Deepfake Detection For Video Using Deep Learning

Faiz Shaikh[1], Shefali Shrivastava[2], Shreema Shetty[3], Prof. Krishnendu Nair[4]
[1, 2, 3] Dept of Information Technology
[4]Professor, Dept of Information Technology
[1, 2, 3, 4] Pillai College of Engineering,
Navi Mumbai, Maharashtra, India

Abstract- *The term DeepFakes is coined using two terms — deep learning and fake. DeepFakes is typically used to refer to images and videos produced synthetically using an algorithm wherein the person's face in the image or video is replaced by another or the audio in a video is changed and the result looks quite authentic. DeepFakes are used to create fake news, hoaxes, pornographic content starring celebrities, conduct frauds and manipulate views of the public during elections. The main machine learning methods used to create DeepFakes are based on deep learning and involve training generative neural network architectures, such as autoencoders or generative adversarial networks (GANs)[1]. Thus, as the threat of Deepfakes increases, there is a need to find a way to detect them. Most image forensics techniques are usually not well suited to videos due to the compression that strongly degrades the data. The project follows a deep learning approach that uses Convolutional Neural Networks and Recurrent Neural Networks. At the frame level, the system employs a convolutional Neural network (CNN) to extract features. These characteristics are used to train a recurrent neural network (RNN), which learns to classify whether or not a video has been manipulated and can detect the temporal irregularities between frames introduced by DF generation tools.*

Keywords- Deepfake, CNN, RNN, LSTM, neural network, deep learning

I. INTRODUCTION

Deepfakes are synthetic media. They use a form of artificial intelligence called deep learning to make images and videos of fake events, hence the name deepfake. Deepfake (derived from "deep learning" and "fake") superimposes an image of the target's face on the source video to create a video of the target doing or saying what the source is doing. It is a technique that can be done. The underlying mechanisms for creating deepfake are deep learning models such as autoencoders and generative hostile networks that are widely used in the field of computer vision**.**These models are used to examine facial expressions and movements of a person and synthesize facial images of another person making analogous expressions and movements[2].The deepfake method usually requires a large amount of image or video data to train the model to create realistic images or videos. Public figures such as celebrities and politicians are the first targets of deepfake because of the large number of videos and images available online. Deep Fakes were used to swap faces of celebrities or politicians to bodies in porn images and videos. The first deepfake video emerged in 2017 where the face of a celebrity was swapped to the face of a porn actor. It is threatening to world security when deepfake methods can be employed to create videos of world leaders with fake speeches for falsification purposes. Deep Fakes therefore can be abused to cause political or religion tensions between countries,to fool public and affect results in election campaigns, or create chaos in financial markets by creating fake news.It can be even used to generate fake satellite images of the Earth to contain objects that do not really exist to confuse military analysts,e.g.,creating a fake bridge across a river although there is no such a bridge in reality[2]. This can mislead a troop who has been guided to cross the bridge in a battle.

How are Deepfakes created

Deepfakes use an autoencoder, which is a form of neural network. An encoder reduces an image to a lower-dimensional latent space, while a decoder reconstructs the image from the latent representation[3]. Deepfakes use this architecture by encoding a person into the latent space with a universal encoder. Key characteristics of their facial features and body posture are included in the latent representation. After that, a model trained particularly for the target can decode it. This means that the target's specific information will be superimposed on the latent space's underlying facial and body traits from the original video.

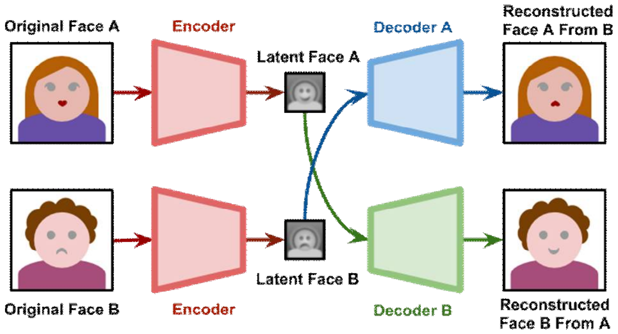
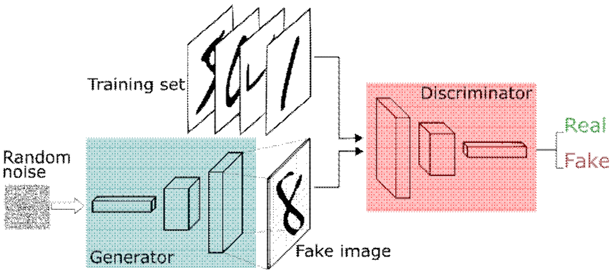


