International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 12 | Dec 2020 www.irjet.net p-ISSN: 2395-0072

A Brief Study on Deepfakes

Hrisha Yagnik[1], Akshit Kurani[2], Prakruti Joshi[3]

[1-3]*Student, Department of Computer Engineering, Indus University, Ahmedabad, India*

Abstract: *The word '**deepfake**' is a portmanteau of the words "deep learning" and "fake" referring to the realistic audiovisual content that is artificially developed mostly using **generative adversarial networks** (GAN). A branch of machine learning that applies neural net simulation to massive data sets, to make a fake. In order to transpose the face onto a target that is provided as though it were a mask, **artificial intelligence** effectively learns what a source face looks like from different angles. Application of GANs is done to pit two AI algorithms against each other, one creating the fakes and the other grading its efforts, teaching the synthesis engine to make better forgeries.*

Keywords: *Deepfakes, Generative Adversarial Networks (GANs), Artificial Intelligence (AI), deepfake detection, deepfake threats.*

1. INTRODUCTION

This technology was invented in 2014 by Ian Goodfellow, who works at Apple now. Deepfakes are the merchandise of AI (AI) applications that merge, combine, replace, and superimpose images and video clips to make fake videos that appear authentic. Deepfake technology can generate, for instance, a humorous, pornographic, or political video of an individual saying anything, without the consent of the person whose image and voice is involved. Deepfakes target social media platforms, where conspiracies, rumors, and misinformation spread easily, as users tend to travel with the gang. At an equivalent time, an ongoing 'infopocalypse' pushes people to think they can't trust any information unless it comes from their social networks, including relations, close friends or relatives, and supports the opinions they already hold. In fact, many of us are hospitable to anything that confirms their existing views albeit they think it's going to be fake.

2. BACKGROUND

2.1 How are Deepfakes made?

University researchers and computer graphics studios have long pushed the boundaries of what's possible with video and image manipulation. But deepfakes themselves were born in 2017 when a Reddit user of an equivalent name posted doctored clips on the location.

It takes a couple of steps to form a face-swap video. First, you run thousands of face shots of the 2 people through an AI algorithm which is called an encoder. The encoder finds and learns similarities between the 2 faces, and reduces them to their shared common features, compressing the pictures within the process. A second AI algorithm called a decoder is then taught to recover those faces from the compressed images. Because the faces are non-identical, you train one decoder to recover the primary person's face, and another decoder to recover the second person's face. To perform the face swap, you merely feed encoded images into the "wrong" decoder. For instance, a compressed image of person A's face is fed into the decoder trained on person B. The decoder then recreates the face of person B with the expressions and orientation of face A. For a convincing video, this has got to be done on every frame.

Another way to form deepfakes uses what's called a generative adversarial network, or GAN. A GAN pits two AI algorithms against one another. The primary algorithm, referred to as the generator, is fed random noise and turns it into a picture. This synthetic image is then added to a series of real images that are fed into the second algorithm, referred to as the discriminator. At first, the synthetic images will not look anything like faces. But repeat the method countless times, with feedback on performance, and therefore the discriminator and generator both improve. Given enough loops and feedback, the generator will start producing entirely life-like faces of completely nonexistent celebrities.

2.2 Who makes Deepfakes?

From academic & industrial academics to amateur enthusiasts to studios of special effects or porn makers. As part of their online campaigns to undermine and disrupt terrorist groups, governments may also be dabbling in technology, or making contact with targeted individuals, for example.

There are at least four main categories of deep-seated producers:1) deep-seated hobbyist groups,2) political actors, such as foreign governments and various activists,3) other malevolent actors, such as fraudsters, and4) legal actors, such as TV companies.

Overall, hobbyists prefer to see AI generated videos as a new type of online comedy, rather than as a way to trick or intimidate people, and to contribute to the advancement of such technology as solving an intellectual puzzle. These deepfakes creations are meant to be entertaining, funny, or politically satirical, and can help with gaining followers on social media.

Various political players, including political agitators, hacktivists, terrorists, and foreign states can use deepfakes in disinformation campaigns to manipulate public opinion and undermine confidence in a given country's institutions. In these times of hybrid warfare, deepfakes are weaponized disinformation aimed at interfering with elections and sowing civil unrest[1].

Deepfakes are also progressively used by fraudsters to carry out currency and stock manipulation, as well as numerous other financial crimes. Criminals have already used fake AI-generated audios to impersonate an executive asking for an urgent transfer of money.

