Deepfake Video Detection using Neural Networks

Akhil Sunil Kumar[1], Amruta Khavase[2], Himesh Rajendran[3]
[1,2,3] *Pillai college of Engineering, New Panvel*

***Abstract* - In past months, free deep learning-based software tools have made the creation of credible face exchanges in videos that leave few traces of manipulation, in what are known as "DeepFake"(DF) videos. Manipulations of digital videos has been demonstrated for many years through the good use of visual effects, recent advances in deep learning have led to a drastic increase in the making real looking of fake content and the accessibility in which it can be created. These so-called AI-synthesized media. Creating the DeepFakes using Artificially intelligent tools are simple tasks. But, when it comes to detection of these DF, it is a major challenge. Because training the algorithm to spot the DeepFake is not simple. We have taken a step forward in detecting the DeepFakes using Convolutional Neural Network and Recurrent neural Network. System uses a convolutional Neural network to extract features at the frame level. These features are used to train a recurrent neural network which learns to classify if a video is manipulated or not and able to detect the temporary inconsistencies between frames introduced by the DeepFake creation tools. Expected result against a large set of fake videos collected from standard data sets. We show how our system can be competitive and results in using a simple architecture.**

***Index Terms* - convolutional Neural network (CNN), recurrent neural network (RNN).**

I.INTRODUCTION

The increasing betterment of smartphone cameras and the availability of good internet connection all over the world has increased the evergrowing reach of social media and media sharing mediums have made the creation and sharing of digital videos more easy than ever before. The growing computational power has made deep learning so powerful that would have been thought impossible a few years ago. Like any new technology, this has created new challenges. So-called "DeepFakes" produced by deepfake generation models that can manipulate video and audio clips. Spreading of the DF over the social media platforms have become very common leading to spamming and speculating wrong information over the platform. These types of the DF will be terrible, and lead to threatening, misleading of common people.

To overcome such a situation, DeepFake detection is very important. We describe a new deep learning-based method that can distinguish AI-generated fake videos from real videos. It is important to develop technology that can detect fakes, so that the DF can be identified and prevented from spreading over the internet. For detection of the DF it is very important to understand the way Generative.

Adversarial Network (GAN) creates the DF. Generative adversarial network takes as input a video and an image of a specific individual (‘target’), and outputs another video with the target’s faces replaced with those of another individual (‘source’). The backbone of DeepFakes are deep adversarial neural networks trained on face images and target videos to automatically map the faces and facial expressions of the source to the target. With proper post processing, the resulting videos can have a high level of real like effect. The generative adversarial network splits the video into frames and replaces the input image in every frame. Further it reconstructs the video. This process is usually done by using autoencoders. We describe a new deep learning-based method that can distinguish DeepFake videos from the real ones. Our method is based on the same process that is used to create the DeepFake by generative adversarial network. The method is based on the properties of the DF videos, due to limitation of computation resources and production time, the DF algorithm can only synthesize face images of a fixed size, and that they must undergo an affinal warping to match the configuration of the source’s face. This warping leaves some distinguishable artifacts within the output deepfake video thanks to the resolution inconsistency between warped face area and surrounding context. Our method detects such videos by comparing the generated face's and their surrounding regions by splitting video into frames and extracting the features with a ResNext Convolutional Neural Network and using the Recurrent Neural Network with Long Short Term Memory capture the temporary inconsistencies between frames introduced by generative adversarial network during the reconstruction of the DeepFake. To train the ResNext CNN model, we simplify the method by simulating the resolution inconsistency in affine face wrappings directly.

II. LITERATURE SURVEY

The current growth in deep fake video and its illegal use is a major threat to democracy and public trust. Due to this there is an increased demand for fake video analysis, detection and intervention. Some of the related word in deep fake detection are listed below:
ExposingDF Videos by Detecting Face Warping Artifacts used an approach to detect artifacts by comparing the generated face's and its surrounding regions with a dedicated Convolutional Neural Network model. Their method is based on the observations that current DF algorithms can only generate images of limited resolutions, which are then needed to be further transformed to match the faces that is to be replaced in the input video.

