Deepfake impact on cyber security and potential methods to detect and mitigate their harmfull effects

Alanas Jakonis C00278356
MSc Cybersecurity, Privacy and Trust

Introduction

The objective of this thesis is to explore the impact of deepfakes on cybersecurity. The aim is to propose and explore potential methods to detect and mitigate their harmful effects. Deepfakes are AI computer generated images, videos and audio recordings that mimic real people, they are often used for malicious purposes. The primary goal of this research is to provide an in-depth analysis of deepfakes and their impact on cyber space and propose effective counter measures to detect the deepfakes and mitigate their harmful effects.

Research questions/hypothesis

- What are the risks and benefits of deepfake technology?
- What are the future implications of deepfake technology on society?
- How can deepfakes be regulated to mitigate the risks ?
- What is the percentage of authentication bypassed by the created deepfakes?
- Which iteration rage can create the most realistic and successful deepfake?
- Which type of created deepfake (voice, image, video) is more successful at bypassing the authentications on various devices?

The hypothesis is that the deepfake technology will bypass the authentication methods.

Summary of literature review

Deepfakes are a rapidly growing technology, they are developing so fast, that the researchers cannot keep up with coming up with new software's to detect them. Deepfakes are ran by AI and ML, therefore the software keeps on improving and learning from its own mistakes, soon, they will be flawless and impossible to detect. However, some scientists state, that we should be able to differentiate between an artificially generated content and a content that has been made using an actual camera (Mahmud & Sharmin, 2021). Researchers are working hard to keep up with deepfake detection methods, some deepfake give-aways include inconsistencies with lighting or unusual pixel formations. Deepfakes can also be discovered by looking at the resolution, or there is also a capsule network-based system to detect deepfakes. Overall, the scientists are usually looking for any type of inconsistencies or unusual movement. There has been a lot of times when deepfakes has been used for a malicious intent (like the Donald Trump's speech, or generation of pornographic images), it is also important that they can be used for greater good, such as rare disease detection or artificial generation of voice for those unable to speak. Based on the gathered research, I think that there is more bad then good when it comes to deepfakes. They can be used for impersonation, blackmail, slander, misinformation, and fraud, not forgetting the deepfakes could cause the "Liar's Dividend" effect, which in the long run could cause more criminal behaviors. At the present moment given that there is minimal to no legislation, deepfakes pose a threat to our society and there needs to be more laws around the creation and online posting of deepfakes.

TECHNOLOGIES

python SOFTWARE FOUNDATION

NVIDIA CUDA

Android Studio

Methodology

This study will look at previous research done on deepfakes to determine the harmful effects of this new emerging technology.

I will also attempt to create multiple deepfakes of myself (voice, image, video) from sources which are already published on the internet, this will show how easy it is to create a deepfake of anyone who posts as little as their face online.

The created deepfakes will range in different amount of iterations in order to determine the minimum number of iterations needed to produce a deepfake that is realistic enough to bypass authentication.

The created deepfakes will be used to attempt to bypass authentication on a mobile, laptop, computer devices and some applications which require authentication.

Each attempt of a bypass will be noted. After the data is collected, it will be analyzed to see the percentages of successful/unsuccessful authentication bypasses, and which iteration rage can create the most realistic and successful deepfake. Each type of created deepfakes (voice, image, video) will also be analyzed to find which was more successful at bypassing the authentications.

