Deepfake Video Detection using Machine Learning

Abhijit Jadhav[1] Sakshi Lokare[2] Priya Shinde[3]

[1,2,3]Student [4]Prof. Kavita Khemnar

[1,2,3,4]Department of Artificial Intelligence & Machine Learning

[1,2,3,4]Sahyadri Valley COE & Technology Pune, Maharashtra, India

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

Keywords: Deepfake Video Detection, convolutional Neural network (CNN), recurrent neural network (RNN)

I. INTRODUCTION

The increasing sophistication of smartphone cameras and the availability of good internet connection all over the world has increased the ever-growing reach of social media and media sharing portals have made the creation and transmission of digital videos more easy than ever before. The growing computational power has made deep learning so powerful that would have been thought impossible only a handful of years ago. Like any transformative technology, this has created new challenges. So-called "DeepFake" produced by deep generative adversarial 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 peculating wrong information over the platform. These types of the DF will be terrible, and lead to threating, misleading of common people. [2]

To overcome such a situation, DF detection is very important. So, we describe a new deep learning-based method that can effectively distinguish AI-generated fake videos (DF Videos) from real videos. It's incredibly important to develop technology that can spot fakes, so that the DF can be identified and prevented from spreading over the internet.

For detection the DF it is very important to understand the way Generative Adversarial Network (GAN) creates the DF. GAN 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 DF 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 postprocessing, the resulting videos can achieve a high level of realism. The GAN split the video into frames and replaces the input image in every frame. Further it reconstructs the video. This process is usually achieved by using autoencoders. We describe a new deep learning-based method that can effectively distinguish DF videos from the real ones. Our method is based same process that is used to create the DF by GAN. The method is based on a property 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 they must undergo an affinal warping to match the configuration of the source's face. This warping leaves some distinguishable artifacts in the output deepfake video due to the resolution inconsistency between warped face area and surrounding context. Our method detects such artifacts by comparing the generated face areas and their surrounding regions by splitting the video into frames and extracting the features with a ResNext Convolutional Neural Network (CNN) and using the Recurrent Neural Network (RNN) with Long Short Term Memory(LSTM) capture the temporal inconsistencies between frames introduced by GAN during the reconstruction of the DF. To train the ResNext CNN model, we simplify the process by simulating the resolution inconsistency in affine face wrappings directly.[4]

II. LITERATURE SURVEY

The explosive growth in deep fake video and its illegal use is a major threat to democracy, justice, and public trust. Due to this there is a increased the 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 [1] used an approach to detects artifacts by comparing the generated face areas and their surrounding regions with a dedicated Convolutional Neural Network model. In this work there were two-fold of Face Artifacts.

Their method is based on the observations that current DF algorithm can only generate images of limited resolutions, which are then needed to be further transformed to match the faces to be replaced in the source video.

Exposing AI Created Fake Videos by Detecting Eye Blinking [2] describes a new method to expose fake face videos generated with deep neural network models. The method is based on detection of eye blinking in the videos, which is a physiological signal that is not well presented in the synthesized fake videos. The method is evaluated over benchmarks of eye-blinking detection datasets and shows promising performance on detecting videos generated with Deep Neural Network based software DF.

Their method only uses the lack of blinking as a clue for detection. However certain other parameters must be considered for detection of the deep fake 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 [3] 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 have used random noise in the training phase which is not a good option. Still the model performed beneficial in their dataset but may fail on real time data due to noise in training. Our method is proposed to be trained on noiseless and real time datasets.

Detection of Synthetic Portrait Videos using Biological Signals [4] approach extract biological signals from facial regions on authentic and fake portrait video pairs. Apply transformations to compute the spatial coherence and temporal consistency, capture the signal characteristics in feature sets and PPG maps, and train a probabilistic SVM and a CNN. Then, the aggregate authenticity probabilities to decide whether the video is fake or authentic.

Fake Catcher detects fake content with high accuracy, independent of the generator, content, resolution, and quality of the video. Due to lack of discriminator leading to the loss in their findings to preserve biological signals, formulating a differentiable loss function that follows the proposed signal processing steps is not straight forward process.

