11] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, 2020. arXiv:2006.07397.

[12] R. K. Singh, P. V. Sarda, S. Aggarwal, D. K. Vishwakarma, Demystifying deepfakes using deep learning, in: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1290–1298. doi:10.1109/ICCMC51019.2021.9418477.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impactof residual connections on learning, 2016. arXiv:1602.07261.