Risk assessment of deepfake technology

BIRAVEEN NEDUNCHELIAN

HA HOANG

SHOAIB AHMED

YUSUF OSMAN

INNHOLDSFORTEGNELSE

Risk assessment of deepfake technology

INTRODUCTION

In this assignment we have chosen to do a risk assessment of the AI technology Deepfake. We will briefly describe the context of the application, method of A.I used, current and expected A.I outcomes. Then we will provide an account of relevant risks and collateral consequences related to using the A.I technology.

WHAT'S AN AI?

Artificial intelligence, also known as AI for short, is about developing computer systems capable of doing tasks that usually requires "human intelligence" to execute. There are many examples of AI that includes, but are not limited to, visual perception, speech recognition, decision making, self-learning and translations. In our report we will focus on visual perceptions, more specifically **Deepfakes**.

In short Artificial intelligence is about introducing intelligence to machines, so that the system is able to learn on its own, recognize problems and solve problems.

LEARNING ALGORITHMS

There are three main learning paradigms in Machine Learning. These include supervised learning, unsupervised learning and reinforcement learning. Each of the learning algorithms have their own benefits and challenges, so deciding which learning algorithm to use is dependent on the desired output.

SUPERVISED LEARNING:

This form of learning uses labeled training datasets. The AI algorithm learns patterns in the labeled dataset to then solve and evaluate problems with unlabeled data to categorize it in labels. This is for example used in spam filters, face detection, stock price prediction, image recognition etc.

UNSUPERVISED LEARNING:

The machine learning algorithm gets an unlabeled dataset which it uses to find patterns and sort into labeled datasets. It does this by creating clusters in the datasets to find patterns and extract

features on its own. This is used in for example customer segmentation, target marketing systems, fraud detection, visualization of graphs, diagrams, charts etc.

Reinforcement Learning:

This learning method is used in interactive learning environment where the system is rewarded or punished by success or failure. This means that the system gets a desired output in an environment that it can perceive and interact with, to reach the desired output by trial and error. This is by example used in autonomous cars, games, robotics, recommendation systems etc.

Where did deepfakes come from, who created it and how did it become mainstream?

According to an article from Wikipedia: "Deepfake technology has been in development by researchers at academic institutions beginning in the 1990s, and later by amateurs in online communities". The term stems from the two words "Deep learning" and "fake". By merging them together we get "deep fake".

A reddit user by the name, "deepfakes", used deep fake to post non-consensual fake adult videos, using an AI to swap celebrities' faces onto real porn actors and actresses. A few years later the technology exploded, and the videos started to show up on social medias all over the internet. Deep fakes have since then had a dark tone to it, but as time has progressed deep fakes have also been used for positive applications.

Technicalities

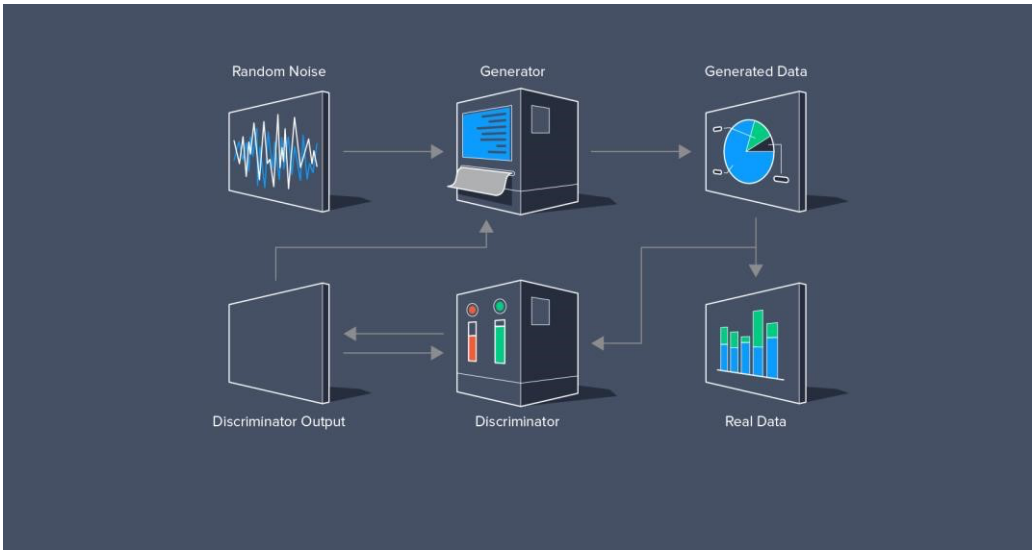
Deepfakes rely on autoencoder, which is a type of neural network, and are created by using two AI algorithms. The algorithms are named the discriminator while the other is named the generator. The discriminator determines whether the media given to it from the generator is real or fake. Together they create a generative adversarial network, or GAN for short.

GAN

GAN is short for Generative Adversarial Network and is a Machine Learning model, which uses two neural networks to compete in order to create new data that in many cases will pass as real data or content. Generative modeling is an unsupervised learning task in Machine Learning that is discovering and learning patterns with the input data. Thereby it can, in some ways, generate output data that in a plausible way have been drawn from the original dataset.

GAN is a smart way to train generative models by framing the problem as a supervised learning problem. The model has two sub models. One of the sub models is called the generative model, and we train that model to generate new data examples. The other sub model is called the discriminator and it tries to classify data as real or fake. Most of the time the discriminator is fooled by the generator and that means the examples are plausible. If the discriminator finds out that the data is fake the generator will update the new information and output some new data examples for the next test.

