A Project Report

on

COMPARATIVE ANALYSIS OF DEEP LEARNING MODEL FOR DEEPFAKE DETECTION

Submitted in partial fulfillment of the

requirement for the award of the degree of

B. Tech (Hons.) with AI & ML

GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Under The Supervision of
Ms. Garima Pandey
(Ast. Prof)

Submitted By

Piyush Chandra
19SCSE1180047

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA

DECEMBER, 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the thesis/project/dissertation, entitled **"COMPARATIVE ANALYSIS OF DEEP LEARNING MODEL FOR DEEPFAKE DETECTION"** in partial fulfillment of the requirements for the award of the Bachelor of Technology Degree submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of September,2021 to December, 2021, under the supervision of Ms. Garima Pandey (Ast. Prof), Department of Computer Science and Engineering, of School of Computing Science and Engineering, Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Piyush Chandra 19SCSE1180047

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Ms. Garima Pandey
(Ast. Prof)

CERTIFICATE

The Final Project Viva-Voce examination of Piyush Chandra 19SCSE1180047 has been held on _____ and his/her work is recommended for the award of Bachelor of Technology.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date: December, 2021
Place: Greater Noida

Acknowledgement

I like to express my special thanks of gratitude to the mentor and the reviewers who gave me the utmost needed support. The paper could not have been completed without the co-operation of the group members.

This project has been costly in effort and time and data hunt. I am really thankful to my colleagues for their willingness and determination. I am also grateful to my institution for helping me put together this paper. The generosity and expertise of one and all have improved this study in innumerable ways and saved me from many errors; those that inevitably remain are entirely my own responsibility.

Abstract

Deepfake detection is the concept of distinguishing a computer manipulated graphic from a real recorded graphic. The technology used for this purpose is deep learning. Deep Learning is a sub branch of artificial intelligence. With technology becoming more readily available, deepfakes are also increasing in use in recent years.

It becomes evident that we need a system that detects deepfakes and prevents its use in suspicious activities. Development of a deepfake detection technology becomes evident to avoid the use of deepfakes in such activities. For this purpose, many tech giants have assimilated huge datasets which consist of videos that were made using deepfakes already available.

To detect a deepfake, requires an equally capable or even better algorithm and detection technique. Generative Adversarial Nets, GANs, is one such technique that might be able to rival other deepfake techniques.

This paper will discuss various methods to apply to detect deep fakes along with the process, libraries used, dataset liabilities and limitations, analysis and efficiency. Since Deep Learning technology is evolving each day with new innovations, this paper provides a comparative study about methods that have already been tested and their limitations with respective models and how it will be possible to make them more efficient.

Contents

CHAPTER 1.1 - Introduction

Fake images and videos can be easily found on the internet nowadays. They spread misinformation to the masses very easily. With progressing technology and its accessibility, there are different kinds of misinformation agents found online. One of these agents are deepfakes. Deepfakes are the product of artificial intelligence which produce videos that look almost identical to the naked human eye. They can impersonate any person, usually world leaders or celebrities, and spread it to stir an emotion among the masses. It has become very easy to influence people with such tactics. The danger posed by such technology is great. It can go as far as creating an entirely new personality and identity of a human with a never seen before face. That person can exist freely anywhere online and make false claims and affect the masses, i.e., exist as a social weapon. Weaponization of such technology can be used for infiltration and spying purposes on social media platforms and act as catalyst for such cybercrimes.

Anyone from an ordinary academic to a professional in Artificial Intelligence can make a deepfake just by following simple steps. Deepfakes can be made using online tools and mobile apps that are easily available online for a small fee or even free in some cases.

