Deepfake Detection and Classification

Sam Koenke
Department of Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA
sskoenke@wpi.edu

Revant Mahajan
Department of Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA
rmahajan@wpi.edu

Pratik Jawahar
Department of Robotics Engineering
Worcester Polytechnic Institute
Worcester, MA
pjawahar@wpi.edu

***Abstract*—Recent advancement in computer vision techniques have led to a surge in deep fake videos and images flooding. Deep fake is a term coined to represent fake images generated using deep neural networks. These videos and images can have a profound impact on world politics, stock market and personal life. With this paper, our team summarizes past work in this field and then focuses on creating a network architecture to classify these deep fakes.A CNN architecture is implemented and trained and many insights into the problem of deep fake classification are discussed.**

***Index Terms*—Deepfakes, Deep Learning, Fake Videos, Deepfake Detection,**

I. Introduction

Social engineering is the planned manipulation of societal behavior, and has been around for as long as people have been able to communicate with each other. As the information age has progressed, it has become easier and easier to reach a larger audience. From the invention of writing, the wide spread use of printing, and now to a modern, digitally connected society; the ability to spread information has become easy and nearly instantaneous. With these changes in the rapid availability of information, one must now, more than ever, have a way to figure out what information is true and what is false.

Reliability theory states that the three most important aspects of information are its confidentiality, its availability and its integrity. Social engineering typically uses information that is released publicly (i.e non-confidential) and highly available through digital means such as the internet and social media. It is the difficult task of determining the integrity of the data that is often left up to the individuals who come in contact with the information.

With so many organizations trying to manipulate social perception and behavior, quite a bit of pressure is being put onto tech companies to minimize the amount of fake information that is being shared via their platforms. While this is certainly an aid to the minimizing the problem, the individuals who come in contact with the information still have a responsibility to determine what they will believe. Often individuals will base their believing of the integrity of information off of the source of the information and the people who provide it. For video and audio content, often this is done based off the visual or audible recognition of someone in the media, but what if you can no longer use your vision or hearing to validate the information? What basis do you now have to verify that piece of information is accurate? It is this vulnerability that deepfakes, a modern social engineering tool, seek to exploit.

Deepfakes, use deep learning to digitally create media content that appears to both look and sound like authentic media. The media created by these deepfakes are sometimes easily detectable, but sometimes good deepfakes are virtually indistinguishable from the real thing.

The potential effects of deepfakes are quite serious. Political manipulation has emerged as one of the most serious types of social engineering. A convincing deepfake video could do damage to political candidates or cause opposing countries to strike one another with military action. In a 2018 speech at the Heritage Foundation, Senator Ricky Rubio gave an indication of how serious of a threat deepfake technology is when he said, "In the old days, if you wanted to threaten the United States, you needed 10 aircraft carriers, and nuclear weapons, and long-range missiles. . . . Today, you just need access to our internet system, to our banking system, to our electrical grid and infrastructure, and increasingly, all you need is the ability to produce a very realistic fake video that could undermine our elections, that could throw our country into tremendous crisis internally and weaken us deeply." [4]

Other uses of deepfake technology are almost endless. Fake evidence generated for use in the judicial system could lead to incorrect verdicts. Fake media containing non-consensual sexual content could be quite damaging to the personal life and career of individuals. Fake business information could be used to manipulate the stock market.

The problem of social engineering through the use of deepfakes is serious. This paper investigates methods of using deep learning to classify potential deepfake content as real or fake.

II. Background

The recent developments in computer vision and image processing algorithms have turned out to be a double-edged sword, bringing about remarkable augmentation in machine capabilities but concomitantly raising questions of morality as well. They have made significant contributions in a multitude of fields such as cancer detection, autonomous vehicles, augmented reality and state of the art surveillance systems. However, they have also become tools for generating fake digital media to a strikingly realistic degree which poses a great threat on various levels. A new term has been coined to represent such morphed videos and images, Deepfakes.

A. What are Deepfakes

Deepfakes are artificially synthesized digital media, primarily in the form of images and videos, using Machine Learning and Artificial Intelligence. A deep fake may refer to a pre-recorded digital medium (video, image, audio, etc.) of a real person that is morphed either partially or completely, to synthesize an artificial subject that borrows features from the human subject, or a completely artificial digital medium, where every feature is autonomously computer generated with no reference medium. Though the artificial subject has realistic features, the probability of all its features matching those of a single real subject is very low, per the canonical definition. However, the artificial subject may be biased to imbibe features from one or more human subjects based on how the learning algorithm is trained.

