Deepfakes Detection Challenge using AI and Machine Learning

Michael Leonardi, Ashutosh Misar,
Sivakumar G. Pillai, Avery Leider, Charles Tappert
Seidenberg School of Computer Science and Information Systems, Pace University
Pleasantville, NY 10570, USA
{ml76729p, am69489n ny00685p, aleider, ctappert}@pace.edu

***Abstract*—Facebook is funding research by delivering a deepfakes detection challenge to the worldwide community of 2.37 Billion monthly active users. This AI challenges the community to find solutions to fake or altered videos appearing on Facebook and other publications. These fake videos have become known as "deepfakes" and are increasing in frequency on social media websites as well as the Internet. This paper provides an overview of deepfake identification methods with the aim of possibly securing a grant for Pace University to fund research to better identify deepfakes. Pace University's team of strong data scientists has the capability of improving the methods currently used to identify deepfakes and ensuring that the worldwide community has a safe and healthy video ecosystem. This is a critical role that will be responsible for the measurement, detection and reduction of negative user experiences ranging from violence and adult content to evolving areas like misinformation and even fake news. In order to achieve the goal of creating better deepfake detection methods, we will look into the existing tools and attempt to combine several of them. This will serve as a framework to produce a more effective tool designed to differentiate both real and fake videos and minimize the spread of fake news.**

***Index Terms*—Machine Learning, AWS, Artificial Intelligence, Deepfake, Generative adaptive networks (GAN), Deep Convolutional Generative Adversarial Network (DCGAN), Concept drift**

I. Introduction

We have put together a Pace University Team to enter the Facebook Deepfake Challenge [1]. We see deepfakes and similar technologies as a new wave of cybersecurity threats, with the potential of affecting every digital audiovisual communication channel. Deepfakes are a growing problem and they affect the security and liberty of anyone connected to them. A deepfake is an AI-generated fake video which shows someone doing or saying fictitious things. This is achieved by changing the face of the actual person in the video and replacing it with a target of your choosing. Neural networks are then used to map the facial expressions of the target thus creating very realistic fake videos [2].

While this may sound like a harmless prank, with the use of social media a deepfake can quickly change the sense of reality for millions. Deepfakes could be used to create a fake emergency or terror attack, ruin a marriage with a video showing infidelity or even affect the upcoming Presidential election. [3]. The spread of deepfake videos has overwhelmingly affected women such as Hollywood actress Scarlet Johansson, whose face has been digitally inserted into pornographic videos viewed millions of times. Many other women who are not public figures have also been the victims of similar fake videos which continue to show up on social media websites. Even Mark Zuckerberg, the creator of

Fig. 1. example of a deepfake

Facebook was recently the victim of a deepfake video showing him saying outlandish things such as "Facebook owns their users" [1]. This among other high profile forgeries such as one of former President Barrack Obama has lead to a growing concern that fake videos are becoming more popular and are altering our sense of what is real. The idea that we can no

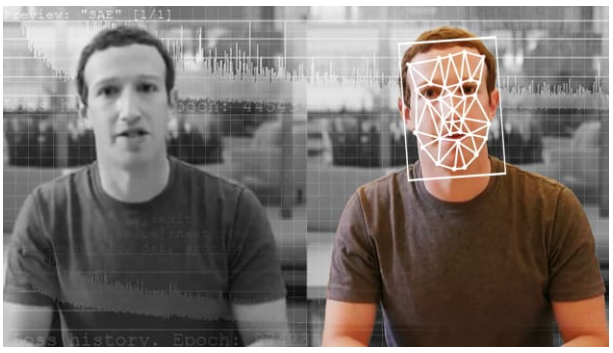

Fig. 2. Mark Zuckerberg deepfake

longer believe what we see and hear is a very disturbing thought and if we don't find better ways to identify if what we are seeing is authentic, we will be forced to do just that.

The tools that are being used to identify which videos are real and which are fake are not good enough. Detection of deep

fakes is also becoming more of a challenge as there are many open source software and apps available which can easily be used to create believable forgeries. This is exponentially increasing the number of fake images and videos that are being uploaded to social media and further creating a need for better detection [4]. This issue is being accelerated as over one third of all online activity is spent watching videos. Some statistics show that over 500 million hours of video are watched on YouTube every day and half a billion people watch videos on Facebook daily as well [1]. The majority of the content is being viewed on mobile devices and 92 percent of users share videos with other people. Even more disturbing is that most of the videos are viewed for less than 10 seconds overall and without any sound. This means that people are relying on what they see and aren't taking the time to listen or see if actions in a video are true or false. The same study shows that viewers retain 95 percent of the message received when they watch it in a video so the need to detect and stop these fakes is growing rapidly.