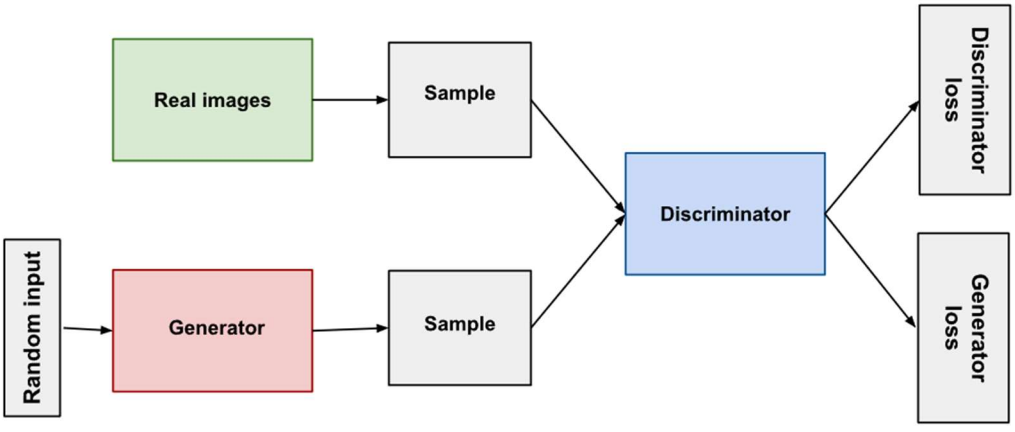
CHAPTER 1.2 - Problem Formulation

Deepfakes pose a great danger to society and various communities. One can simply change one's views towards someone and something if they are able to pose as someone or something credible and claim a false piece of information as true. Due to this reason, it is very important to realize that we need to have a program that can differentiate between a deepfake or a real source when human senses have failed. There have already been some cases of use of deepfakes that have caused controversies in the past. For instance:

Case I: Deepfake footage of American President Nixon's address to the nation about the Apollo 11 missions was released online. Most people who had already seen the address live on television were able to know that it was not a real video. But most new generation people had already believed that such a telecast had taken place. Eventually it was revealed that it was a video released by MIT to spread awareness about media misinformation.

Case II: Social media company had banned fake images and videos on its platform at the time of US elections to avoid people from getting misinformed about the presidential candidates. Deepfakes were being used to misinform people about the elections.

The proposed model uses Generative Adversarial Nets, GANs, are generative models. They create new data instances that resemble the training data.



Creation and detection of deepfakes has always been an area of research. Many papers have suggested ways to created and detect deepfakes. This paper will discuss some of them and analyse the use of one such technique in detection of deepfakes.

CHAPTER 1.3 - DEEPFAKE CREATION

A. Challenges

Deepfake creation is a rather challenging task. One must have access to reasonable hardware to be successful in creating a deepfake. While this challenge can be tackled by the use of cloud computing engines and platforms such as Google Cloud and Amazon Web Services which provide access to state-of-the-art hardware. Creating and detecting a deepfake, both tasks require a powerful mathematical processor. While the CPU can perform mathematical calculations just fine, it will take a considerable amount of time. To speed up this process, Cloud service providers provide Graphical Processing Units (GPUs) that are much well equipped to perform heavy mathematical calculations in a much shorter time. The biggest challenge that stands in the way is to choose a model that works efficiently and has a good accuracy. A good model choice is linked to detecting deepfakes with ease. It'll be very influential and beneficial to have a system that can avoid the weaponization of such technology. Unfortunately, with recent advances made towards its creation, deepfake detection is still lagging behind in innovation.

B. Techniques involved

Various techniques have been explored over the years. Some of the most influential and hugely dominant techniques in deepfakes projects are Autoencoder-Decoder pairs, GANs and CNNs. They have been used as go-to techniques for processing of images and creation or detection of Deepfakes. All the techniques implemented are just variations of those stated above.

CHAPTER 2.1 - PREVIOUS WORKS

Scientific as well as technological advancements have made it increasingly difficult to differentiate between a deepfake and a real video. Many celebrities and even world leaders have been victims of this technology. The type of manipulated media created by deepfakes usually fall into 3 categories[2].

