Deepfakes Detection Challenge using AI and Machine Learning

Nikhil Dikshit, Tonya Fields, Ning Yang , Michael Leonardi, Sivakumar Govindarajan
, Ashutosh Misar, Avery Leider, Charles Tappert
Seidenberg School of Computer Science and Information Systems, Pace University
Pleasantville, NY 10570, USA
{nd20961n, tf12345n, ny00685p, am69489n, ml76729p, aleider, ctappert}@pace.edu

***Abstract*—Facebook is funding research by delivering to the worldwide community [1] of 2.37 Billion Monthly Active Users a deepfakes detection challenge. This AI challenges the community to find solutions to fake or altered videos that appearing on Facebook and other publications. This paper is written to get a grant for Pace University from Facebook to fund research to better identify deepfakes. Pace University's team of strong data scientists will be responsible for improving the methods we use to identify deepfakes and ensuring that we are developing 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. By using the algorithms and machine learning tests that Siva and Nikhil previously worked on in their earlier research on mortgage credit data, and using it instead on deepfakes data, and using great ideas from AMS, we can find great wealth quickly.**

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

I. Introduction

Here we have put together a Pace University Team to enter the Facebook Deepfake Challenge [2]. 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. [3]. 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 [4]. 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. An example of the power of this technology can be seen in the NBC video which shows examples of videos being altered [**?**]. The video shows Jake Ward of NBC news having Richard Nixon's face and voice while altering the famous resignation speech. Former President Barrack Obama is shown saying some strange things but is actually being represented by comedian Jordan Peele in an even more believable fake. The idea that we can no 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.

Currently 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 deepfakes. 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 [5].

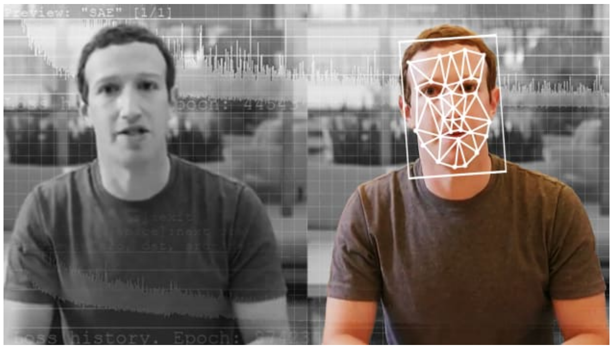

Fig. 1. example of a deepfake

A current tool featured on the Facebook AI Challenge website for Deepfakes [2] 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. [2]

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. This warping leaves distinct artifacts due to the resolution inconsistency between warped face area and surrounding context. 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 [4].

II. Literature Review

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 simply by downloading an app on your phone. This has led to instances where fake news can be widely distributed over social networks and posing a significant challenge of detection [5]. 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 [5] 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.

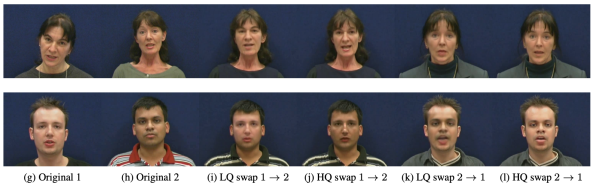
(g) Original 1 (h) Original 2 (i) LQ swap 1 → 2 (j) HQ swap 1 → 2 (k) LQ swap 2 → 1 (l) HQ swap 2 → 1

Fig. 2. 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 [5].

A second method of exposing deepfakes was recently created using deep learning to detect warping within the image [4]. 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 [4].

III. Project Requirements

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 [2]. 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 anaconda platform and are mostly broken up into Jupyter notebooks to help setup controlled environments.

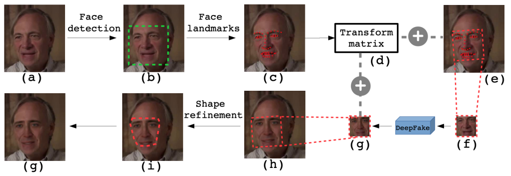

Figure 1. *Overview of the DeepFake production pipeline. (a) An image of the source. (b) Green box is the detected face area. (c) Red points are face landmarks. (d) Transform matrix is computed to warp face area in (e) to the normalized region (f). (g) Synthesized face image from the neural network. (h) Synthesized face warped back using the same transform matrix. (i) Post-processing including boundary smoothing applied to the composite image. (g) The final synthesized image.*

Fig. 3. Li, Lyu Deepfake Production Pipeline

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.

References

[1] A. Hutchinson, "Facebook reaches 238 billion users," April 2019.

[2] zuckerberg, "Deepfake challenge," April 2019.

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

[4] 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.

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