Fig 1.1: Deepfakes generated using autoencoder

The addition of a generative adversarial network to the decoder is a popular improvement to the aforementioned architecture. In an adversarial interaction, a GAN trains a generator, in this case the decoder, and a discriminator. The generator generates new images from the source material's latent representation, while the discriminator determines whether or not the image is generated. As a result, the generator produces images that closely resemble reality, as any flaws would be detected by the discriminator. In a zero-sum game, both algorithms advance over time. Deepfakes are difficult to combat because they are continually changing; if a flaw is identified, it may be fixed. [4]

Deepfake Detection

Deepfakes are increasingly detrimental to privacy, society security and democracy. Methods for detecting deepfakes have been proposed as soon as this threat was introduced.[5] As the quality of deepfakes increases, it becomes more and more difficult for humans to detect them. Early attempts were based on handcrafted features obtained from artifacts and inconsistencies of the fake video synthesis process. Recent methods, on the other hand, applied deep learning to automatically extract salient and discriminative features to detect deepfakes With the extensive research done on face manipulation in images and videos using AI this field is growing at a faster pace. It will get easier and faster to create deepfakes of high quality, the result would be a huge volume of deepfake content which may be too great for human detection alone. Therefore, there is a need to rely on algorithms to determine whether a content is deepfake or not. Deepfake detection research is ongoing. Not only are algorithms automatic ,they can also potentially detect cues that are hard for humans to find.



II. LITERATURE SURVEY

A. An Examination of Fairness of AI Models for Deepfake Detection: Loc Trinh, Yan Liu thoroughly measured the predictive performance of popular deepfake detectors on racially aware datasets balanced by gender and race. Large disparities in predictive performances across races, as well as large representation bias in widely used FaceForensics++ was found.This paper echoes the importance of benchmark representation and intersectional auditing for increased demographic transparency and accountability in AI systems.

*B. Deepfake UCL:Deepfake Detection via Unsupervised Contrastive Learning:*In this paper, Sheldon Fung, Xuequan Lu, Chao Zhang, Chang-Tsun Li have done the deepfake detection via unsupervised contrastive learning.The model first generates two different transformed versions of an image and feeds them into two sequential sub-networks, i.e., an encoder and a projection head.The unsupervised training is achieved by maximizing the correspondence degree of the outputs of the projection head.To evaluate the detection performance of the unsupervised method, unsupervised features are used to train an efficient linear classification network.

*C. DeepFake Detection by Analyzing Convolutional Traces:*Luca Guarnera, Oliver Giudice, Sebastiano Battiato proposed the assumption that local correlation of pixels in Deepfakes are dependent exclusively on the operations performed by all the layers present in the GAN which generate it; specifically the (latter) transpose convolution layers. This detection method is based on features extracted through the EM algorithm. The under-lying fingerprint has been proven to be effective to discriminate between images generated by recent GANs architectures specifically devoted to generating realistic people’s faces.

D. Deepfake Detection using Capsule Networks with Long Short-Term Memory Networks: In this paper Akul Mehra describes a model that exploits the inconsistencies and identifies real and fake videos and is our contribution towards deepfake detection.Capsule Network is introduced to detect spatial inconsistencies in a single frame and then combined with LSTM to detect the spatio-temporal inconsistencies across multiple frames.

*E. Methods of Deepfake Detection Based on Machine Learning:*In this work, Artem A. Maksutov; Viacheslav O. Morozov; Aleksander A have described a summary of indicators that can be used to decide whether video or photo was changed. Their choice of building model is face warping artifacts detection that is one of the best indicators of fake video/photo right now. A great part of present deepfake algorithms can synthesize only low quality resolution faces. Such transformations leave distinctive artifacts that can be detected.

*F. The Deepfake Challenges and Deepfake Video Detection:*This paper proposes deepfake video detection using CNN and LSTM based on eye blinking rate tested on UADFV publicly available dataset.The VGG16 and ResNet-50 based CNN model are trained on a training dataset that contains open and closed eyes regions.The eye blinking rate enables to detect fake video from real videos.