III. PROPOSED SYSTEM

There are many tools available for creating the DF, but for DF detection there is hardly any tool available. Our approach for detecting the DF will be 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 to upload the video and classify it as fake or real. This project can be scaled up from developing a web-based platform to a browser plugin for automatic DF detections. Even big application like WhatsApp, Facebook can integrate this project with their application for easy pre detection of DF before sending 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 simple system architecture of the proposed system: -

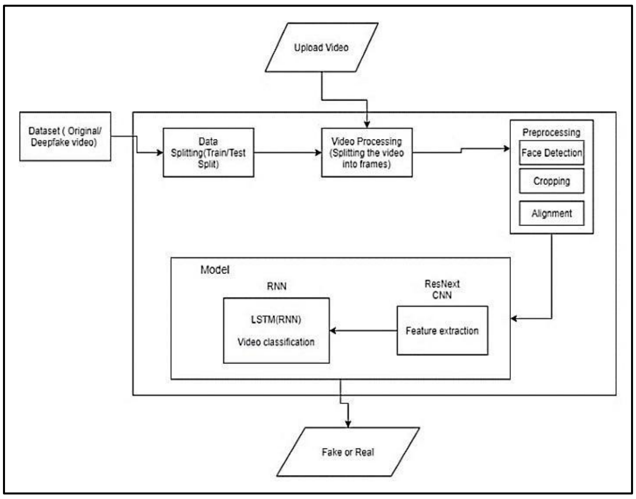

Fig. 1: System Architecture

A. *Dataset:*

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

B. *Preprocessing:*

Dataset preprocessing includes the splitting the video into frames. Followed by the face detection and cropping the frame with detected face. To maintain the uniformity in the number of frames the mean of the dataset video is calculated and the new processed face cropped dataset is created containing the frames equal to the mean. The frames that doesn't have faces in it are ignored during preprocessing.

As processing the 10 second video at 30 frames per second i.e total 300 frames will require a lot of computational power. So for experimental purpose we are proposing to used only first 100 frames for training the model.

C. *Model:*

The model consists of resnext50_32x4d followed by one LSTM layer. The Data Loader loads the preprocessed face cropped videos and split the videos into train and test set. Further the frames from the processed videos are passed to the model for training and testing in mini batches.

D. *ResNext CNN for Feature Extraction*

Instead of writing the 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 fine-tuning the network by adding extra 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 sequential LSTM input.

E. *LSTM for Sequence Processing*

Let us assume a sequence of ResNext CNN feature vectors of input frames as input and a 2-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 this problem, we are proposing to the use of a 2048 LSTM unit with 0.4 chance of dropout, which is capable to do achieve our objective. LSTM is used to process the frames in a sequential manner so that the temporal analysis of the video can be made, by comparing the frame at 't' second with the frame of 't-n' seconds. Where n can be any number of frames before t.

F. Predict:

A new video is passed to the trained model for prediction. A new video is also preprocessed to bring 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.