A current tool featured on the Facebook AI Challenge website for Deepfakes [1] works with the GAN model which uses two machine learning models as adversaries to create and then detect what is fake in a video. The generator which creates the fake, and then the discriminator which detects the fake, go back and forth creating and detecting higher quality fakes until the discriminator can no longer determine which is real and fake. This is a method of "unsupervised learning" and in a sense has succeeded in creating a blueprint for how to create better fakes but fails to maintain the ability to identify a fake once it is a high enough quality. [1]

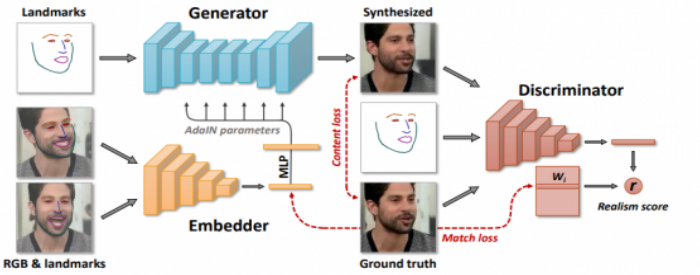

Fig. 3. GAN Model example

The challenge has now become to create tools which effectively distinguish deep fake videos from real ones before they can be distributed particularly via social media. A deep learning based method has been able to achieve some level of digital artifact detection when the face of the victim is warped onto the subject in a deepfake video. Distinct artifacts are left due to resolution inconsistency between warped face area and surrounding area. As such, these artifacts can be used to detect DeepFake Videos by comparing the generated face area and the surrounding area with a convoluted neural network [2].

II. Literature Review

A number of the current techniques for detecting deepfake videos are targeted towards face swapping videos as these account for the majority of the videos circulated online. They use frame-level binary classification problems which are based on physical/physiological aspects of the videos [5]. The physical/physiological method focuses on exploiting the observation that many deepfake videos lack normal eye blinking. This is due to the use of online portraits to generate the fake which normally don't have closed eyes. The subjects also tend to have incoherent head poses which are based off of the facial landmarks extracted from the real videos.

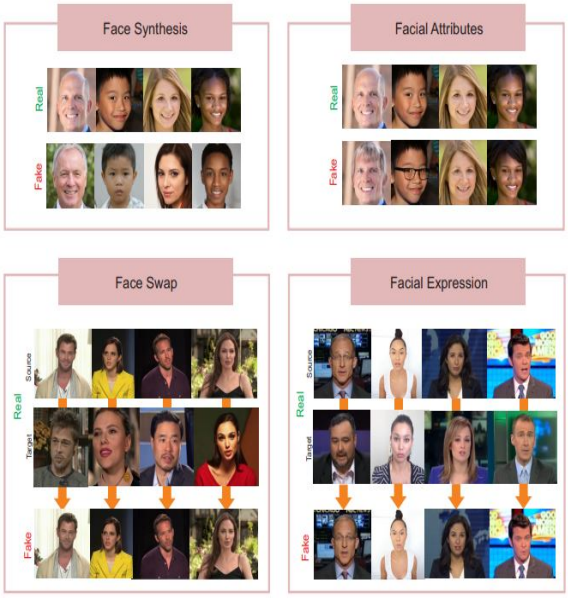

Fig. 4. Four Cluster Groups of Face Manipulations

• Face synthesis: this manipulation creates entire nonexistent faces, sometimes through powerful Generative Adversarial Networks (GAN). These techniques achieve astonishing results, generating high-quality facial pictures with a high level of realism. [6]

• Face swap: This alteration method consists of the substitution of the face of one person with the face of another person. There are two different approaches to this manipulation, the classical computer graphics-based techniques like FaceSwap7 , and the novel deep learning techniques known as DeepFakes8. Youtube is an example of where you can see this type of realistic video manipulation. [6]

• Facial attributes: this manipulation consists of modifying some attributes of the face like the colour of the hair or the skin, the gender, the age, adding glasses, etc .This manipulation method is sometimes dole out through GANs like the StarGAN approach proposed in .One example of this sort of manipulation is the popular FaceApp mobile application. [6]

