Paper ID: 20245

International Youth Science Journal

IYS Journal: Volume 1, Issue 1, May 2024, Paper ID: 20245

https://iysjournal.netlify.app

Accuracy Of Deepfake Detecting Reverse Engineering Algorithms

Tuhin Sarkar
Narayana E-techno and Olympiad School
Admission Number: **5788914**
School Referral Email: blrwf.etechno@narayanagroup.com Teacher (s) or Mentor/Mentee: N/A

Abstract:

The advent of deepfake technology, which uses artificial intelligence to create hyper-realistic and fabricated video and audio, has introduced significant challenges for social media platforms and their users. This paper explores the harms of deepfakes, including the erosion of trust in digital media, the potential for political manipulation, the facilitation of cyberbullying, and the threat to personal privacy. This study further examines possible solutions that include limiting access to artificial intelligence, furthermore, the urgency for effective countermeasures. To mitigate these threats, the paper discusses a comprehensive approach involving technological solutions, policy interventions, law enforcement, and public education. Specifically, the advancement in detection algorithms, robust regulatory frameworks, and increased digital literacy among users are put forward as essential strategies to combat the proliferation of deepfakes on social media.

I. Introduction

The way of communication and human interaction has changed. Social media platforms are free for the end user, as these platforms are mainly monetized by advertisement revenue. Due to the massive audience on social media, every business irrespective of its size has a presence on social media. Actors, celebrities, and even government officials are prone to defamation and the misleading effects of AI-generated deepfakes. Political figures have an even more significant impact due to fabricated and deepfake content.

Deepfakes have garnered widespread attention for their potential use in creating child sexual abuse material, celebrity pornographic videos, revenge porn, fake news, hoaxes, bullying, and financial fraud.[1]

Hence, it is a must to address this growing issue of breached online security and privacy. Deepfakes produced by artificial intelligence are often so realistic that they are practically impossible to notice by the human eye. However, it is possible to be a vigilant netizen if certain protocols are followed and public education is promoted on this matter.

This research focuses on the progress of (A.I) generated, deepfake, and fabricated content, how real they are, what are the possible harms caused by them, how it can be prevented, and the accuracy of reverse engineering and machine learning algorithms at detecting fabricated content.

II. Defining (A.I) generated content, and fabricated deepfakes.

Deepfakes: an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said. [2]

In simple terms, deepfake or synthetic media is a form of digitally created or manipulated content that is used to alter reality in a convincing manner.

This technology employs advanced machine learning algorithms and artificial intelligence (A.I) to produce realistic images, videos, or audio samples that can make it appear as though someone is saying or doing things that they never actually did or said.

The primary techniques used for deepfakes include:

1. **Face Swapping:** This includes replacing the person's face with the face of another person. This method depends on neural networks to map and blend facial features seamlessly.

2. **Voice Cloning:** Synthesizing a person's voice to generate audio that mimics their speech pattern, voice, tone and intonation.

3. **Image Synthesis:** Generating entirely new images or videos that look and feel real but are completely fabricated.

III. Materials and Methods •

Sample Deepfake Clips used for testing:

1. Viral Deepfake Videos Thrive Of Aamir Khan & Ranveer Singh Endorsing Political Parties – Business today

2. Amit Shah Fake Video: Debunking the Fake Video of Amit Shah On Reservations – The Indian Express

3. AI Deepfake Video Of Actress Rashmika Mandanna Going Viral – Business Today

4. Anderson Cooper, 4K Original/(Deep) Fake Example - LipSynthesis

5. Fake Obama created using AI video tool - BBC News

6. This is not Morgan Freeman - A Deepfake Singularity

7. President Joe Biden's Magical Pistachio Story (Deepfake AI)

8. Deepfake example. Original/Deepfake close shot Bill Gates.

9. 2024 Deepfake Example in 4k - ORIGINAL/DEEPFAKE - Bill Gates

10. Trump 4k Deepfake example - LipSynthesis

Samples:
https://drive.google.com/drive/folders/1ZxVZA6e1kp4gbte9ylJDGYsSbmTprDTG?usp=sharing

- Tools used:

Note: The only tools that are tested in this paper are free-to-use and available to the public. The reason being, accessibility and accuracy of my research. If a tool is used that is paid, it will make it harder for the public to test and validate the accuracy of my research. Many reputable deepfake detectors are not tested as they are not available to the general public, and require to schedule a product demo with a company email address. Do note that the situation of the tools available for free in the market will change and the accuracy of those will improve.