1. Face-swap: the faces of two different people are swapped to make it look like that particular person is doing the actions of the person whose actions they truly are. It replaces the likeness of a person with someone else's likeness. Recreation of a new physical appearance of someone is the true objective.

2. Lip-sync: unlike the category discussed earlier, this means that instead of the replacing whole face of the person with someone else's, the image or video of a person is made to seem like that person is saying something they actually haven't. The lip movement can be replicated according to those words we want the person to say.

Fig. 1. Deepfake image created using PGGAN[5]

3. Puppeteering: the person in target is animated to mimic the movements of someone recorded doing those exact movements. It almost mirrors the person in front of the camera doing those actions.

A recurrent convolutional model (RCN) was proposed based on the integration of the

convolutional network DenseNet [3] and the gated recurrent unit cells to exploit temporal discrepancies across frames (see Fig. 4). The proposed method is tested on the FaceForensics++ dataset, which includes 1,000 videos and shows promising results.

In a model proposed by Guera and Delp[4], a temporal-aware pipeline method that uses CNN and long short-term memory (LSTM) to detect deepfake videos. CNN is employed to extract frame-level features, which are then fed into the LSTM to create a temporal sequence descriptor. A fully-connected network is finally used for classifying doctored videos from real ones based on the sequence descriptor

Deepfake videos are made in low resolution. The higher the resolution, the better hardware and more time are needed in creating a high-quality video. As explained in section II(A), a better GPU is required to process such a video. And quite similarly an equally, if not more, powerful unit would be needed to detect a deepfake that is produced in a good quality resolution.

A method to identify Deep Network Generated (DNG) fake images is proposed by Li et al.[22]. DNG fake images are in RGB color space with no explicit associations among the color components and there are some clear differences between fake images and real images in other color spaces such as HSV and YCbCr. Also, the DNG fake images are dissimilar from the real images while considering red, green, and blue components together. Hence this method analyzes the disparities in color components of images by separating image into R, G and B components and also transforming image into HSV and YCbCr color space. Then images in R, G, B, H, S, Cb are filtered using a high pass filter and the co-occurrence matrix is computed on each filter residuals. Finally, classifier is trained using a feature vector generated by concatenating the extracted co-occurrence matrixes. The GAN models used in this method for generating fake images are DCGAN, WGAN-GP and PGGAN and real image datasets considered are CelebFaces Attributes (CelebA) and Labeled Faces in the Wild (LFW). See Fig. 3.

Methodology	Datasets	Performance	Limitations
Differences in color components of deepfakes and real images are analyzed for detecting deepfake images (Li et al.,2018 [22])	Real image datasets: Celeb A, HQ-CelebA and LFW GANs used for generating Deepfake images: DCGAN, WGAN-GP and PGGAN.	Accuracy is > 98%	Method is not evaluated on any practical case scenarios of deepfake images.
Detection using Convolutional Neural Network (Do et al.,2018 [23])	Real image dataset: Celeb A GANs used for generating Deepfake images: DCGAN and PGGAN Evaluation dataset: Images from AI Challenge Contest	Accuracy: 80% and Area under the ROC Curve (AUROC) is 0.807	Performance is not evaluated on deepfake images generated by WGAN-GP, BEGAN etc.
Ensemble of neural network classifier(Tariq et al.,2018 [24])	Real image dataset: Celeb A GANs used for generating Deepfake images: PGGAN	Accuracy is 93.99% and 99.99% for small resolution images(64x64) and higher resolution images respectively	Performance is not evaluated on deepfake images generated by DCGAN, WGAN-GP, BEGAN etc.
Two methods based on (1) Color Image Forensics (2) Saturation based Forensics (McCloskey and Albright,2018 [25])	Method 1: Real image dataset: Celeb A GANs used for generating Deepfake images: PGGAN Method 2: Real image dataset: ImageNet dataset Deepfake images: LSUN dataset Evaluation dataset: GAN Crop image dataset and GAN Full image dataset of Standards and Technology's Media Forensics Challenge 2018	Method 1: AUROC 0.56 and 0.54 for GAN Crop image datasets and GAN Full image dataset respectively Method 2: AUROC 0.7 for both the evaluation datasets.	It is evident from AUROC that method gives comparatively a poor performance.
Designed a creating Computer Generated Face Identification (CGFace) model based on customized CNN (Dang et al.,2018 [26])	Real image dataset: Celeb A GANs used for generating Deepfake images: PGGAN and BEGAN	Accuracy:98% AUROC 0.81	Performance is not evaluated on deepfake images generated by DCGAN, WGAN-GP etc.