• Facial expression: this manipulation consists of modifying the facial expression of the person, transferring the facial expression of one person to different person. One of the

foremost common techniques of face manipulation is called Face2Face and acts in real time. Recent approaches have shown its potential, producing high-quality videos of an individual (Obama) altering what he is really saying in a target video. [6]

Fig. 5. Generation of complex deepfakes using face manipulation methods

Generative adversarial networks (GAN) have been advancing and creating newer, higher quality fake videos. These neural networks have created a growing concern as they can quickly and easily generate believable deepfakes which detection tools have difficulty recognizing. [4]. The deep fake video is created with an input video of a specific individual who is the target and generates another video with the target's face replaced with another individual [2]. Coupling this with the GAN model which is trained to translate between faces of the target and the source and you can create a very realistic believable fake video. This has led to instances where fake news can be widely distributed over social networks and posing a significant challenge of detection.

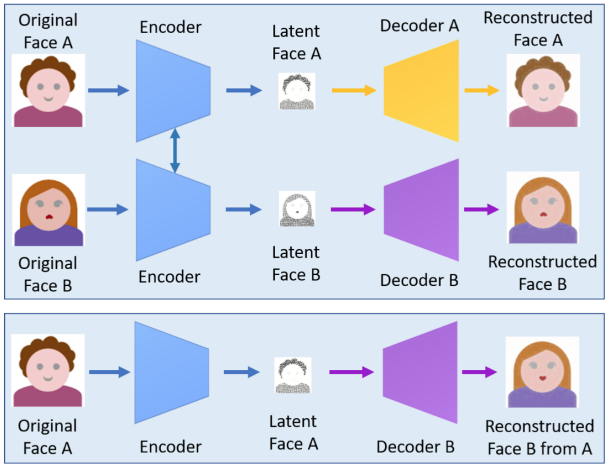

Fig. 6. A deepfake endcoder-decoder creation model using two pairs. Different decoders are used with the same network encoder

Researchers have now begun working on creating databases for the detection of fake videos by using GAN based face swapping algorithm. The authors of the deepfakes database [4] have taken data from a VidTIMIT database which includes 10 videos and audio recordings of 43 people. They then used this data to create 16 pairings where the subjects have similar facial features and swapped the faces of both subjects. Each pair were also used to train two GAN networks, a high quality image with and input/output image size of 128x128 and a low quality (LQ) input/output image size of 64x64.

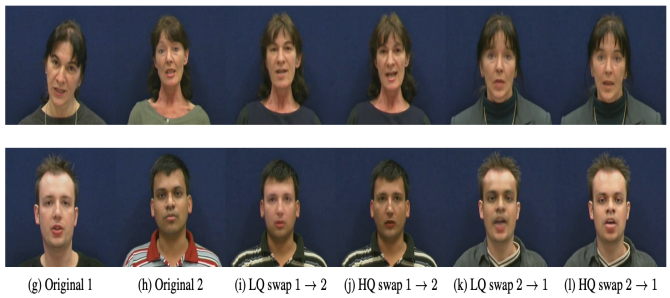

Fig. 7. Screenshot of the original videos from VidTIMIT database and low (LQ) and high quality (HQ) Deepfake videos.

Different blending techniques were used along with histogram normalization to adjust for lighting conditions when creating the videos. The result were extremely realistic deepfakes that effectively mimic facial expressions, blinking and mouth movements and thus wouldn't be detected with the current methods [4].

A second method of exposing deepfakes was recently created using deep learning to detect warping within the image [2]. This method is based on a limitation of computing resources where the deepfake algorithm can only synthesize images of a fixed size and require warping in order to fit the face of the source.

Warping consists of scaling, rotation and shearing of the image to make it match the poses of the targets face. This warping process creates digital artifacts that show resolution inconsistency in the area surrounding the face as a result of the compression step in fake videos. A Convoluted Neural Network (CNN), can be trained to detect the presence of such artifacts [2].

