DEEPFAKE DETECTION USING GAN

[1]Ms.A.Punidha, [2]Deepika T, [3]Abhishek Sharma, [4]Dividharshana K,
[5]Arul Karthikeyan M, [6]Hari Prasath M

Department of Computer Science and Engineering
Coimbatore Institute of Technology
Coimbatore, India.

***Abstract*- Nowadays, people faced an emerging problem of AI synthesized face swapping videos, widely known as the Deep Fakes. These kinds of videos can be created to cause threats to privacy, fraudulence, and so on. Sometimes good quality DeepFake video recognition could be hard to distinguish with people's eyes. These can be used to manipulate public opinion during elections, commit fraud, and discredit or blackmail people. Therefore, there is a need for automated tools capable of detecting false multimedia content and avoiding the spread of dangerous false information. That's why researchers need to develop algorithms to detect them. One of the frequently studied deep learning techniques is GAN. These networks are commonly used to create DeepFake videos but are not used to detect them. Here exploring a solution based on a GAN discriminator to detect DeepFake videos. Train a GAN and extract the discriminator into a custom module for DeepFake detection. Test different discriminator architectures using different data sets to see how discriminator performance varies with different settings and training methods.**

***Keywords*- deep learning; neural networks, face swapping indicators, GAN-Generative Adversarial Networks, PCL- Pair-wise self-consistency learning, LSTM-Long Short-Term Memory.**

I. INTRODUCTION

Deepfakes are considered the major threat to AI in the ever-growing social media platforms. There are many scenarios where these realistic face-swapped deepfakes are used to create political distress, fake terrorism events, revenge porn, and blackmail people are easily envisioned. Some examples are Brad Pitt, and Angelina Jolie nude videos.

It becomes very important to spot the difference between the deepfake and pristine video. We are using AI to fight AI. Deepfakes are created using tools like FaceApp and Face Swap, which use pre-trained neural networks like GAN or Auto encoders for these deepfakes creation. Our method uses an LSTM-based artificial neural network to process the sequential temporal analysis of the video frames and pre-trained Res-Next CNN to extract the frame-level features. ResNext Convolution neural network extracts the frame-level features and these features are further used to train the Long Short Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real.

To emulate the real time scenarios and make the model perform better on real-time data, we trained our method with a large amount of balanced and combination of various available datasets like FaceForensic++, Deepfake detection challenge, and Celeb-DF.

Further to make the ready to use for the customers, we have developed a front-end application where the user will upload the video. The video will be processed by the model and the output will be rendered back to the user with the classification of the video as deepfake or real and confidence of the model.

II. RELATED WORK

The Literature Survey was carried out in great detail using existing ALPR methodologies and techniques. MD Shohel Rana and Mohammad Nur Nobi.[1] explore various approaches in deep fake detection and have described the problems raised by Deepfake methods to improve the performance. The methods mentioned are by using Machine Learning- The tree-based ML approaches, Deep Learning - GAN-image artifacts, Statistical Measurements, Blockchain- MFCC

Artem A. Maksutov. [2] this paper presents an algorithm that can decide whether a photo was changed with DeepFake face swapping technology or not with high accuracy. As for model, DenseNet169 is used with face warping artifacts indicator. It should work correctly with the present algorithms of Deepfake so we decided to try it. The photo of people is collected from the internet. Negative examples of photos are used to add some noise to them. Due to resolution limits, DeepFake algorithm cannot produce small moving parts with good quality. Therefore, sometimes it produces artifacts on hairs, eyebrows, eyelashes, or some small skin defects.

Asad Malik [3]. This paper discussed the different Neural Networks, DeepFake Production methods, DeepFake Detection methods, and the limitations of both. The methodology used here is Generative Adversial Neural Network (GANN) with Discriminator and Generator. The former is used to produce the deepfakes and the latter detects them whether it is fake or not. The deepfakes are produced by Full face synthesis, Attribute Manipulation, Identity Swap, and Expression Swap. They efficiently produce and detect the deepfakes which are produced by these methods.