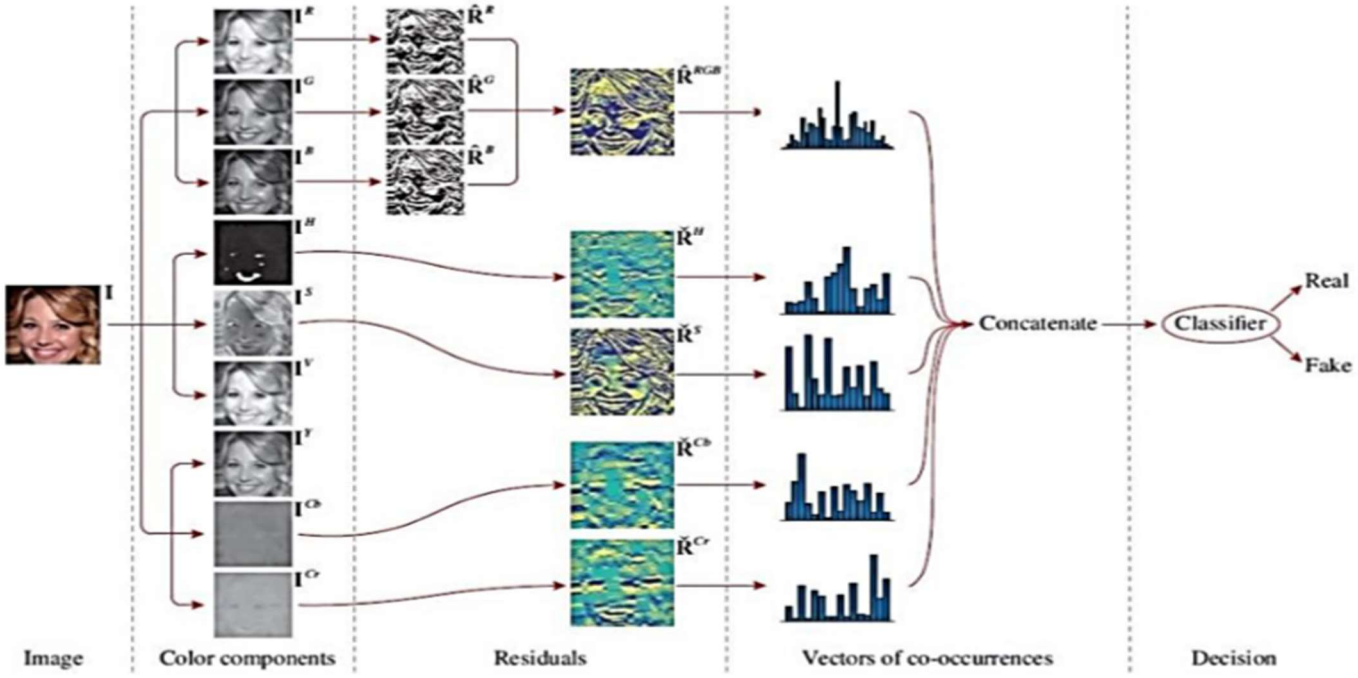


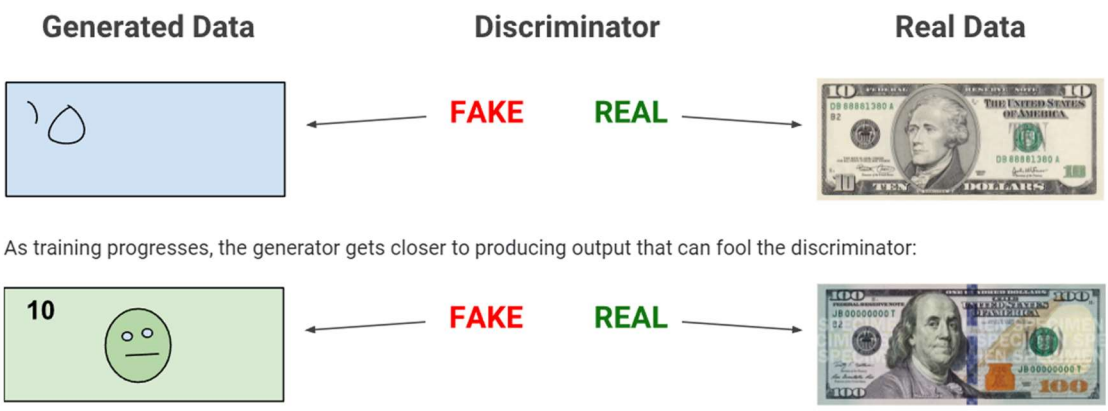
Fig. 3 Block Diagram of a method proposed in Ref. [6]

CHAPTER 3.1 - Proposed Model

GANs pair a generator, which learns to produce target output, with a discriminator, which learns to distinguish true data from the output of the generator.

- A generative adversarial network (GAN) has two parts:
- The **generator** learns to generate plausible data. The generated instances become negative training examples for the discriminator.
- The **discriminator** learns to distinguish the generator's fake data from real data. The discriminator penalizes the generator for producing implausible results.

The generator tries to fool the discriminator and the discriminator tries to keep from being fooled. When training begins, the generator produces obviously fake data, and the discriminator quickly learns to tell that it's fake:



The model will identify the facial features such as the smile and eyes and mark the face, eyes and smile of the face in the input video.

The workflow of the proposed model will be taking videos as input which are not labelled as real or fake. The trained deep learning model will use its discriminator to determine the video as fake or real and give the output. The video id of the input will be marked with real or fake in a csv file.

Generative Adversarial Nets consist of two neural networks i.e., a generator neural network(G) and a discriminator neural network (D).A general block diagram of GAN for generating fake images is shown in Fig.2. Generator takes some random noise (n) as input and attempts to produce fake images G(n) which