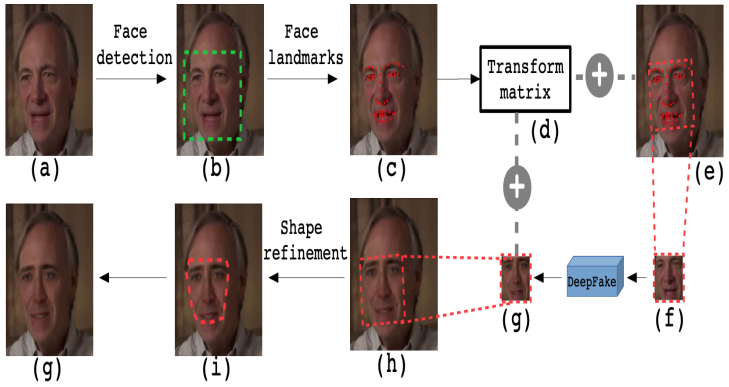

Fig. 8. Li, Lyu Deepfake Production Pipeline Overview: (a) Original image. (b)Detected face area. (c)Face landmarks. (d)Transform matrix is computed to warp face area (e) normal face area (f) Synthesized face generated from nerual network. (h) Synthesized face warped back to transform matrix. (i)Boundary smoothing of composite image.

Other protective measures have been studied recently to aid in the detection of deepfakes. Adding specifically designed

patterns known as the adversarial perturbations that are imperceptible to the human eyes but can help result in detection. High-quality AI face synthesis models require a large number of training images collected using automatic face detection methods known as face sets [5]. The adversarial perturbations can pollute a face set to have few actual faces and many non-faces lowering the quality of the training data obtained for the AI producing the fake. The larger the data set, the more images will be included which have the adversarial perturbations included disrupting the synthesis of the fake. The non-face images distort the view of the face in the image or video allowing for easier detection. This method could be used on video sharing platforms as a service to process images and videos before they were uploaded online [5].

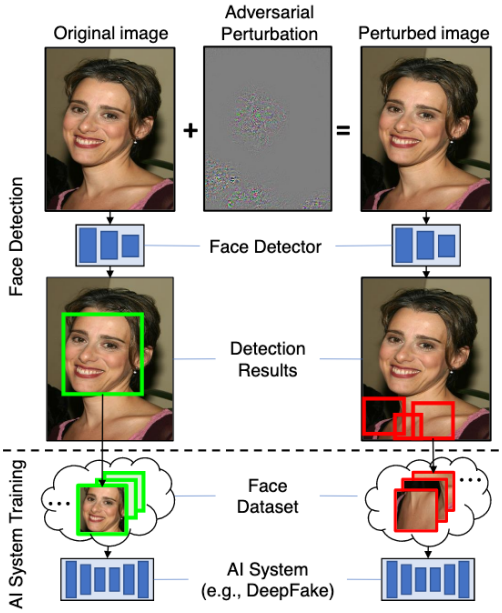

Fig. 9. Example of adversarial perturbations disrupting face detectors

Audio deepfakes have also improved in quality with the use of AI-generated synthesis. Combining the use of audio and video impersonating the target significantly make deepfake videos more convincing and dangerous [5]. Audio signals are different from video signals so they require different methods to detect fakes. Samples were gathered from a variety of devices capturing human speech translations to text. The text was then used to synthesize text-to-speech on the same devices. This was also done using a range of different human speaker profiles which increased the diversity of the synthesized voices [7].

A technique for distinguishing human speech from synthesized speech that leverages higher-order spectral correlations revealed by bi-spectral analysis was established. Many correlations are not present in most recorded human speech but are present in speech that is synthesised with several state of the art AI systems. The human speech had a glaring difference in bicoherent magnitude and phases then the synthesized speech. This was calculated by normalizing the magnitude and phase and characterizing them into four statistical moments. The moments are mean, variance, skewness and kurtosis which reduces the recordings to an 8-D feature vector [7]. The research showed that the artifacts were a result of the fundamental properties of the synthesis process which means they would be difficult to eliminate as a countermeasure. GAN model synthesizes speech in using a different mechanism but the bispectral statistics are still present. This will allow the bispectral analysis to be used to identify synthesized audio until the possible next evolution of GAN networks.