As mentioned previously GAN uses two AI models, a “generator” and a “discriminator”. We have attached a picture that visualizes the way GAN works:



picture taken from https://www.toptal.com/machine-learning/generative-adversarial-networks

GAN learns in a way that humans also learn. You try something, you get some feedback, adjust your strategy, and try again

The GAN network can generate data from scratch. You can feed it some random noise and it can output some realistic data. The network can also produce new data with limited amount of data. Especially when it is difficult to find data or buy data, considering how expensive it is. Imagine that you are going to make a game with a cat and mouse, but you don't have enough data to create a realistic picture of a cat or mouse. GAN can thereby be a good tool as it can produce new data with the little input it gets.

PROBLEMS AND CONTROVERSIES

As technology is advancing it is getting more and more complicated to detect real content from fake content. Deepfake technology gives scammers more realistic tools to fool people with fake news, blackmails and other manipulations techniques. This has grown a great concern to the general people whether this technology is for the better or worse case.

Well-known people in politics have been targeted by deep fake creators that make fake videos of them. This can create political confusion that can sway the voters during elections if they cannot spot the difference between the fake and real videos. The majority of people are getting their everyday news from the media which has also made it easier to target audiences with fake news and misinformation.

Deepfake can also be used to blackmail people by generating fake materials which incriminates the target. This can be used to make the target send money in crypto currency or manipulate them to do favors. Another issue that has arisen with this technology is celebrity pornography or revenge pornography created with deepfake. Pornographic videos are then fabricated to portray public figures which can ruin or change their reputation.

The concerns regarding deep fakes are therefore regarding the misinformation that can warp someone’s perception of reality and the material damage like financial loss, credibility of content or ruined reputations.

POSITIVE OUTCOMES

Using deepfakes is not all about disadvantages. There are also positive outcomes that come with using deepfakes, such as being a supplement to bring history and art "alive" for a wider audience. An example of this is when a Scottish company named *Cereous* managed to train its deepfake algorithms on audio recordings of John F. Kennedy the former president of the

USA. The company was able to create "lost" audio of the speech in Dallas the day he was assassinated. This helps to get thousands, or millions of people interested in art and history.

Other positive uses with deepfakes are emerging in education and entertainment. There was a health charity from the UK that used deepfake to have David Beckham delivering an antimalaria message in nine languages last year.

Nvidia a computer company used deep fake technology to create graphics for video games and they have also announced a new product called *Maxine.* This product is made to help with video calls. they are using Face alignment, so it looks like you are staring at the person on the screen like you are having a normal conversation,

The technology of deep fakes will also have beneficial impacts on other areas such as the field of medicine. A professor has predicted that the use of deep learning can be used to synthesize realistic data that will help researchers develop new ways of treating diseases without actual patient data.

SOLUTION

The solutions to deepfakes are difficult to explain since there are not any specific solutions out there, but researchers have managed to come through with something that can be considered as a set of solutions to deep fakes. These two solutions are the use of tech to detect fake videos and to improve media literacy

Tech solutions can be explained by trying to detect deepfakes using similar AI that was used when it was developed. Some researchers found out that analyzing the blinks in videos could be one way to detect a deepfake from an unaltered video. This is because there are not many photographs of celebrities blinking

Another solution to this problem could be to increase media literacy among vast swathes of the population. What this does is that they can spot "fake news" when it is published. How to achieve this is a mystery.

GRAY AREAS

When it comes down to whether it is okay to use deepfake or not is debatable because you never know for sure if deepfakes are used politically right. This can lead to some gray areas of this technology.

With digitalization moving rapidly, laws and verdicts are being applied to protect individual's privacy. We see this with GDPR and the Screms II verdicts in the EU that have developed as a byproduct of digitalization in the world. Active consent by each individual is being focused and taken seriously, which makes it safer and more comfortable for the general users to trust the development of new technology. However, this does not stop the bad guys from taking advantage of the loopholes. This can therefore develop further laws that can protect the privacy, identity and reputation of people.

News is on our faces all the time. Fake news and real news. At this point of time, it is hard to distinguish between what is real and what is fake. It is very clear that the law against fake news needs to develop further to protect people, but with the technology developing at such a fast pace, it is nearly impossible for the politicians to stay ahead of it or on a par with it.

SUMMARY

In conclusion we can see that deepfake has both positives and negatives sides. The rise of deepfake is inevitable and this may become a problem if it is used in the wrong hands. Many people are already afraid of its power and there will certainly be a time where we cannot see the difference between real or fake anymore. The technology is dangerous but would be an amazing asset in the right hands.

The problem lies in the attitude and expectations from the people. We cannot stop the development of new technology. However, it is possible to educate people to get a healthier view of the topics.

SOURCES:

intro:

what's an ai:

https://www2.deloitte.com/no/no/pages/technology/articles/tre-ting-vite-kunstig-intelligensai.html

Where did deepfakes come from and who created it?

https://www.businessinsider.in/tech/news/chris-um-who-created-tom-cruise-deepfake-videoson-tiktok-now-owns-an-ai-company/articleshow/85170929.cms

https://en.wikipedia.org/wiki/Deepfake

https://www.youtube.com/watch?v=fOxgk9JCDq4

Problems/affection/controversies:

https://www.youtube.com/watch?v=ENdRAI-JrDc

https://www.technologyreview.com/2020/12/24/1015380/best-ai-deepfakes-of-2020/

Solution:

https://cybersecurity.att.com/blogs/security-essentials/deepfakes-are-a-problem-whats-thesolution

https://boingboing.net/2020/10/09/nvidia-has-a-plan-to-use-ai-to-deepfake-your-face-to-lookthe-right-way-on-video-calls.html

https://www.toptal.com/machine-learning/generative-adversarial-networks

https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positivefor-humanity/?sh=237c0bf52f84

https://www.roydswithyking.com/deepfakes-and-the-law-what-can-you-do-if-your-faceappears-in-a-deepfake-video/