Behavioral Analysis and Deepfake Detection

Marlee Bryant
The University of Alabama
mabryant4@crimson.ua.edu

ABSTRACT

Deepfake technology, intended for positive use in art and education, has unfortunately been corrupted for use in nonconsensual pornography and the spread of misinformation. Detection algorithms must constantly evolve to keep up with the development of more advanced Deepfake creation algorithms. The first section of this paper provides an overview of Deepfakes, including their origins, creation algorithms, and criminal uses. Following this, detection algorithms are discussed, including visual detection, algorithmic detection, and the shortcomings of these detection approaches. Finally, a unique approach to Deepfake detection is presented in the form of behavioral analysis. The application of behavioral analysis to Deepfake detection is explored, followed by details of the algorithmic approach.

1. DEEPFAKES

1.1 Origins

The term Deepfake refers broadly to falsified images, videos, or audio created by a deep learning artificial intelligence algorithm. The practice of creating Deepfakes emerged in the 2010's, and the term was coined in 2017 by a Reddit user, anonymously posting falsified pornographic content behind the username "deepfake," explaining that "deep" refers to the deep learning algorithm used to create the "fake" content [3]. Unlike modified content created by human manipulation, using a tool like Photoshop for example, Deepfakes require little expertise by the creator to produce an incredibly realistic result.

1.2 Algorithm Overview

The most common form of Deepfakes, and the kind which will be the focus of this article, are modified videos created using a face-swapping technique. For this variety of Deepfake, the input consists of a target video, which serves as the basis for the output, and a large dataset of photos depicting the face which will be "swapped" into the output video. A Generative Adversarial Network (GAN), depicted in Figure 1, uses facial recognition to generate a mask of the new face to replace the original for each frame of the video [3]. The first part of the algorithm is the Generator, which uses knowledge acquired from analysis of the input dataset to generate the mask, attempting to match the expression, pose, skin tone, and lighting conditions of the target video.

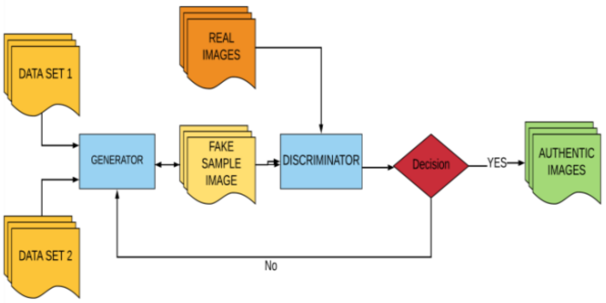

Figure 1. GAN algorithm depiction

The second part of the algorithm, the Discriminator, analyzes the generated frame for indications that it was modified, such as distortion of facial features or mismatched tones in adjacent pixels, and produces a confidence rating indicating how realistic the image is. These two phases of the algorithm are run iteratively until a sufficiently realistic image is produced, which replaces the original frame [3].

1.3 Training and Requirements

Several factors of the training process determine how realistic the output will be, but in general the tools necessary to create a Deepfake are easily accessible. In terms of hardware, the algorithm requires an extremely large number of computations and complex graphical analysis, so a powerful CPU and GPU are necessary [3]. Computers with advanced processing power are easily obtainable due to the prevalence of online gaming, which requires similar hardware capabilities. The software for creating Deepfakes is equally accessible, with multiple open-source options available to anyone with internet access [3]. Equipped with these tools, anyone could create a Deepfake, but without an understanding of the process behind deep learning, the result will likely be low quality.

To generate a realistic Deepfake, the most important factors are dataset quality and time for training. A quality dataset must be large and varied, containing photos which align well with the lighting, expressions, and facial orientations depicted in the original video [1]. Extensive training time is essential, since a greater number of iterations through the Discriminator allows a higher confidence score to be achieved [1]. Experiments conducted by the Deepfake lab compared different dataset sizes, shown in Figure 2, and durations of training, shown in Figure 3, to exemplify the importance of each of these factors.