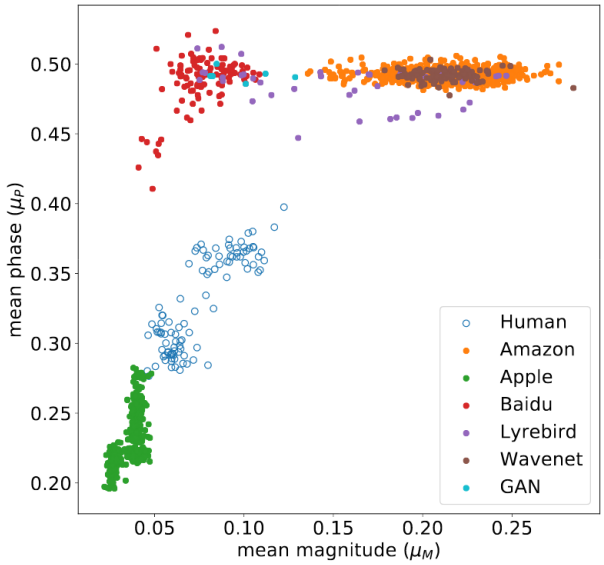

Fig. 10. Biocoherence Magnitude and Phase.The open blue circles respond to human speech and all other circles respond to synthesized speech

III. Methodology

The Facebook DeepFakes challenge has provided data sets and benchmarks which have helped to speed up the progress of AI. The goal of the challenge is to product tools that can be used to effectively identify fake videos and the legitimacy of information that is being presented online [1]. Facebook offers a Github with 217 code repositories and allows the challenge participants to create deepfakes to test with a variety of data sets. These datasets can be analyzed using the python anaconda platform and are mostly broken up into Jupyter notebooks to help setup controlled environments.

Pytorch is an open source deep learning platform that uses tensors and allows the usage of a GPU to provide maximum flexibility and speed in computing. This allows the user to build and train a small neural network that can classify images. We will also be using this tool to help analyze some of the datasets provided by Facebook.

The AI technologies that power deep fakes and other tampered media are rapidly evolving, making deep fakes so hard to detect that, at times, even human evaluators can't reliably tell the difference. The Deep fake Detection challenge is designed to incentivize rapid progress in this area by inventing new ways of detecting and preventing manipulated media.

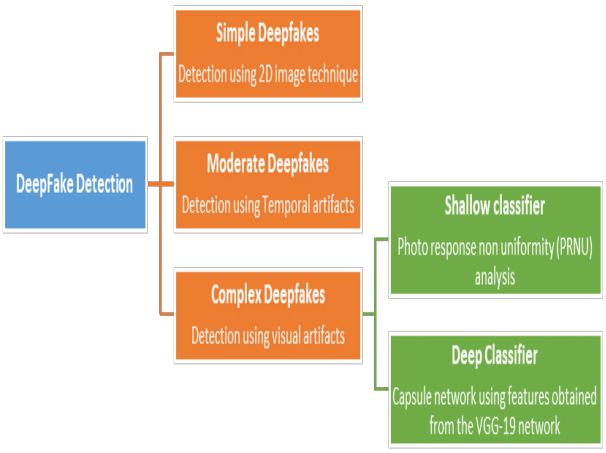

Fig. 11. Hieracrhy of Deepfake Detection techniques

Technical environment:

Four stages for preparing a machine learning model:

1)Preprocessing input data

2)Training the deep learning model

3)Storing the trained deep learning model

4)Deployment of the model

Among these stages, training the machine learning model is the most computationally intensive task. We are combining multiple analysis methods so we can't depend on a particular technological environment. For instance, in our first method we used the kaggle platform for creating our model. Kaggle kernels allow us to build, train, and test up to a certain extent. It provides all the computational power and space required for data in the online/cloud platform but it requires an internet connection. The other methods general technological environment is as follows:

o Quad core Intel Core i7 Skylake processor or higher (Dual core is not the best for all kind of methods/scenarios, but manageable)(CPU Notebook ¡= 9 hours run-time).

o 16GB of RAM (8GB is okay but for tenser flow, cnn/ rnn models execution but, it could extensively add more time to train the model especially with a large data set).

o M.2 PCIe or regular PCIe SSD with at least 256GB of storage, though 512GB is best for performance. The faster you can load and save your applications, the better the system will perform. (SATA III will get in the way of the system's performance).

o Premium graphics cards, GTX 980 or 980Ms would be the best for a laptop, and GTX 1080s or GTX 1070s would be the best for the desktop setup . (GPU Notebook ¡= 9 hours run-time).