2.3 Technology required to make Deepfakes

It is hard to form an honest deepfake on a typical computer. Most are created on high-end desktops with powerful graphics cards or better still with computing power within the cloud. This reduces the time interval from days and weeks to hours. But it takes expertise, too, not least to the touch up completed videos to scale back flicker and other visual defects. That said, many tools are now available to assist people make deepfakes. Several companies will make them for you (deepfakesweb.com) and do all the processing within the cloud. There's even a mobile app, Zao which lets users overlap their faces to a long list of TV and movie characters on which the system has trained.

2.4 Deepfake detection

Poorly-made video deepfakes are easy to identify – the lips are out of sync, the speaker isn't blinking, or there could also be a flicker on the screen. But, because the technology improves over time and NLP algorithms become more advanced, it's getting to be harder for people to identify deepfakes and other advanced impersonation scams. Ironically, AI is one among the foremost powerful tools we've to combat AI-generated attacks. AI can understand patterns and automatically detect unusual patterns and anomalies – like impersonations – faster and more accurately than a person's can.

But, we can't just believe in technology. Education and awareness amongst people is additionally incredibly important. Therefore it is promising to find out that 61% of IT leaders are already educating their workers about the danger of deepfakes and 27% have plans to continue to do so.

While several studies use features extracted supported visual artifacts, image quality, lipsync, blinking, or warping for classification (and might soon be obsolete), this work gives a generalizable statistical framework with guarantees on its reliability[2].

Another mainstream series of detection algorithms rely on data augmentation. Leveraging the imperfection of synthesized videos, e.g., warping artifacts, one can effectively distinguish Deepfake videos from benign counterparts[9, 10].

FaceForensics++ is an effective dataset of facial forgeries that enables training deep learning-based approaches. And research is the approach of detection fake video through trained CNN (Convolution Neural Network). These methods promise trustworthy results but require a lot of data and need to be improved periodically. Thus, we focused on research, which does not require as much data, but is likely to be used more widely[6].

We exploit missing reflections, and missing details in the eye and teeth areas for Deepfake detection. We again detect facial landmarks and crop the input image to the facial region[4]. We will repose on the information-theoretic study of authentication to cast deepfake detection as a hypothesis testing problem specifically for outputs of GANs, themselves viewed through a generalized robust statistics framework. The answer to the present problem has got to be driven by individuals until governments, technologists, or companies can find an answer. If there isn't an instantaneous push for a solution, though, it might be too late. What we should always do is demand that the platforms that propagate this information be held accountable, that the government enforces efforts to make sure technology has enough positive use cases to outweigh the demerits and have enough sense to not share them. Otherwise, we may find ourselves during a cyberwar that a hacker started supporting nothing but an augmented video.

3. EVALUATION

3.1 Consequences of Deepfakes

Artificial intelligence technology, to many, is scary. It's full of ethical dilemmas and uncertain future applications. Perhaps one of the more alarming possibilities enabled by AI development is the creation of deepfakes.

Deepfakes undermine credence in the information we see. They also present plenty of opportunities for misconduct and malicious use. It's hard to ascertain how we could hope to point out such functionality during a positive light — or how anyone could ever come to believe it.

The intent of generating such videos can be harmless and have advanced the research of synthetic video generation for movies, storytelling and modern-day streaming services. However, they can also be used maliciously to spread disinformation, harass individuals or defame famous personalities[7]. The extensive spread of fake videos through social media platforms has raised

significant concerns worldwide, particularly hampering the credibility of digital media[8].

But, as with any tool, there are boons and banes to deepfakes too.

3.1.1 Positive Impacts

• Educational benefits

Deepfake technology has a positive educational potential. It could revolutionize our lessons in history with interactivity. With deep samples of historical figures, it could preserve tales and help catch interest.

• Unifying the global audience

Since deepfakes can replicate voices and alter images, translated films that use the original actors can be allowed. The voices sound like that of the original. And, crucially, the gestures of the lips also lead to the words spoken.

• The entertainment industry

Consider the days that the ghost has been given up by an actor. The role of CGI can be filled by deep fake technology, recreating the likeness of unavailable past actors. So with their actor, the character does not have to pass away. The recreation of the late Peter Cushing in Star Wars: Rogue One(2017), who died in1994, for instance.