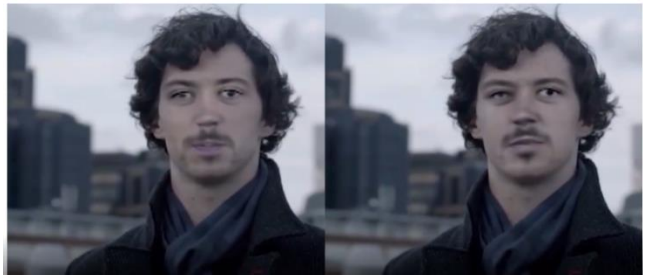

Figure 2. Comparison of Deepfake training set size, 200 images on the left, 2000 images on the right

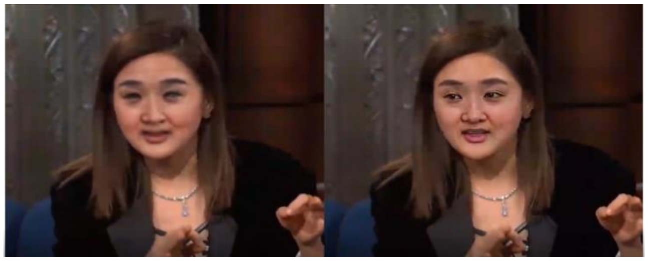

Figure 3. Comparison of Deepfake training time, 4 hours on the left, 48 hours on the right

From these experiments, flaws can be observed even with improved training, indicating an even larger dataset and training time are necessary for highly realistic output.

1.4 Criminal Use

While Deepfake technology was intended for use in art and education, its notoriety comes from prevalence in criminal activity. Most commonly, Deepfake technology is abused to create nonconsensual pornographic videos, oftentimes targeting celebrities. Additionally, Deepfakes have been used to manipulate social media news consumers, by depicting politicians or celebrities making compromising statements which do not align with their beliefs [6]. Due to the lack of thorough verification of social media news posts, as well as the potential for individual creator's content to go "viral," these posts can lead to widespread controversy and misinformation. Businesses can also be affected, even suffer financial losses because of Deepfakes depicting high ranking employees. Even for the average individual, Deepfakes can do harm, often in the form of "revenge porn" created by disgruntled ex-partners. The potential for misuse of Deepfakes abounds, making efforts to create effective and efficient algorithms for identifying Deepfakes incredibly important.

2. DETECTION TECHNIQUES

2.1 Visual Detection

Some indicators of a Deepfake video can be identified by a trained eye, but visual detection becomes increasingly difficult for Deepfakes which are skillfully made. The first visual indictors are found by observing the skin on and around the face. Overly smooth skin on the face occurs as a result of low resolution used to generate the mask, which often needs to be resized to fit the target video [2]. This indicator is especially difficult to use with faces wearing makeup, since the smoothness could be a result of makeup rather than algorithmic manipulation. Another identifying feature on the skin is mismatched skin tones on the neck or ears, but most Deepfake creators chose a target video with a subject whose skin tone closely matches the mask face, making skin tone variations nearly impossible to detect [2].

The best place to find Deepfake indictors visible to the naked eye is the area around the eyes. If the training set does not include photos of the mask face with closed eyes, the algorithm cannot properly replicate blinks. Comparison of the blinking rate in the video to natural human blinking rate can swiftly identify one of these poorly made Deepfakes [2]. Unfortunately, most creators are aware of this identifier due to its use in early Deepfake detection algorithms, so they ensure closed-eye photos are included in their training set. Additionally, due to low resolution of the generated mask, small moving parts such as eyebrows and eyelashes are difficult to realistically depict. Oftentimes flickering or warping of the mask can be seen in these areas when movement occurs, but newer Deepfake creation algorithms using higher resolution can easily avoid this [2]. While awareness of these visual indictors is important for quickly identifying a potential Deepfake, algorithmic analysis is usually necessary for a higher degree of certainty.