Ruben Tolosana [4]. DeepFake Production methods along with the public datasets and the method for detecting the deepfakes produced by those methods. The various methodologies used for deepfake recognition are Full face synthesis, Attribute manipulation, Identity swap, and Expression swap. The existing methodologies of deep fake production techniques and the improvements in those techniques are to produce and detect those deepfakes.

Siwei Lyu. [5], This paper discusses the current three major types of DeepFake videos. Head and upper-shoulder, Face swapping involves generating a video of the target with the faces replaced by synthesized faces. Lip syncing is to create a falsified video by only manipulating the lip region so that the target appears to speak something that s/he does not speak in reality. It discusses the various methods to detect the deepfakes like DNN splicing detection methods, physical/physiological aspects of deepfake videos, and data-driven, which directly employ various types of DNNs trained on real and DeepFake videos but capturing specific artifacts. Introduces the deepfake problem and discusses the advantages of Deepfake technology.

Deng Pan. [6] this paper explained the essential concepts in different deep learning solutions to automatically classify and hence detect deep fake videos. Specifically, we utilize FaceForensics++ as the source video data and used this data to train two neural networks using pre-processed images. The training of each network produces four models, each corresponding to one of four different mainstream deepfake software platforms. The method used here is deep learning and the process involves Data Pre-process, Split Dataset, Train model, and Evaluate modelAll rates were above 85% and some rates could achieve over 98% on all datasets (except for Neural Textures).

T.Jung. [7] Detects Deepfakes generated through the generative adversarial network (GANs) model via an algorithm called DeepVision to analyze a significant change in the pattern of blinking, which is a voluntary and spontaneous action that does not require conscious effort. The methods used are GAN and deep learning DeepVision is a new integrity verification method performing on a frame basis.

M.A. Hoque. [8] This paper discusses the different types of modification of images/videos and surveys the corresponding methods and tools. They highlight the ongoing efforts to detect fake images and videos using advanced machine learning tools and fact-checking. It provides a holistic blockchain-based solution. The methodologies used here are Neural networks, image classification, and blockchain-all the data is stored in a blockchain

Duha A.Sultan. [9] An inclusive study is presented on the existing techniques used for creating and detecting fake materials and analyzing these techniques that are used by several researchers in addition to the great role of artificial intelligence and deep learning in improving them. Techniques used here Two-step learning approach, Convolutional neural network (CNN), Pair-wise self-consistency learning (PCL), YOLO-CNN, and Optical flow-CNN. The use of CNN to extract visual artifacts in a frame and LSTM to extract temporal features across multiple frames gave the best results for the classification of real or fake videos.

Haya R. HASAN [10] In this paper, a general framework using Ethereum smart contracts to trace and track the provenance and history of digital content to its original source even if the digital content is copied multiple times. The methodology used is InterPlanetary File System (IPFS), Smart contract is written in Solidity language, Blockchain network costs Gas which is paid in Ether tokens. The cost estimate is minimal and is always under 0.095 USD per transaction. It can also be used in other types of digital content such as audio, photos, images, and manuscripts

III. PROPOSED METHODOLOGY

The proposed system employs a method to detect that the photos or videos are changed or morphed in social networks by using ResNext Convolution neural network extracts the frame-level feature sand these features are further used to train the Long Short Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real.

METHODOLOGY

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised, or unsupervised. Deep learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks, and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.

ARCHITECTURE

The dataset contains 3000+ videos and photos. The dataset is then preprocessed. In the pre-processing step, face-only videos are split and put in another folder, which is only taken for training and modeling. The face-only video dataset is given to the model. In the end, the model defines that the given picture or video is real or fake as output as shown in figure 5.