G. Deepfake Video Detection using Neural Networks: Detecting the DF using CNN and RNN. 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.

H. FaceForensics++: Learning to Detect Manipulated Facial Images: This research paper shows how manipulated images can be detected using trained forgery detectors. It is particularly encouraging that even the challenging case of low-quality video can be tackled by learning-based approaches, where humans and hand-crafted features exhibit difficulties.

I. Deepfake Video Detection Using Recurrent Neural Networks: This research paper presents a temporal-aware system to automatically detect deepfake videos.It proposes a two-stage analysis composed of a CNN to extract features at the frame level followed by a temporally-aware RNN network to capture temporal inconsistencies between frames introduced by the face-swapping process. A simple convolutional LSTM structure is used which can accurately predict if a video has been subject to manipulation or not with as few as 2 seconds of video data.

J. MesoNet: a Compact Facial Video Forgery Detection Network: In this research paper the method to detect forged videos was placed at a mesoscopic level of analysis. The architecture is based on well-performing networks for image classification that alternate layers of convolutions and pooling for feature extraction and a dense network for classification.

III. PROPOSED WORK

We have proposed a system where users can upload a video to test whether it is a deepfake or not. The trained model will assess the uploaded video and give results based on it's analysis.

3.1 System Architecture

The system architecture is given in Figure 3.1. Each block is described in this Section.

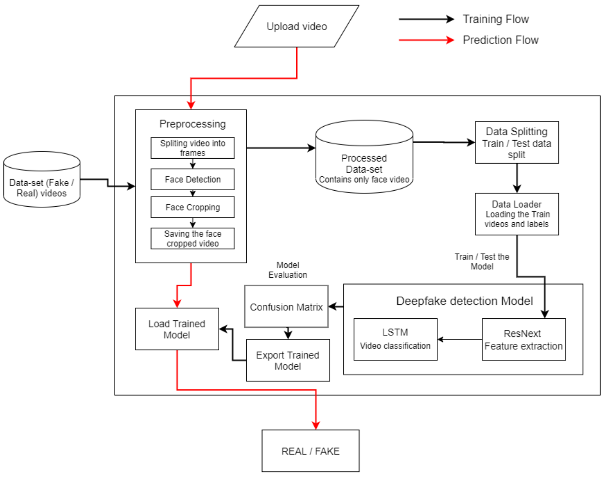

Upload video
Training Flow
Prediction Flow
Data-set (Fake / Real) videos
Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
ResNext Feature extraction
Load Trained Model
Export Trained Model
REAL / FAKE

Fig. 3.1 Proposed system architecture

A. Dataset:

We have used Celeb-DF (v2) as our primary focus is detecting celeb deepfakes as they are the most common victims.

B.Preprocessing:

Dataset preprocessing involves splitting the video into frames. Faces are subsequently detected and the frame is cropped with the detected faces. To maintain a consistent frame count, the dataset video is averaged and a new processed face cropped dataset is created with frames equal to the average. Frames that do not contain faces are ignored during preprocessing.

C. Model:

The model is made up of resnext50 32x4d and one LSTM layer. The Data Loader loads the preprocessed face cropped films and divides them into two sets, one for training and one for testing. Furthermore, The model receives frames from the processed videos.Mini batches are used for teaching and testing.

D. ResNext CNN for Feature Extraction:

We use the ResNext CNN classifier for extracting features and reliably recognising frame level characteristics. Following that, we'll fine-tune the network by adding additional required layers and setting a correct learning rate to ensure that the gradient descent of the model is properly converged.After the final pooling layers, the 2048-dimensional feature vectors are used as the sequential LSTM input.

E. LSTM for Sequence Processing:

The LSTM is used to analyse the frames sequentially in order to do a temporal analysis of the video by comparing the frame at 't' second with the frame at 't-n' seconds. Before t, n can be any number of frames.

F. Predict:

The trained model is given a new video to forecast. A new video is also preprocessed to bring in the trained model's format. The video is divided into frames, then face cropped, and instead of storing the video locally, the cropped frames are transmitted immediately to the trained model for identification.