2.2 Algorithms for Detection

Algorithms for Deepfake detection all analyze the features of the potential fake content at different levels to determine whether they align with the rules of real video content. These detection methods have expanded and advanced over time to analyze the content in 4 different styles: visual, local, deep, and temporal [4]. The original approach to Deepfake detection relies on some of the same factors as visual detection, blinking rate and facial pose. As was previously mentioned, the blinking rate in Deepfake videos sometimes does not align with the natural human blinking rate, which can be assessed more accurately by an algorithm than simple visual analysis [4]. Another visual detection technique analyzes several points on the face to determine if the pose of inner facial features, like the eyes and nose, lines up with the pose of the outer facial features, like the chin and forehead, exemplified in Figure 4 [4]. Local feature-based detection is similar to visual feature-based detection, but rather than analyzing whole sections of the face, each frame is analyzed pixel by pixel in search of distortions which would not be found in a video created by usual means, like flickers of the original face appearing behind the mask [4].

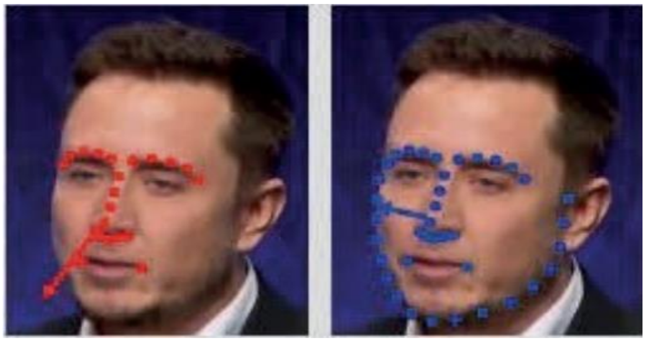

Figure 4. Facial pose analysis for Deepfake detection

Newer and more advanced techniques for Deepfake detection, falling into the categories of deep and temporal feature-based detection, use complex AI systems like those used to create the Deepfake. Deep detection methods analyze each frame at the pixel level, much like local detection, but are able to perform more thorough analysis through the use of a layered Convolutional Neural Network (CNN) [4]. One of such algorithms, known as MesoNet, detects compressed video conditions through the process displayed in Figure 5. Temporal feature-based detection takes an entirely different approach by analyzing the video in series of sequential frames, where the other algorithms analyze one frame at a time. The algorithm used for processing sequential data is a Recurrent Neural Network (RNN), in which the output of each batch of processing is used as input for the next batch, shown in Figure 6 [4].

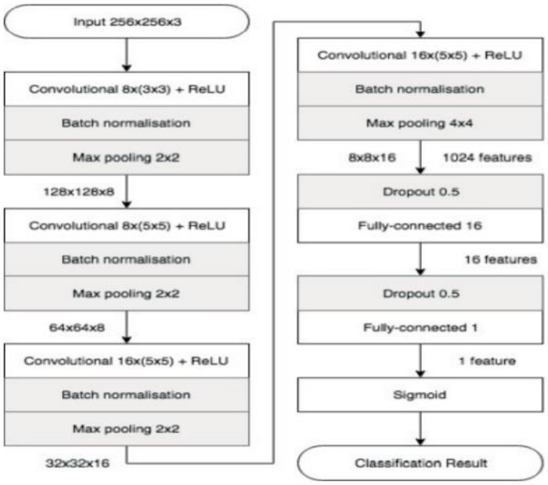

Figure 5. Depiction of MesoNet algorithm

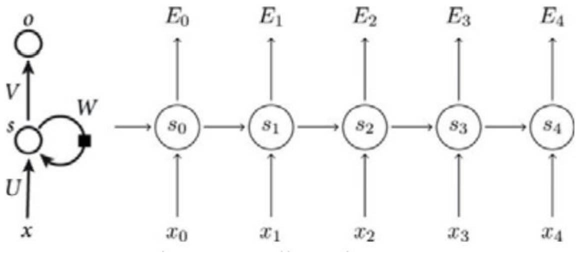

Figure 6. RNN algorithm depiction

2.3 Shortcomings

As quickly as Deepfake detection algorithms have advanced, Deepfake creation algorithms have advanced to avoid detection. Visual feature-based detection is mostly obsolete for modern Deepfakes since the creation algorithm has advanced to avoid obvious mistakes in blinking and facial pose misalignment. Since local feature-based extraction analyzes individual pixels in more detail, it is still effective in some cases, but modern Deepfakes are often so realistic that deeper analysis of each pixel is required to recognize mistakes. For these extremely realistic Deepfakes, only deep and temporal feature-based analysis is effective, some algorithms with up to 99% accuracy [4]. While this level of detection success is impressive, Deepfake creation algorithms will continue to advance with the goal of avoiding detection by these algorithms, so the accuracy is likely to decrease over time. This cat and mouse trend between Deepfake creation and detection will certainly continue until a detection algorithm is developed which is nearly impossible to evade.