- Tools used:

1. Deepware AI - https://deepware.ai/

III. Effectiveness of tools

As deepfakes get more sophisticated with time and more computing power. The line between real and fake starts to fade. That is not to say that there are no solutions to combat this. In such scenarios, deepfake tools are to be used.

The effectiveness and accuracy of deepfake detection algorithms may vary. It is always possible that a fabricated video will be identified as real. Deepfake content gets more advanced and sophisticated as computing power and computing time increases.

1. Deepware (A.I)

Deepware (A.I) is a company that was launched in mid-2018 by their parent company Zemana to develop advanced deepfake detection available to everybody.

a. Features: Deepware (A.I) comes with only one feature, that is, deepfake content detection. The U.I is very simple with minimal distractions. No external advertisements are shown. Though in the navigation bar at the top, there is a button that takes you to an external-paid website called "PAGI GEN" which works as an (A.I), Machine Learning, and Deepfake generating company for movie production.

b. Tests:

i. The **first deepfake** that was tested was a deepfake of **Donald Trump**. Upon testing, it showed up as **real**. **(NO DEEPFAKE DETECTED),** which is not true.

- Model Results: ***Avatarify: NO DEEPFAKE DETECTED (40%), Deepware: NO DEEPFAKE DETECTED (2%), Seferbekov: NO DEEPFAKE DETECTED (1%), Ensemble: NO DEEPFAKE DETECTED (1%)***.

ii. The **second deepfake** that was tested was of **Rashmika Mandanna**. Upon testing, the media could not be displayed as it was marked to be "Adult Content". It came up as real **(NO DEEPFAKE DETECTED)**, which is not true.

- Model Results: ***Avatarify: NO DEEPFAKE DETECTED (24%), Deepware: NO DEEPFAKE DETECTED (16%), Seferbekov: NO DEEPFAKE DETECTED(46%), Ensemble: NO DEEPFAKE DETECTED(28%)***.

iii. The **third deepfake** that was tested was that of **Ranveer Singh.** Upon testing, It came up as **real (NO DEEPFAKE DETECTED)**, which was not true.

- Model Results: ***Avatarify: NO DEEPFAKE DETECTED (12%), Deepware: NO DEEPFAKE DETECTED (1%), Seferbekov: NO DEEPFAKE DETECTED (6%), Ensemble: NO DEEPFAKE DETECTED (2%).***

iv. The **fourth deepfake** that was tested was that of **Barack Obama**. Upon testing, It came up as **(DEEPFAKE DETECTED).** Most likely due to a invisible signature put on it.

- Model Results: ***Analyst: DEEPFAKE DETECTED, Avatarify: NO DEEPFAKE DETECTED (19%), Deepware: NO DEEPFAKE DETECTED (0%), Seferbekov: NO DEEPFAKE DETECTED (49%), Ensemble: NO DEEPFAKE DETECTED (12%).***

v. The **fifth deepfake** That was tested was that of **Anderson Cooper**. Upon testing, it came up as **(SUSPICIOUS).**

- Model Results: ***Avatarify: SUSPICIOUS(72%), Deepware: NO DEEPFAKE DETECTED(0%), Seferbekov: NO DEEPFAKE DETECTED(3%), Ensemble: NO DEEPFAKE DETECTED (0%).***

vi. The **sixth deepfake** That was tested was that of **Morgan Freeman.** Upon testing, it came up as **(DEEPFAKE DETECTED).** Which was most likely due to this being a popular deepfake video and an invisible signature in the video itself.

- Model Results: ***Analyst: DEEPFAKE DETECTED, Avatarify: NO DEEPFAKE DETECTED (18%), Deepware: NO DEEPFAKE DETECTED (0%), Seferbekov: NO DEEPFAKE DETECTED (0%), Ensemble: NO DEEPFAKE DETECTED (0%).***

vii. The **seventh deepfake** That was tested was that of **Joe Biden**. Upon testing, it came up as **(SUSPICIOUS)**

- Model Results: ***Avatarify: NO DEEPFAKE DETECTED (29%), Deepware: NO DEEPFAKE DETECTED (34%), Seferbekov: SUSPICIOUS (71%), Ensemble: SUSPICIOUS (58%)***.

viii. The **eight deepfake** That was tested was that of **Bill Gates**. Upon testing, it came up as **(NO DEEPFAKE DETECTED)** which is not true.