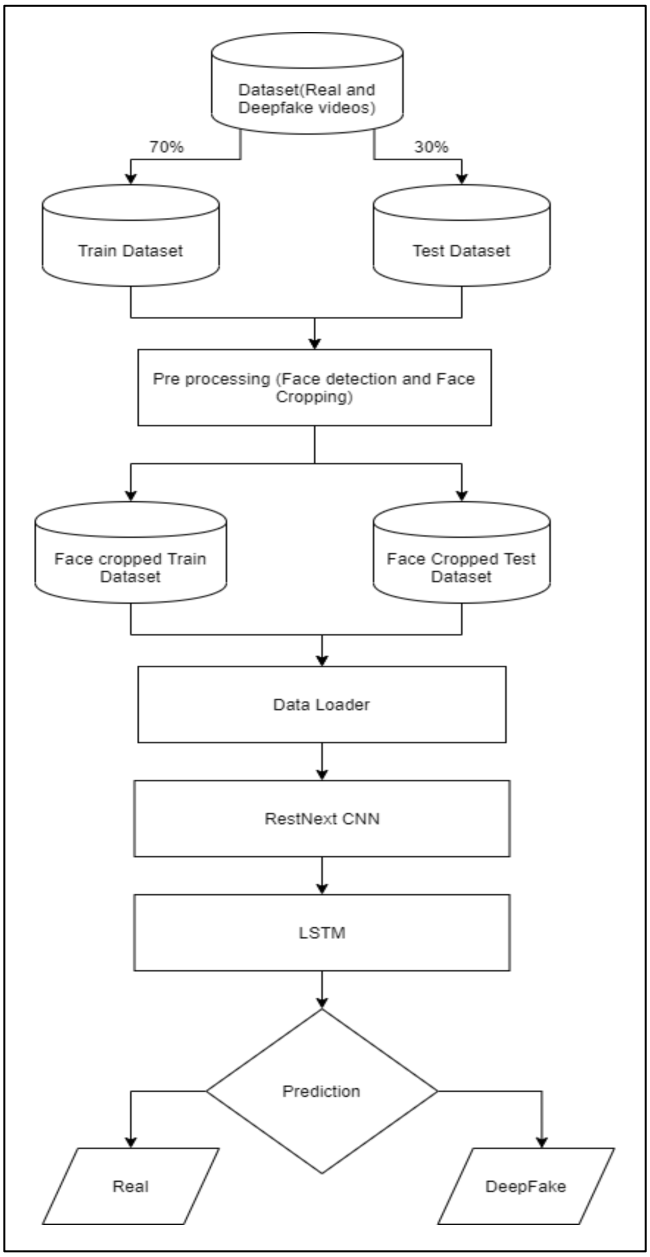
Dataset(Real and Deepfake videos)
70%
30%
Train Dataset
Test Dataset
Pre processing (Face detection and Face Cropping)
Face cropped Train Dataset
Face Cropped Test Dataset
Data Loader
RestNext CNN
LSTM
Prediction
Real
DeepFake

Fig. 2: Training Flow

IV. Result

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 the figure 3.