3. BEHAVIORAL ANALYSIS

3.1 Overview of Algorithm

Behavioral analysis algorithms which categorize actions based on similarity exist for use in action-based video indexing, or “Intelligent Fast-Forward” [5]. This variety of behavioral analysis works by taking space-time measurements at multiple temporal scales in the video. Each action is represented on different temporal scales, for example in the action of a person walking limb movement would exist on the high temporal scale and movement of the body as a whole would exist on the low temporal scale [5]. Specific actions, performed at similar speeds, would be captured at the same temporal scales, simplifying comparisons between similar actions. This stochastic temporal process for each action can be used to generate an empirical distribution associated with each action, which is the key value required for determining the similarity of each action [5]. To obtain this value, the target videos must be altered through blurring and segmenting to isolate the temporal direction. In order to compare similar actions occurring in different scenarios, it is important to isolate the behavioral information from the photometric information, like lighting conditions and colors of clothing and background. Isolating this information is done by obtaining the gradient normal to the generated temporal surface and normalizing it to a length of one. The behavioral aspects of the gradient are represented by its direction and the photometric aspects are represented by its magnitude, so normalization erases this magnitude generated by information unnecessary for the comparison [5]. The resulting gradient measurements can be utilized for comparing the degree of similarity between two actions independent of their setting.

3.2 Application to Deepfake Analysis

Behavioral habits are intrinsic and unique to each person and contain subtleties which are difficult to identify through human analysis. When these behavioral habits are captured on video and therefore converted into data, AI algorithms can analyze them thoroughly and identify these subtleties which are otherwise undetectable. For this reason, behavioral analysis could be the key to creating a Deepfake detection algorithm which would be incredibly difficult to evade. The subject of the original video used to generate the output for the Deepfake is almost always a different person than the person who the mask depicts. It follows that this individual would have different behavioral patterns, including gestures, head and body movements, and expressions. If a unique behavioral pattern can be identified for the individual used to create the mask, this can be compared to the Deepfake video to determine if the behavioral patterns align. The primary limitation to this approach is the availability of a large dataset of videos depicting the individual and their behaviors, which would be necessary to identify distinct behavioral patterns. The most threatening use of Deepfakes is fake videos of celebrities and politicians created to spread misinformation and spawn dissent. Fortunately, large datasets of videos depicting these types of people are easy to obtain, making behavioral analysis an ideal solution for identifying this variety of Deepfakes.

4. REFERENCES

[1] Andrej Cattaneo, Ivana Riva, Noura Sammoura, Maria del Pilar Suarez Anzorena, Arthur van der Werf and Yueling Wu. Deepfake Lab, Unraveling the mystery around Deepfakes. https://deepfakelab.theglassroom.org/

[2] Artem A. Maksutov, Viacheslav O. Morozov, Aleksander A. Lavrenov and Alexander S. Smirnov. Methods of Deepfake Detection Based on Machine Learning. IEEE (2020).

[3] Digvijay Yadav and Sakina Salmani. Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. Proceedings of the International Conference on Intelligent Computing and Control Systems (2019).

[4] Kurniawan Nur Ramadhani and Rinaldi Munir. A Comparative Study of Deepfake Video Detection Method. 3rd International Conference on Information and Communications Technology (2020).

[5] Lihi Zelnik-Manor and Michal Irani. Statistical Analysis of Dynamic Actions. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, No. 9 (2006).

[6] Stamatis Karnouskos. Artificial Intelligence in Digital Media: The Era of Deepfakes. IEEE Transactions on Technology and Society, Vol. 1, No. 3 (2020).