• Art world

AI software could help us create virtual museums. This will encourage people who otherwise would not be ready to encounter them face to face to access the world's masterpieces. We should share the planet's compelling, profound artwork.

• Medical community

This will offer a boost to data protection, thus assisting with the emergence of current diagnostic and tracking procedures. Hospitals can produce deep-fake patients by using the technology behind deep-fakes. That is, patient data that is practical for research and experimentation, but does not place actual patients at risk. So rather than actual patient data, researchers can use true-to-life deep-fake patients. From this there is space to examine new diagnostic and monitoring techniques. Or even train other AI to help with medical decisions.

• Training

It takes masses of data to train artificial intelligence. Often there are questions as to where all this knowledge comes from. There's also the matter of algorithmic prejudice induced by a lack of knowledge diversity. We can, however, generate true-to-life, diverse artificial data with deep-fake technology. And so the issue of requiring further training data could cease. Or for onboarding team members, consider some preparation. For example, customer service training. It's possible that we could one day have realistic, deepfake virtual humans. This could provide deepfake examples of real customers. That way, you'll train your new team members, without throwing them headfirst into real customer interactions.

3.1.2 Negative Impacts

• Deepfakes are a serious threat to our society, form of government , and business because they 1) put pressure on journalists struggling to filter real from fake news, 2) threaten national security by disseminating propaganda and interfering in elections, 3) hamper citizen trust toward information by authorities, and, 4) raise cyber security issues for people and organizations.

• It's highly probable that the journalism industry is going to face a huge consumer trust issue thanks to deepfakes. Deepfakes pose a greater threat than "traditional'' fake news because they're harder to identify and other people are inclined to believe the fake is real. The technology allows the assembly of seemingly legitimate news videos that place the reputation of journalists and therefore the media in danger.

Purveyors of deepfakes can try to make it difficult for people to determine whether a video comes from a source that is subject to normative constraints. Indeed, this is precisely what has happened with the phenomenon of fake news. With text-based news, there have never been any technical constraints preventing someone from writing a story about something that never happened. Thus, text-based news is only trustworthy if we can tell that the source is subject to normative constraints[5].

• During the spike in tensions between India and Pakistan in 2019, Reuters found 30 fake videos on the incident; mostly old videos from other events posted with new captions. Misattributed video footage such as a real protest march or violent skirmish captioned to suggest it happened somewhere else is a growing problem, and will be augmented by the rise of deepfakes[1].

• The intelligence community is concerned that deepfakes will be used to threaten national security by disseminating political propaganda and disrupting election campaigns. Putting words in someone's mouth on a video that goes viral may be a powerful weapon in today's disinformation wars, intrinsically altered videos can easily skew voter opinion.

• Deepfakes are likely to hamper digital literacy and citizens' trust toward authority-provided information, as

fake videos showing officials saying things that never happened make people doubt authorities.

• Nonetheless, the foremost damaging aspect of deepfakes might not be disinformation intrinsically, but rather how constant contact with misinformation leads people to feel that much information, including video, simply can't be trusted, thereby leading to a phenomenon termed as "information apocalypse" or "reality apathy". In other words, the best threat isn't that folks are going to be deceived, but that they're going to come to take everything as deception.

4. POSSIBLE SOLUTIONS TO DEEPFAKES

The news articles reviewed indicate that there are four methods of fighting deep fakes:1) legislation and regulation,2) corporate strategies and voluntary action,3) education and training, and4) anti-deep fake technology, including deep fake identification, authentication of content, and deep fake prevention.

Legislation and legislation are also transparent forms of profound falsification. Deepfakes are not currently explicitly addressed by civil or criminal legislation, while legal scholars have proposed adapting existing laws to cover libel, slander, identity theft, or using deepfakes to mimic a government official. The social media companies of today enjoy strong immunity for the content shared on their sites by users. One legislative alternative may be to remove the legal immunity of social media sites from the content shared by their users making not only users but also websites more accountable for posted information.

Recent studies have analyzed how easy it is to detect audio and text-to-video fake content. Recently, a fake detection method was proposed that exploits the inconsistencies that exist between the dynamics of the mouth shape (visemes) and the spoken phoneme. They focused on some particular visemes in which the mouth must be completely closed and observed that this did not happen in many manipulated videos. Their proposed approach achieved good results, especially as the length of the video increases[3].