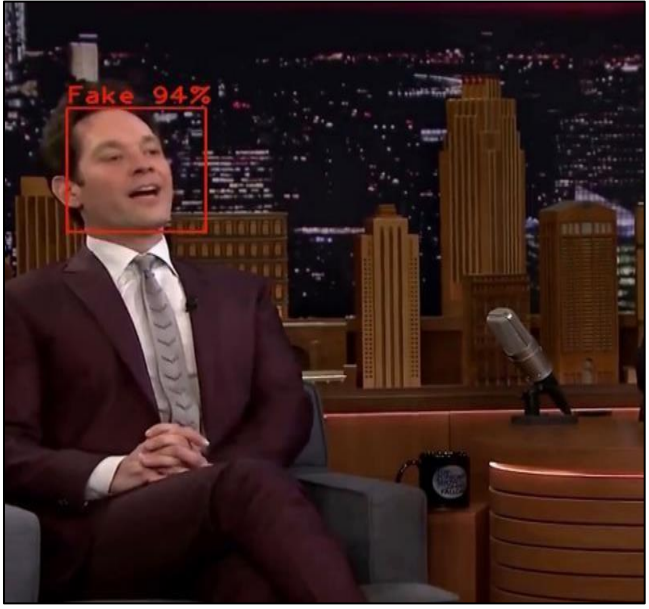

Fig. 3: Expected Results

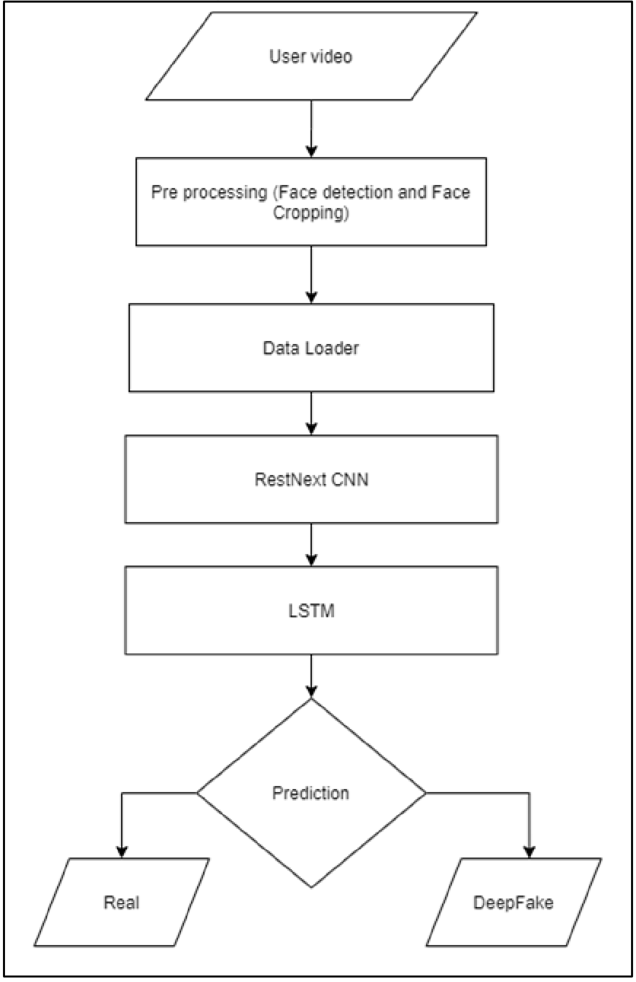

Fig. 4: Prediction flow

V. Conclusion

We presented a neural network-based approach to classify the video as deep fake or real, along with the confidence of 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. The system was tested using a number of metrics, and the results demonstrated that it is more effective at identifying deep fakes. The evaluation outcomes implied that, even when endured with fraudulent faces, the suggested system can recognize faces with high accuracy of 91.82% and a F1 score of 91% which has been developed using an easily understandable CNN network rather than complex techniques like SVM, Boosting, Transfer Learning and other advanced architectures.

VI. Limitations

Our method has not considered the audio. That's 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.

References

[1] Yuezun Li, Siwei Lyu, “ExposingDF Videos By Detecting Face Warping Artifacts,” in arXiv:1811.00656v3.

[2] Yuezun Li, Ming-Ching Chang and Siwei Lyu “Exposing AI Created Fake Videos by Detecting Eye Blinking” in arxiv.

[3] Huy H. Nguyen , Junichi Yamagishi, and Isao Echizen “ Using capsule networks to detect forged images and videos ”.

[4] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari and Weipeng X “Deep Video Portraits” in arXiv:1901.02212v2.

[5] Umur Aybars Ciftci, ˙Ilke Demir, Lijun Yin “Detection of Synthetic Portrait Videos using Biological Signals” in arXiv:1901.02212v2.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[7] David G¨uera and Edward J Delp. Deepfake video detection using recurrent neural networks. In AVSS, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

[9] An Overview of ResNet and its Variants https://towardsdatascience.com/an-overview-of-resnetand-its-variants-5281e2f56035

[10] Long Short-Term Memory: From Zero to Hero with Pytorch:https://blog.floydhub.com/long-short-termmemory-from-zero-to-hero-with-pytorch/

[11] Sequence Models And LSTM Networks https://pytorch.org/tutorials/beginner/nlp/sequence_models_tutorial.html

[12] https://discuss.pytorch.org/t/confused-about-the-imagepreprocessing-in-classification/3965

[13] https://www.kaggle.com/c/deepfake-detectionchallenge/data

[14] https://github.com/ondyari/FaceForensics

[15] Y. Qian et al. Recurrent color constancy. Proceedings of the IEEE International Conference on Computer Vision, pages 5459–5467, Oct. 2017. Venice, Italy.

[16] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-toimage translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5967–5976, July 2017. Honolulu, HI.

[17] R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch, “Transferable deep-CNN features for detecting digital and print-scanned morphed face images,” in CVPRW. IEEE, 2017.

[18] Tiago de Freitas Pereira, Andr´e Anjos, Jos´e Mario De Martino, and S´ebastien Marcel, “Can face anti spoofing countermeasures work in a real world scenario?,”in ICB. IEEE, 2013.

[19] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, “Distinguishing computer graphics from natural images using convolution neural networks,” in WIFS. IEEE, 2017.

[20] F. Song, X. Tan, X. Liu, and S. Chen, “Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients,” Pattern Recognition, vol. 47, no. 9, pp. 2825–2838, 2014.

[21] D. E. King, “Dlib-ml: A machine learning toolkit,” JMLR, vol. 10, pp. 1755–1758, 2009.