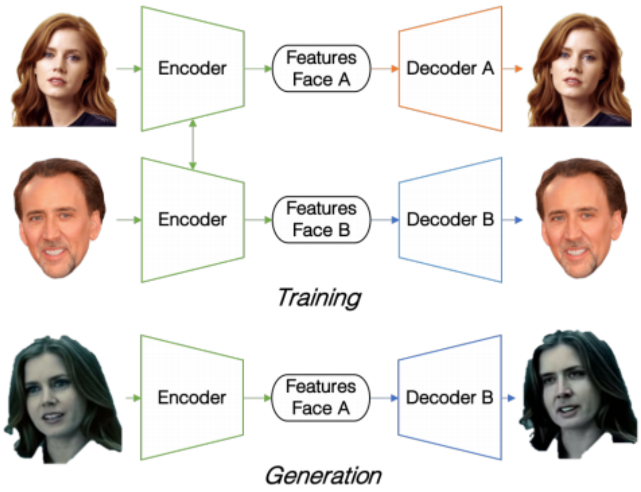

Fig. 1. Fake Images using Auto Enoders

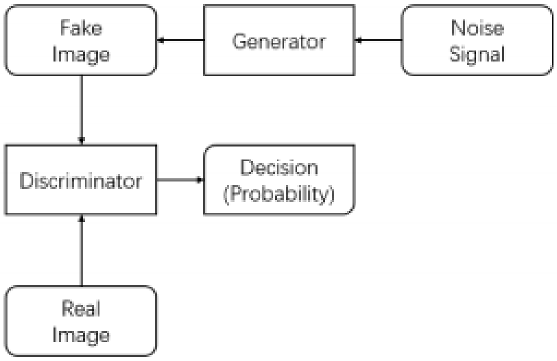

Fig. 2. Fake Images using GAN

B. How are Deepfakes generated

- Auto-encoders are known for dimensionality reduction and compact representations of images. They are able to form a compressed representation of images with a minimized loss function [5][8]. An Encoder Decoder network is used to generate fake images from two different real images. A common encoder is trained to form a compressed representation of the input images. For each image class, a different decoder is trained to get back the original image, or as close as possible from this compressed representation. After the training is done, an input image of one of the known classes is fed into the encoder and then to a decoder, of a different class, to generate the fake image.

- Generative Adversarial Networks are another method by which deep fakes can be created, through two neural networks competing against each other[12]. The first network is called a generator, and it tries to generate fake images. The second neural network is called a discriminator, which tries to classify if an image is fake or not. The feedback is fed back to the generator so it can be trained to produce images, that can fool the discriminator.

C. Detecting Deepfakes

Since the ethical implications of deepfake algorithms gained priority, multiple methods of detecting and classifying a given video as being legitimate or fake have come up. The initial approaches of detection that used weak and imperfect morphing features as indicator flags, have however been countered with improved and more realistic deepfake algorithms, making detection of deepfakes a daunting task. Some methods employed previously include the following methods.

- The use of 'softbiometric signatures' has been one of the most primitive techniques used to identify deepfakes. This method involves studying intrapersonal features of the original subject portrayed in the video and forming a class of implicitly distinct stochastic features such as head motion, facial gestures and characteristics while saying a particular phrase or while portraying a particular emotion. This data is used to train the learning algorithm and is then tested against videos of comedic impersonators of the subject and against intrinsically generated Deepfakes. A one-class SVM model is then used to distinguish the set of features from the real subject and those from an impersonator or a deepfake [1]. This method achieved 92% detection accuracy however, generating the softbiometric signatures is a highly complex and personalized process which means each subject must have his own training dataset. Generating such a training dataset is a very arduous task compared to generating a deepfake, meaning this method is quite obsolete with respect to the speed at which deepfakes are being developed.
- Semantic inconsistencies and weakly morphed features are another common and effective means of detecting

deepfakes. This technique exploits the occasional imperfections in image morphing that arise due to excessive movement of the subjects features, or due to swiftly changing image characteristics such as ambient lighting or momentary drops in pixel resolutions. Under these circumstances, the image morphing algorithm cannot find exact matches of facial features in order to morph the subjects features accurately, resulting in errors such as a double chin or a double layered jaw line for that given set of frames. Such frames may also be prone to semantic inconsistencies in the Neural Network that was used to create the deepfake, such as mismatched earrings [14], which are blatant indicators of a fake video.