are similar to real image dataset (x) whereas, the discriminator D aims to discriminate images generated by G from real images. The discriminator takes both real images and fake images as input and it estimates the probability of a sample coming from real image dataset rather than from fake images generated by the Generator. The discriminator will yield a probability value 1 when it is convinced an image is real and a 0 when it detects a fake image. The aim of discriminator is to maximize the number of times it correctly classifies the type of image it receives as input however the generator is trying to make the discriminator less correct. Thus, both networks are playing a game against each other, challenging to see who is superior at achieving their specific goal. So, discriminator is trained to maximize the probability that it properly discriminates images into real or fake, while generator is trained to minimize the probability that fake images generated by it are determined by discriminator as fake images, i.e., to minimize 1-D(G(n)). Thus, both the networks play a minimax game between them and it can be expressed mathematically as following value function as given in Equation (1) as:

$$\min G \max D\ V(D,G) = \mathbb{E}x{\sim}pdata(x)[logD(x)] + \mathbb{E}n{\sim}pn(n)[\log(1 - D(G(z))]$$

where $pdata(x)$ is data distribution of real images (x) and $pn(n)$ is noise(n) distribution. Once the necessary training is done, generator would be capable of producing natural and realistic looking fake images by using noise signals n, whereas the ability of D to differentiate deepfake images from real ones will also be improved.