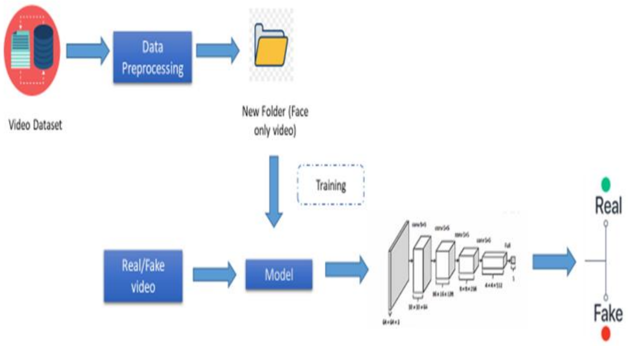


Fig 1. The architecture of DeepFake Detection

IV. IMPLEMENTATION

. PREPROCESSING

Data preprocessing is a critical stage in machine learning that improves the quality of the data to encourage the extraction of valuable insights from the data. Preparing (cleaning and arranging) raw data in order to make it acceptable for creating and training machine learning models is known as data preprocessing. In the pre-processing, the videos are converted into frames. And average frames per video are calculated. A new folder with Face-only videos is resulted from the pre-processing step.

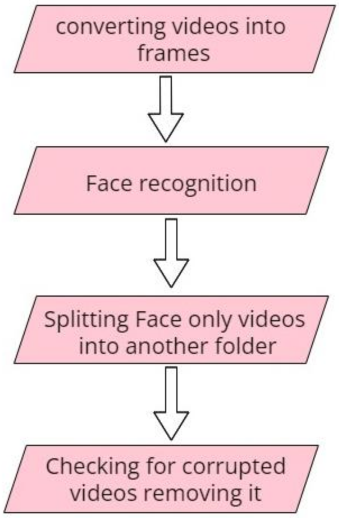
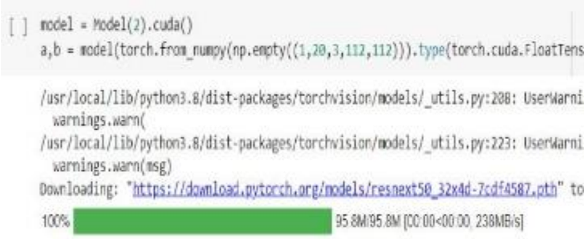

converting videos into frames
Face recognition
Splitting Face only videos into another folder
Checking for corrupted videos removing it

Fig 2. Pre-processing Steps

RESNET CNN FOR FEATURE EXTRACTION

Face-only videos are given as to the ResNet CNN model, a classifier used for extracting the features and accurately detecting the frame-level features. ResNet CNN model download is shown in Figure 3. Following this, 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. Using a trained model helps to reduce size and training difficulty



```
[ ] model = Model(2).cuda()
    a,b = model(torch.from_numpy(np.empty((1,20,3,112,112))).type(torch.cuda.FloatTens

    /usr/local/lib/python3.8/dist-packages/torchvision/models/_utils.py:208: UserWarni
      warnings.warn(
    /usr/local/lib/python3.8/dist-packages/torchvision/models/_utils.py:223: UserWarni
      warnings.warn(msg)
    Downloading: "https://download.pytorch.org/models/resnext50_32x4d-7cdf4587.pth" to
    100%                                   95.8M/95.8M [00:00<00:00, 238MB/s]
```

Fig 2. ResNet model download

LSTM FOR SEQUENTIAL PROCESSING

Long Short-Term Memory (LSTM) is a variety of Recurrent Neural Networks (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. LSTM processes the frames sequentially and then compare the features of the frame at different time.By comparing the frame with the segregated graph as in Figure 4, it depicts whether the video is deepfake or not. After training, any video can be passed to the model for prediction