- Intrapersonal feature analysis such as, analysis of eyes of the subject in a deepfake video has been the most successful and versatile method of detection thus far. This turns out to be an implicit short coming of GAN generated fake videos, where it was noticed that GAN generated video subjects rarely ever blinked [6]. It was also noticed that the eye movement did not accurately match the head pose for a GAN generated video. This is primarily because the GAN used to generate the video is trained using still images of the subject and as it goes with most portrait images of people, the eyes are rarely ever shut. Other intrapersonal features such as head pose, movement, color of the eyes etc. have been effective flags that are significantly difficult for GANs to forge, thereby serve as good indicators for detecting deepfakes [13] [7].
- Speech and audio analysis proves to be yet another effective tool, that strays away from any of the methods mentioned above. Studying audio quality and looking for semantic details in speech are the two main avenues for determining flags that might predict if an audio stream is real or forged. These two parameters however, are not fully incongruous when compared to how image features are analyzed to detect fakes. Looking for abrupt, unnatural changes in audio quality, ambient noise that doesn't match the setting of the video and semantic errors such as unnatural pauses, use of uncommon phrases and the sync between the video and audio feeds are being studied as potential indicators of deepfakes [2].

III. The Database

The data sets used for this project are sourced from the FaceForensics++ Database [cite:FF++], created and provided by Google JigSaw. This data set is created from a selection of 1000 videos containing 509, 914 sourced from www.YouTube.com to imitate real scenarios. These are then manipulated or forged by two computer graphics based methods (*Face2Face and FaceSwap*) and two Neural Network based (*DeepFakes and NeuralTextures*) approaches[11].

A. FaceSwap

FaceSwap is a graphics based approach to transfer the face region from a source video to the target. The face region is detected based on some sparse landmarks. These regions are then used to morph the source features with a 3D template model using shape blending techniques. This model is then fed back onto the target image and a correlation is formed to minimize the difference between the model and the extracted landmarks using the textures of the input image. This is performed for all source and target pairs for all frames in the video.[11]

B. DeepFakes

DeepFakes have become widely popular in recent times as a face morphing or swapping framework and are being studied extensively to develop deep learning algorithms. The implementation used in the FaceForensics++ data set is the *faceswap github*. In this method, two auto encoders and a shared encoder are used to reconstruct features of source and target frames respectively. The images are then cropped and aligned before blending the source features to the target image using Poisson image editing.[11]

C. Face2Face

This is a computer graphics based re-enactment system that projects the source expressions while maintaining features of the target image. This is done by first selecting keyframes manually. These frames are then used to generate a dense reconstruction of the face which enables easier simulation of the face under different illumination conditions or with changed expressions.[11]

D. NeuralTextures

This is a Neural Network based approach that makes use of GANs that rely on tracked geometry of the source and target frames. This information is extracted using *Face2Face*, before being fed into the Generative Network that performs the feature morphing from the source to the target.[11]

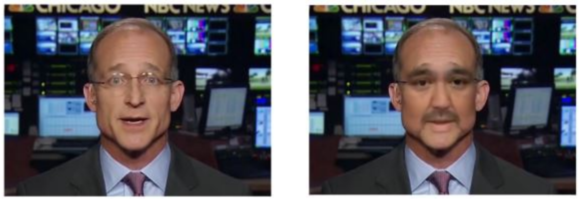

Fig. 3. Example from FaceSwap of real (left) and fake (right)

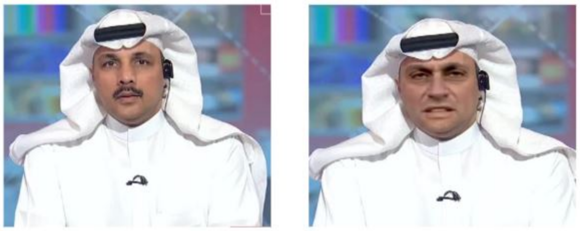

Fig. 4. Example from DeepFake of real (left) and fake (right)