- Model Results: ***Avatarify: NO DEEPFAKE DETECTED (20%), Deepware: NO DEEPFAKE DETECTED (0%), Seferbekov: NO DEEPFAKE DETECTED (2%), Ensemble: NO DEEPFAKE DETECTED (0%)***.

ix. The **ninth deepfake** That was tested was that of **Amit Shah**. Upon testing, it came up as **(SUSPICIOUS).**

- Model Results: ***Avatarify: NO DEEPFAKE DETECTED (0%), Deepware: NO DEEPFAKE DETECTED (20%), Seferbekov: DEEPFAKE DETECTED (97%), Ensemble: SUSPICIOUS (67%).***

x. The **tenth deepfake** That was tested was that of **Amir Khan**. Upon testing, it came up as **(SUSPICIOUS)**.

- Model Results: *Avatarify:* ***NO DEEPFAKE DETECTED (39%), Deepware: NO DEEPFAKE DETECTED (25%), Seferbekov: SUSPICIOUS (75%), Ensemble: SUSPICIOUS (55%).***

c. Rating: After testing, 4 out of 10 deepfakes showed up as **(no deepfake detected)**. 4 out of 10 showed up as (**suspicious**). 2 out of 10 showed up as **(deepfake detected)**. The accuracy of Deepware (A.I) is not on point. Upon scanning, 4 out of 10 fake videos show up as real. In this research, only fake videos were scanned and not real ones. As Deepware Scanner is still in beta, the results may not be very accurate.

III. Conclusion

It is important to note that deepfakes are not perfect. After all, it's an (A.I), though deepfakes of celebrities or famous figures may look more real due to more pictures being available of them online, that can be fed to a (A.I). It is vitally important to note that (A.I) generated or fabricated content have flaws that can and sometimes cannot be noticed by the human eye.

It is crucial to know that some (A.I) generated and fabricated content can look indifferentiable to real content to the human eye. Hence, tools such as Deepware scanner may also be used due to it being free and available to the general public without the requirement for a product demo or a trial using a company email address. Additionally, people can train themselves to be (A.I) educated and aware by taking tests that educate and help people recognize the difference between real and fake.

Not to mention (A.I) scams that are currently on the rise. In such harassment or cyberbullying, the victim receives a text from an unknown anonymous person saying that they have the victim's explicit images, and demand some amount of ransom from the victims, in exchange for keeping these images private. Some other kinds of scams include a friend, family or a close relative calling in desperate need of help and requests some amount of money in utmost urgency.

The current infrastructure of 2024 is not enough to accurately detect professionally made deepfakes. Deepfake content detectors are not accessible to the general public and most of the deepfake content detectors online require a product demo trail with a company email address. With increasing computer power and computing time, the deepfakes get harder to spot by both humans and machines.

Policymakers should create strong criminal and civil liability for people that distribute nonconsensual intimate audiovisual content, including AI-generated content, as well as for people that threaten to do so. Penalties should be particularly severe when the victim is a minor.[3]

This trend can be prevented through public education. Even in 2024, a lot of computer science (IT) school books lack basic knowledge about ethics related to fabricated content. If cybercrimes like such need to be prevented, public education on this matter is a must.

[1] WikiPedia. "Deepfake", https://wikipedia.com, Accessed 26.05-2024. https://en.wikipedia.org/wiki/Deepfake

[2] Merriam webster. "deepfake." Accessed 26.05-2024. https://www.merriam- webster.com/dictionary/deepfake

[3] IBM Newsroom. "Here's What Policymakers Can Do About Deepfakes, Right Now," February 28, 2024. Accessed May 26, 2024. https://newsroom.ibm.com/Blog-Heres-What-Policymakers-Can-Do-AboutDeepfakes#:~:text=Policymakers%20should%20create%20strong%20criminal,the%20victim%20is%20a%20 minor