Exposing AI Created Fake Videos by Detecting Eye Blinking describes a new method to expose fake face videos generated with deep neural network models. This method is predicated on the detection of the movement of the eyelids within the videos, which may be a physiological signal that's not well presented within the developed fake videos. The method is evaluated over scores of the eyes blinking detection datasets and shows promising performance on detecting videos generated with Deep Neural Network based software DeepFakes. Their method only uses the shortage of blinking as a clue for detection. However certain parameters must be considered for detection of the deepfake like teeth enchantment, wrinkles on faces etc. Our method is proposed to consider all these parameters.

Using capsule networks to detect forged images and videos uses a method that uses a capsule network to detect forged, manipulated images and videos in different scenarios, like replay attack detection and computer-generated video detection. In their method, they used random noise in the training phase which is not a Still the model performed beneficial in their dataset but may fail on real time data due to noise in training. Our method is said to be trained on noiseless and real time datasets.

Detection of Portrait Videos using Biological Signal method to extract biological signals from facial regions on authentic and fake portrait videos. Apply the transformations to calculate the spatial coherence and temporary consistency, capture the signal characteristics in feature sets and PPG maps, and train a probabilistic SVM and a CNN. Then, the mixture authenticity probabilities to make a decision whether the video is fake or authentic. The Fake catcher detects the fake content with high accuracy, independent of the generator, content, resolution, and quality of the video. Due to lack of change detector leading to the loss in their findings to preserve biological signals, formulating a spotable loss function that follows the proposed signal processing steps is not a straight process.

III. PROPOSED SYSTEM

There are many tools available for creating the DeepFakes, but for DeepFakes detection there is hardly any tool available. Our approach for detecting the DF will be a great contribution in avoiding the percolation of the DF over the world wide web. We will be providing a web-based platform for the user for uploading the video and detect if its fake or real. This project is often scaled up from developing a web-based platform to a browser plugin for automatic DF detections. Even big applications like WhatsApp, Facebook can integrate this project with their application for easy pre-detection of DF before sending it to another user. One of the important objectives is to evaluate its performance and acceptability in terms of security, user-friendliness, accuracy and reliability. Our method is focusing on detecting all types of DF like replacement DF, retrenchment DF and interpersonal DF. figure.1 represents the straightforward system architecture of the proposed system:

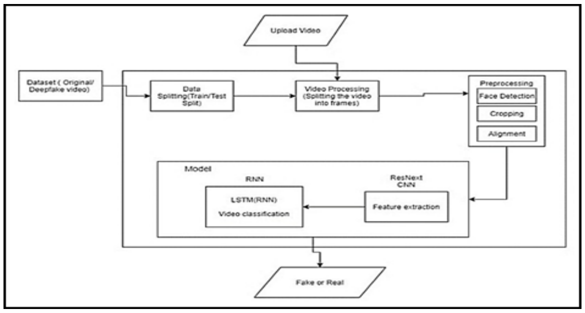


Fig. 1: System Architecture

1. Dataset:

We are using a mixed dataset which consists of equal amounts of videos from different dataset sources like YouTube, FaceForensics++[14], Deep fake detection challenge dataset[13].Our newly prepared dataset contains 50% of the original video and 50% of the manipulated deep fake videos. The dataset is split into two parts 70% train and 30% test set.

2. Preprocessing:

Dataset preprocessing includes splitting of the video into frames. It is followed by face detection and cropping the frame with detected face. To maintain the equality in the number of frames the mean of the dataset video is calculated, and the new processed face dataset is created containing the frames equal to the mean. The frames that do not have faces in it are ignored during preprocessing. As processing the 10 second video at 40 frames per second i.e total 400 frames will require a lot of computational power. So, for experimental purposes we are proposing to use only the first 100 frames for training the model.