Datasets:

We are using a dataset provided by the Deep fake Detection Challenge which is over 470 GB [1]. It similarly has 50 smaller files, each 10 GB in size (small chunks) where the data is comprised of .mp4 files, split into compressed sets. A metadata.JSon accompanies each set of .mp4 files, and contains filename, label (REAL/FAKE), original and split columns, listed under Columns which are described as follows:

Metadata Columns:

o filename - the filename of the video

o label - whether the video is REAL or FAKE

o original - the initial video is listed here

o split - this can be always capable "train"

We have two separate data sets as follows:

1. train sample videos.zip - a zip file containing a sample set of coaching videos and a metadata.json with labels. A deepfake might be either a face or voice swap (or both) within the training data, and this can be denoted by the string "REAL" or "FAKE" within the label column. we should always train our model using local or cloud resources. (for training model)

2. test videos.zip - a zip file containing a small set of videos to be used as a public validation set. we are going to be predicting the probability whether or not a specific video could be a deep fake with the use of trained model. (for predicting deep fake video).

We are going to use four combined strategies to help identify the deepfake. We grouped the detection methods into three major categories: simple deepfakes, moderate deepfakes and complex deepfakes.

In simple Deepfakes we are comparing 2D image face point and then locating a face landmark within the picture. In order to compare the 2D image face point, we are going to capture the first frame of the face in the video file and locate face within the picture with the assistance of a face recognition package taken from the Github. Once the face

Fig. 12. Example of the face being extracted in the video frame

has been located, the face recognition package will identify specific face landmarks and loop through all of the frames of the video marking the landmarks in each frame. There will be some cases where a face could not be found and in such a scenario, that frame will be eliminated. [8]

This data will be appended to a list called the frame list and random frames will be selected and run through an identifier such as the random forest algorithm. This will create our dataset of real and fake videos. The dataset can then be used to train the detection model to differentiate the images in the fake video from the images in the real video and then tested versus the data provided in the deepfakes detection challenge. This system prediction model will work to detect and identify moderate level deep fakes such as the type created by various

open source applications. This model however will not be able to detect GAN deep learning generated deepfakes as they are using a neural network which can effectively hide the distinguishing features during deepfake creation. The results of the training data that was used show 19.25 percent were REAL videos and 80.75 percent were FAKE videos when tested.

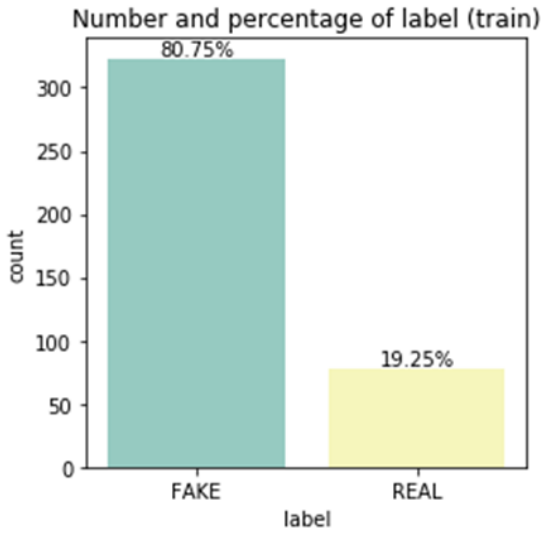

Fig. 13. Distribution of data over the training set

Generally moderate deepfakes can be defined as those deepfakes which uses two or more face manipulation techniques. During the process of compressing the videos, the frame data gets adversely degraded to which it is impossible to use most of the image recognition algorithms. Apart from that, it is inefficient to implement the algorithms which are designed to identify only still fake images due to the temporal characteristics of the videos.

In this section, we will focus on "DeepFake" video detection algorithms by classifying them below:

- Video detection Algorithms using temporal features
- Video detection Algorithms using visual artifacts within frames

Video detection Algorithms using temporal features across Video Frames: The analysis of the gathered observations pinpoints that, during the synthesis process of Deep Fakes, temporal coherence was not implemented accurately.

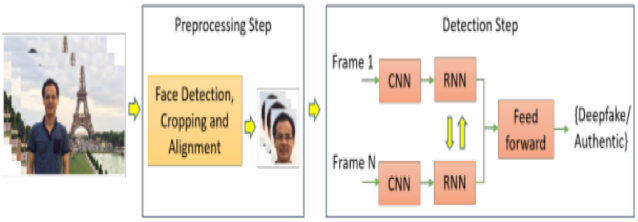

Fig. 14. Deepfake Detection with the help of a CNN or RNN

We perform video manipulations on a frame by frame basis in order to ensure that the lower level artifacts transform themselves into temporal artifacts with discrepancies across various frames.