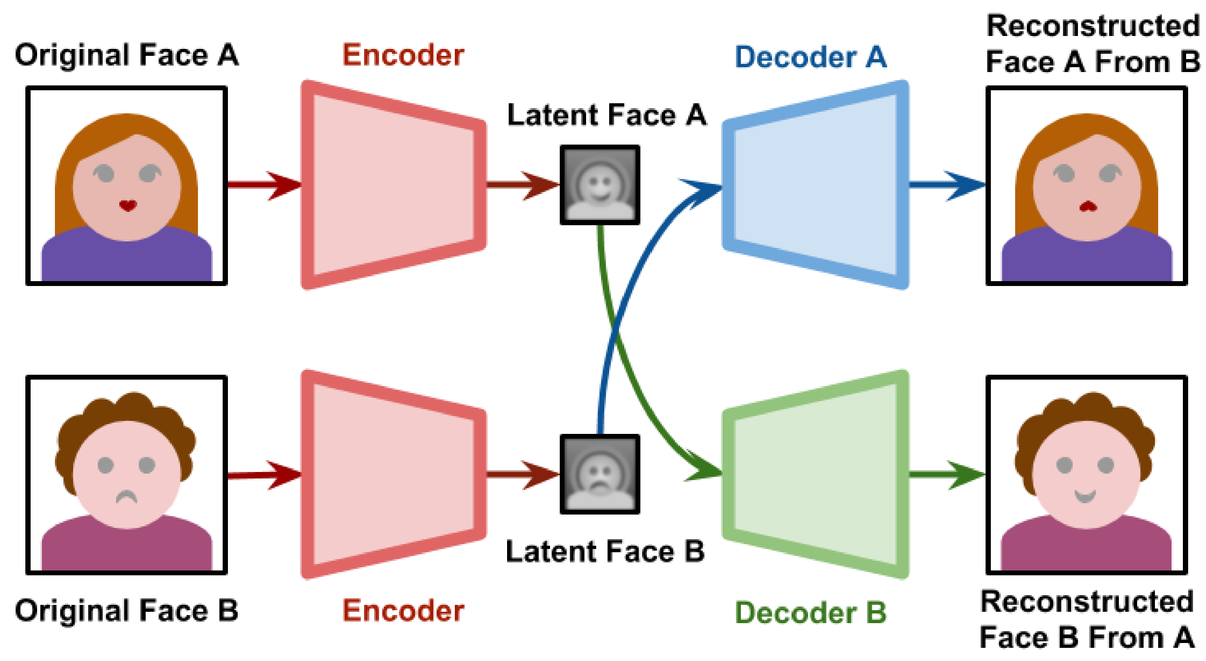

Figure 1: Example of deepfake creation

Early indications/next steps

Early indications suggest that the deepfakes can pose a great treat on society and cyber space as a whole. Deepfakes have been increasing in volume over the years. There is documented cases where a high profile person is getting tricked by a deepfake.

From the research done so far, another early indication I can see is that creating a deepfake is very easy and accessible to public.

Another early indication I have noticed is that the deepfakes have been becoming more and more realistic and believable over the years to the point where it is very hard to distinguish between a deepfake and a real image, video or voice.

For the next step of my thesis, I will focus on creating my own deepfake of myself this will be a deepfake image of me , deepfake video of me and deepfake voice of me. Using these deepfakes I will try to bypass authentication of smart devices such as computer, laptop and phone. I will also use these deepfakes to bypass some applications which use either image or voice authentication.

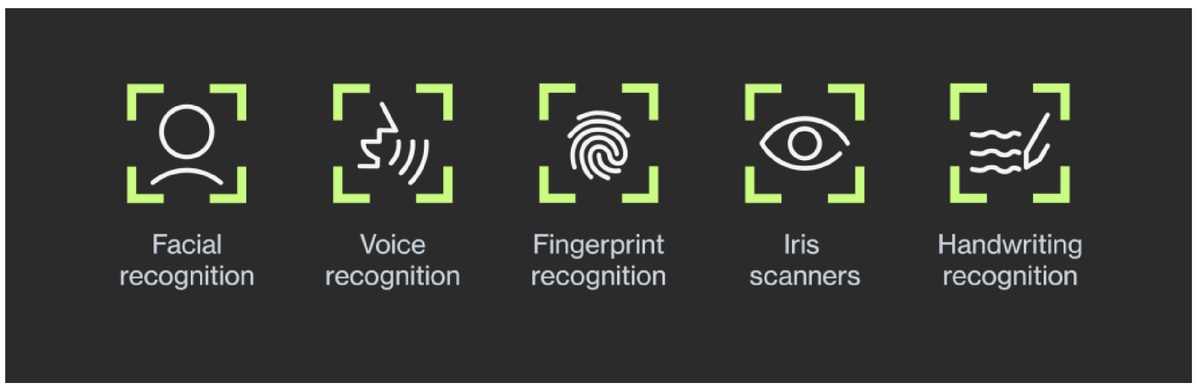

Figure 2: Examples of common biometric authentication

ACKNOWLEDGEMENTS

Alanas Jakonis
jakonis77@gmail.com
+353857735199