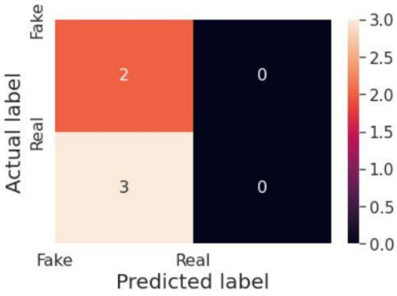
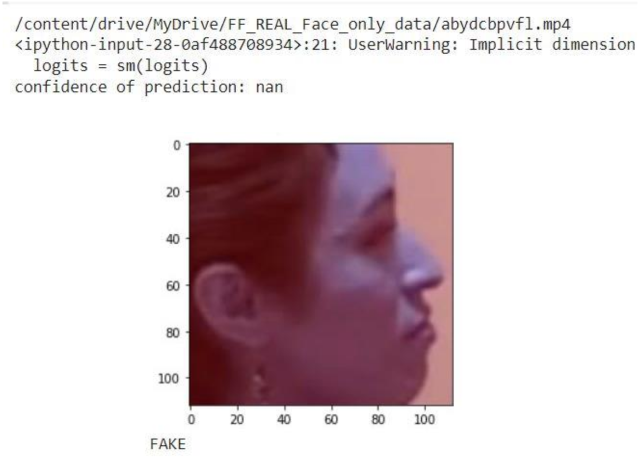


Fig 4. Labeling

RESULT

The output of the model is going to be whether the video is a deepfake or a real video along with the confidence of the model. One example is shown in figure 5.



```
/content/drive/MyDrive/FF_REAL_Face_only_data/abydcbpvfl.mp4
<ipython-input-28-0af488708934>:21: UserWarning: Implicit dimension
  logits = sm(logits)
confidence of prediction: nan
```

```
FAKE
```

Fig 5. Result

V.CONCLUSION AND FUTURE SCOPE

One of the critical disadvantages of current deepfake detection methods is the synthesized videos can be more realistic if they are accompanied by realistic voices, which combines video and audio synthesis together in one tool In the face of this, the overall running efficiency, detection accuracy, and more importantly, false positive rate, have to be improved for wide practical adoption.

The detection methods also need to be more robust to real-life post-processing steps, social media laundering, and counter-forensic technologies. There is a perpetual competition for technology, know-how, and skills between the forgery makers and digital media forensic researchers. The future will reckon the predictions we make in this work.

.

REFERENCES:

1. M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.
2. Maksutov, Artem A., et al. "Methods of Deepfake Detection Based on Machine Learning." 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) (2020): 408-411.
3. Malik, Asad & Kuribayashi, Minoru & Abdullahi, Sani & Khan, Ahmad. (2022). DeepFake Detection for Human Face Images and Videos: A Survey. IEEE Access. 10. 18757 - 18775. 10.1109/ACCESS.2022.3151186.
4. Ruben Tolosana, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection” Volume 64, 2020, Pages 131-148, ISSN 1566-2535
5. Lyu, Siwei. (2020). Deepfake Detection: Current Challenges and Next Steps. 1-6. 10.1109/ICMEW46912.2020.9105991
6. D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake Detection through Deep Learning," 2020 IEEE/ACM International Conference on Big Data Computing, Applications, and Technologies (BDCAT), Leicester, UK, 2020, pp. 134-143
7. Jung Tackhyun “DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern”, Department of Computer Science and Engineering, Konkuk University, Seoul, South Korea, no. 8, pp. 83144-83154, (2020)
8. Hoque, Mohammad & Ferdous, Md. Sadek & Khan, Mohsin & Tarkoma, Sasu. (2021). Real, Forged or Deep Fake? Enabling the Ground Truth on the Internet. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3131517.
9. D. A. . Sultan and L. M. . Ibrahim, “A Comprehensive Survey on Deepfake Detection Techniques”, Int J Intell Syst Appl Eng, vol. 10, no. 3s, pp. 189–202, Dec. 2022.
10. Haya R. Hasan And Khaled Salah “Combating Deepfake Videos Using Blockchain and Smart Contracts”, in IEEE Access,vol.7,pp.41596-41606,2019,doi: 10.1109/ACCESS.2019.2905689.