Steps for face manipulation detection:

1.Pre-processing stage in which the objective is to identify, manipulate and align faces on a sequence of frames

2.Distinguishing the manipulated and the authentic facial images with the help of a combination of recurrent neural networks (RNN) along with convolutional neural network (CNN). As there exist temporal inconsistencies and intra-frame inconsistencies between frames of such deepFake videos, it is appropriate to utilize the temporal-aware pipeline algorithm which implements long short term memory (LSTM) and CNN to identify fake content.

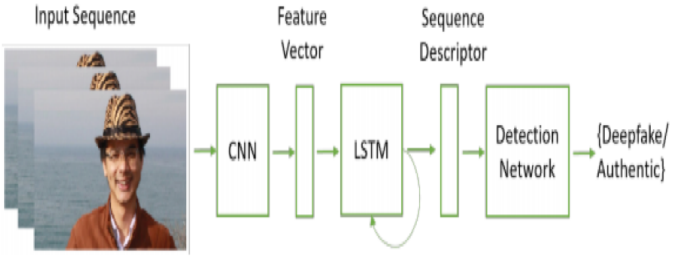

Fig. 15. Deepfake Detection with LSTM and CNN

In order to extract temporal features from our desired video sequence, our deepfake detection algorithm utilizes long short term memory (LSTM), CNN and a sequence descriptor for representational purposes. In order to calculate the probabilities of the frame sequence, the sequence descriptor acts as our input, which is collected with the help of a detection network consisting of fully-connected layers. Once completed, the authentic or deepfake classification of the frames is done for identification purposes.

Visual Artifacts within Video Frame

In order to analyze a video and collect its discriminant features, the videos are generally decomposed frame by frame, and analysis of visual artifacts within each frame takes place. It is then further classified as shallow or deep, which makes it possible for us to differentiate the authentic video data from fake videos. [9]

Deep classifiers

In order to match the configuration of the authentic videos, DeepFake videos are generally generated with the limited resolution, which is achieved via using affine face wrapping algorithms like scaling, shearing, rotating, etc. CNN models like ResNet50, ResNet152, ResNet101, or VGG16 are used to detect the artifacts, which are the result of the resolution inconsistency among the wrapped face and the surrounding context. [10]

In order to differentiate the authentic videos from the fake videos, features of the VGG-19 network are utilized by the capsule network. Before the VGG-19 network can be

implemented to extract latent features for the capsule network, the pre-processing step needs to identify the facial regions in the frames and then scale them to the size of 128x128. The capsule network contains two output capsules along with three primary ones in order to handle real and fake images. Statistical pooling provides the forgery identification capability making it a crucial part of the capsule network.

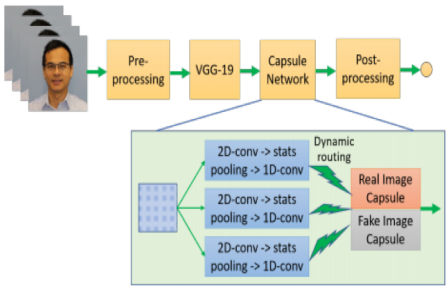

Fig. 16. Deepfake Detection with Capsule Network VGG-19

Shallow classifiers

In order to isolate the culprit's face, we extract the video frame by frame and crop them as required. These cropped video frames are then distributed sequentially and evenly in eight groups. For each group, we then calculate an average frame photo response non-uniformity (PRNU) pattern. Normalized cross-correlation points are calculated by comparing the PRNU pattern of each group with the PRNU patterns of the remaining seven groups. [9]. Finally, in order to contain the culprit's face in the video, we extract the frames from the video and crop them identically to the same pixels. DeepFake identification algorithms generally depend on the artifacts or the inconsistency of natural features among the authentic and fake videos or images. The photo response non-uniformity (PRNU) analysis can help us to identify the deepFake images or videos from the authentic data. A factory