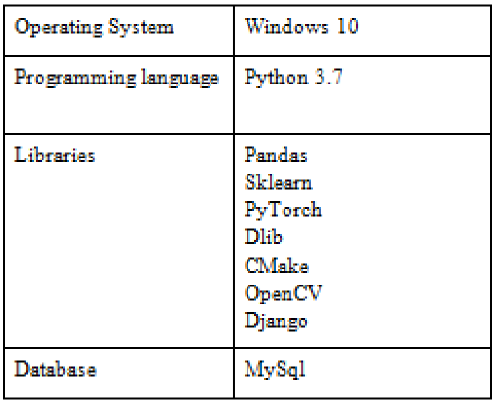
IV. REQUIREMENT ANALYSIS

The implementation detail is given in this section.

4.1 Software

The software's minimum requirement are:

Table 2 Software Details

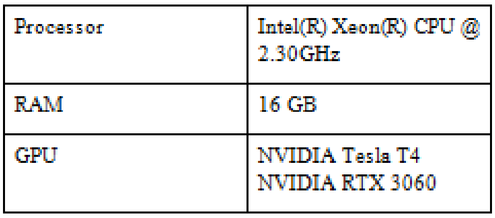


Operating System	Windows 10
Programming language	Python 3.7
Libraries	Pandas Sklearn PyTorch Dlib CMake OpenCV Django
Database	MySql

4.2 Hardware

The minimum hardware requirement are:

Table 2 Hardware Details



Processor	Intel(R) Xeon(R) CPU @ 2.30GHz
RAM	16 GB
GPU	NVIDIA Tesla T4 NVIDIA RTX 3060

4.3 Dataset and Parameters

The dataset used in our project is Celeb-DF (v2): A New Dataset for DeepFake Forensics. Celeb-DF (v2) dataset contains real and DeepFake synthesized videos having similar visual quality on par with those circulated online.Celeb-DF includes 590 original videos collected from YouTube with subjects of different ages, ethic groups and genders, and 5639 corresponding DeepFake videos. The DeepFake videos in Celeb-DF are generated using an improved DeepFake synthesis algorithm, which is key to the improved visual quality. Specifically, the basic DeepFake maker algorithm is refined in several aspects targeting the following specific visual artifacts observed in existing datasets.[6] FaceForensics++ is a forensics dataset consisting of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures. We have used 100 videos of each type of manipulation method.[7]

V. ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Prof. Krishnendu Nair for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Satish Kumar Verma and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] https://testbook.com/current-affairs/artificial-intelligence-and-the-technology-of-deep-fake/

[2] Nguyen, Thanh & Nguyen, Cuong M. & Nguyen, Tien & Duc, Thanh & Nahavandi, Saeid. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey.

[3] https://www.seniorcare2share.com/what-is-a-deepfake/

[4] https://towardsdatascience.com/what-the-heck-are-vae-gans-17b86023588a

[5] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Saeid Nahavandi: Deep Learning for Deepfakes Creation and Detection. CoRR abs/1909.11573 (2019)

[6] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, Siwei Lyu: Celeb-DF: A New Dataset for DeepFake Forensics. CoRR abs/1909.12962 (2019)

[7] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. CoRR abs/1901.08971 (2019)

[8] S. Fan, R. Wang, T.-T. Ng, C. Y.-C. Tan, J. S. Herberg, and B. L. Koenig. Human perception of visual realism for

photo and computer-generated face images. ACM Transactions on Applied Perception (TAP), 11(2):7, 2014.

[9] H. Farid. A Survey Of Image Forgery Detection. IEEE Signal Processing Magazine, 26(2):26–25, 2009.

[10] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormahlen,P. Perez, and C. Theobalt. Automatic face reenactment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4217–4224, 2014.

[11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift.arXiv preprint arXiv:1502.03167, 2015.

[12] B. Balas and C. Tonsager. Face animacy is not all in the eyes:Evidence from contrast chimeras. Perception, 43(5):355–367, 2014.

[13] M. Barni, L. Bondi, N. Bonettini, P. Bestagini, A. Costanzo, M. Maggini, B. Tondi, and S. Tubaro. Aligned and nonaligned double jpeg detection using convolutional neural networks. Journal of Visual Communication and Image Representation, 49:153–163, 2017.

[14] B. Bayar and M. C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pages 5–10.ACM, 2016.

[15] F. Chollet. Xception: Deep learning with depth wise separable convolutions. arXiv preprint, pages 1610–02357, 2017.

[16] F. Chollet et al. Keras. https://keras.io, 2015.

[17] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. University of Montreal, 1341(3):1, 2009.