3. Model:

The model consists of resnext50_32x4d with one Long-Short Term Memory layer. The Data Loader loads the preprocessed face cropped videos and splits the videos into train and testset. Further the frames from the processed videos are passed to the model for training and testing in mini batches.

4. ResNext CNN for Feature Extraction

Instead of rewriting the classifier, we are proposing to use the ResNext CNN classifier for extracting the features and accurately detecting the frame level features. Following, we will be tuning the network by adding required layers and selecting a proper learning rate to properly converge the gradient descent of the model. The 2048-dimensional feature vectors after the last pooling layers are then used as the LSTM input in sequence.

5. LSTM for Sequence Processing

Let us assume a sequence of ResNext CNN feature vectors of input frames as input and a node neural network with the probabilities of the sequence being part of a deep fake video or an untampered video. The key challenge that we need to address is the de- sign of a model to recursively process a sequence in a meaningful manner. for the problem we are proposing to the use of a 2048 LSTM unit with 0.4 chance of dropout, which is capable of achieving our objective. LSTM is used to process the frames in a sequential manner so that the temporary analysis of the video can be made, by comparing the frame at 't' second with the frame of 't-n' seconds. Here n(frames) can be any number of frames before t(seconds).

6. Predict:

The new video is passed to the trained model for the prediction. The new video is preprocessed to bring the new video in the format of the trained model. The video is split into frames followed by face cropping and instead of storing the video into local storage the cropped frames are directly passed to the trained model for detection.

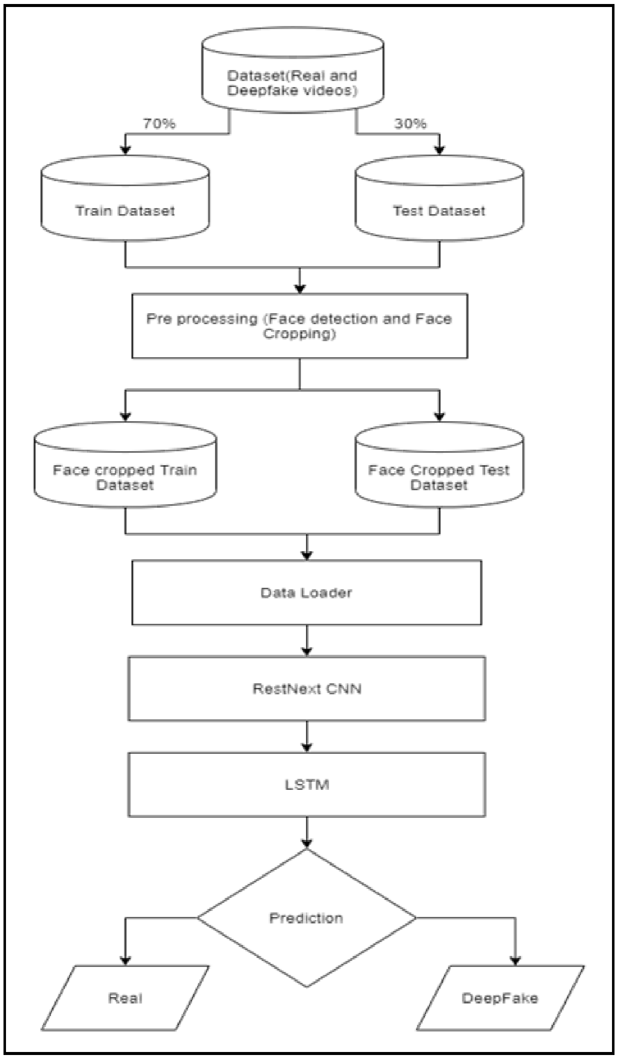
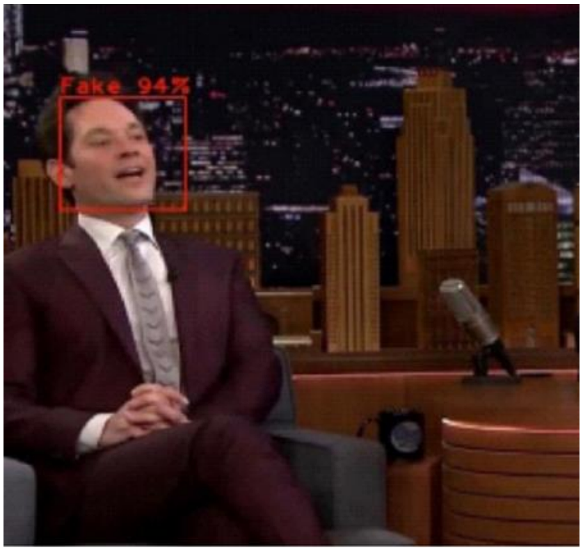