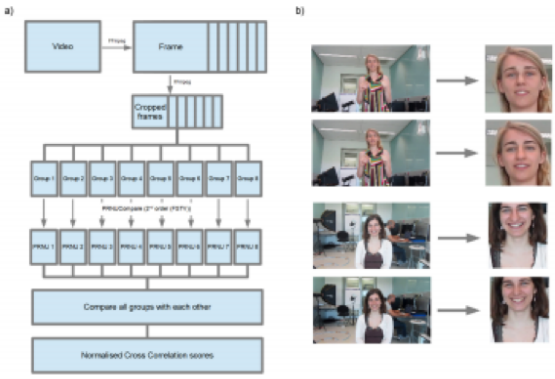

Fig. 17. Detection with PRNU Patterns

anomaly of light-sensitive sensors in digital cameras causes PRNU noise patterns. This is a popular way to differentiate the digital images as every digital camera has a unique PRNU noise pattern. It also acts as a fingerprint of images [10]. As the swapped face is supposed to modify the original PRNU pattern in the face area of video frames, digital forensics takes enormous advantage of that in order to perform their analysis. We sequentially divide the frames into eight classes of identical size and an average PRNU pattern is generated for each class using the second-order (FSTV) method [Baar et al., 2012] with the tool 'PRNUCompare'. Conversion of the videos into frames takes place in order to crop the facial area, which needs evaluation. The average PRNU pattern is calculated after the cropped frames are separated sequentially into eight groups [9]. The PRNU patterns are compared to one another, and normalized cross-correlation scores are then returned. The variations in correlation scores are compared, and the average correlation score for each video is then calculated. This can now be used as a baseline to determine which videos are suspected to be fake.

IV. Conclusions

As we continue to expand in the multimedia system age, having data with integrity is crucial to our lives. Given the recent developments in producing manipulated data at scale,(text, images, videos, and audio) we want the complete involvement of the analysis community to develop strategies and systems to help counteract the threat. These strategies and systems will notice and mitigate the side effects of manipulated transmission helping us identify if what we are viewing is authentic. Technology to control pictures is advancing quicker than our ability to determine what is real from what has been altered so a task as massive as this won't be resolved by one person alone or with any single method.

In this study, we focused on the influence of using different methodologies in the detection of deepfake videos and proposing a homogenous benchmark for follow up work. Most systems find the face manipulation detection task simple to perform due to the GAN "fingerprints" contained within the fake images however, no single technology is able to detect a deepfake on its own. What if we are able to remove those fingerprints? Most methodologies for fake media detection are dedicated on controlled setups with training and testing detection systems considering the same at the image compression level. This tactic appears to be less suitable for real scenarios. New methods will need to be developed to identify such images and video variations as there is a high degradation of the fake detection performance, particularly in random scenarios.

In order to create better tools to detect deepfake videos several of the methods above will need to be combined to provide better analysis. GAN and DCGAN network created deepfakes employ deep learning to create very realistic images that can mimic expression, blinking and movement of the target. In order to detect the GAN or DCGAN created deepfakes, a similar GAN must be created and trained to both create and detect fake videos. During this process, the combination of various methods can be programmed and tested using the provided github data from Facebook. The larger the

dataset used for this process, the higher quality deepfakes can be created and analyzed. This is the next step in the process for Pace University to further explore the deepfakes detection challenge as training a GAN or DCGAN network takes considerable time.

Having a testing mechanism that runs several detection methods simultaneously would be ideal. The overall goal would be to have a filter of sorts that can run tests on all videos being uploaded to social media and major websites to determine the authenticity of videos. Combining the bi-spectral analysis of audio along with face warping artifacts within videos are examples of two phases of the mechanism which would begin a step by step process to help eliminate fake videos being posted. The next phase of this work will need to utilized these combined techniques in order to begin developing a suitable mechanism to identify deepfake videos.

REFERENCES

[1] zuckerberg, "Deepfake challenge," April 2019. [Online]. Available: https://ai.facebook.com/blog/deepfake-detection-challenge/

[2] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[3] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[4] P. Korshunov and M. Sebastien, "Vulnerability assessment and detection of deepfake videos," in *The 12th IAPR International Conference on Biometrics (ICB)*, 2019, pp. 1–6.

[5] S. Lyu, "Deepfake detection: Current challenges and next steps," *arXiv preprint arXiv:2003.09234*, 2020.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1–11.

[7] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis," in *CVPR Workshops*, 2019.

[8] Kaggle.com, "a quick look at the first from of each video," Jan 2020. [Online]. Available: https://www.kaggle.com/brassmonkey381/a-quick-look-at-the-first-frame-of-each-video

[9] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)*, 2018, pp. 133–136.

[10] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2307–2311.