Social media firms need to impose ethics and move away from the fact that having divisive content pushed to the top of the feed is financially a win because it maximizes advertising interaction time. Although few social media companies still have deep-faced policies, they can cooperate to prevent misinformation from being equipped with their platforms and thus proactively implement straightforward, common policies to block and delete deep-faced content. Actually, many businesses do not delete contested content, but rather down-rank it to make it harder to locate, by being less popular in news feeds for consumers. The rise of hate speech, false news, and misinformation polluting digital networks, on the other hand, has prompted several businesses to take further action, such as suspending user accounts and investing in faster technology for detection. Deepfake technology decreases the amount of information that videos carry by increasing the probability that realistic fake videos depicting events that never occurred will be produced. Thus, an obvious strategy for increasing the amount of information that videos carry is to decrease the probability of realistic fake videos being produced. Although there are fewer and fewer technical constraints on the production of realistic fake videos, it is still possible to impose normative constraints[5].

Subtle indicators for the identification of deepfakes have been suggested by media forensic experts, including a number of imperfections such as face wobble, shimmer and distortion; waviness in the motions of a person; inconsistencies in expression and mouth movements; irregular fixed object movements such as a microphone stand; inconsistencies in lighting, reflections and shadows; blurred edges; angles and facile blurring; missing facial features such as a recognized mole on a cheek; clothing and hair softness and weight; excessively smooth skin; missing hair and teeth details; face symmetry misalignment; pixel level inconsistencies; and an individual's unusual behavior doing anything. Although differentiating between a real video and a fake is becoming more and more difficult for individuals, AI can be instrumental in detecting deep fakes. AI algorithms, for example, can analyze Photo Response Non-Uniformity (PRNU) patterns in footage, i.e. imperfections peculiar to the light sensor of particular camera models, or biometric data such as blood flow indicated in a video by subtle changes that occur on the face of an individual. New algorithms for fake detection are based on mammalian auditory systems.

5. CONCLUSIONS

According to the study, deepfakes are a serious threat to society, the form of government and businesses because they put pressure on journalists struggling to filter real from fake news, threaten national security by disseminating propaganda that interferes in elections, hamper citizen trust toward information by authorities, and lift cyber security issues for people and organizations.

On the opposite hand, there are a minimum of four known ways to combat deepfakes, namely 1) legislation and regulation, 2) corporate policies and voluntary action, 3) education and training, and 4) anti-deepfake technology. While legislative actions are often taken against some deepfake producers, it's not effective against foreign states. Rather, corporate policies and voluntary action like deepfake-addressing content moderation policies, and quick removal of user-flagged content on social media platforms, also as education and training that aims at improving digital media literacy, better online behavior

and critical thinking, which create cognitive and concrete safeguards toward digital content consumption and misuse, are likely to be more efficient.

So in this limited scope of research we can say that there are a lot of factors involved that determine the nature of Deepfakes. Arguments, both in favor and against, can be made but they will only prove to be imperative if the development of and access to Deepfake generating tools are governed properly. If closely monitored then Deepfakes can be used to benefit modern-day humanity rather than cause its downfall.

REFERENCES

[1] Mika Westerlund "The Emergence of Deepfake Technology: A Review" Technology Innovation Management Review, Volume 9, Issue 11, November 2019

[2] Sakshi Agarwal and Lav R. Varshney "Limits of Deepfake Detection: A Robust Estimation Viewpoint" arXiv, May 2019

[3] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales and Javier Ortega-Garcia "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection" arXiv, June 2020

[4] F. Matern, C. Riess and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," IEEE Winter Applications of Computer Vision Workshops (WACVW), ISBN:978-1-7281-1393-7, 2019

[5] Fallis, D. "The Epistemic Threat of Deepfakes" Springer, August 2020

[6] T. Jung, S. Kim and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," IEEE Access, Volume. 8, ISSN: 2169-3536, April 2020

[7] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing obama: learning lip sync from audio" ACM Transactions on Graphics (TOG), Volume 36, Issue 4, July 2017.

[8] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, Julian McAuley "Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples" arXiv, November 2020

[9] Yuezun Li, Siwei Lyu "Exposing DeepFake Videos By Detecting Face Warping Artifacts" arXiv, May 2019

[10] Chaofei Yang , Lei Ding, Yiran Chen , Hai Li "Defending against GAN-based Deepfake Attacks via Transformation-aware Adversarial Faces" arXiv, June 2020