Dataset(Real and Deepfake videos)
70%
30%
Train Dataset
Test Dataset
Pre processing (Face detection and Face Cropping)
Face cropped Train Dataset
Face Cropped Test Dataset
Data Loader
RestNext CNN
LSTM
Prediction
Real
DeepFake

Fig2: Training Flow



IV. RESULT

Fig. 2: Training Flow The output of the model is going to be whether the video is deepfake or a real video along with the confidence of the model. One example is shown in figure 3.

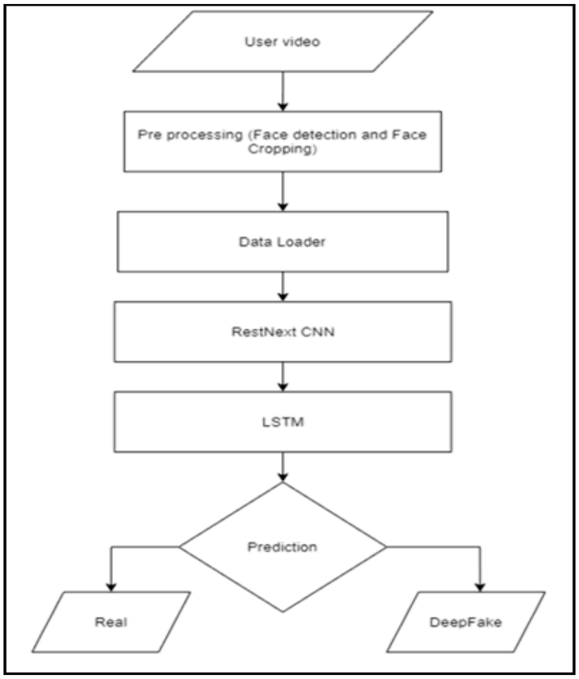


Fig. 4: Prediction flow

V. CONCLUSION

We presented a neural network-based approach to classify the video as deep fake or real, along with the confidence of the proposed model. The proposed method is inspired by the way the deep fakes are created by the GANs with the help of Autoencoders. Our method does the frame level detection using ResNext CNN and video classification using RNN along with LSTM. The proposed method is capable of detecting the video as a deep fake or real based on the listed parameters in paper. We believe that it will provide a very high accuracy on real time data.

VI. LIMITATIONS

Our method has not considered the audio. That is why our method will not be able to detect the audio deep fake. But we are proposing to achieve the detection of the audio deep fakes in the future.

REFERENCE

[1] Haonan Chen, Yaowu Chen, Xiang Tian, Rongxin Jiang, "Cascade face spoofing detector based on face anti spoofing RCNN and improved retinex LBP", December 9,2019.

[2] Xin Yang, Yuezun Li and Siwei Lyu, "Exposing deep fakes using inconsistent headposes", 2019.

[3] S.K Yarlagadda, D.Guera, D.M Montserrat, F.M. Zhu, E.J. Delp, "Shadow removal detection and localization for forensics analysis" , 2019.

[4] Khodabakshsh, Raghavendra Ramachandra, Christoph Busch, "Subjective evaluation of media consumer vulnerability to fake audiovisual content", 2019.

[5] Khodabakshsh, Raghavendra Ramachandra, Kiran Raja, Pankaj Wasnik,Christoph Busch, "Fake face detection methods: Can they be generalized" , 2018.