UTeM

اونيۏرسيتي تيکنيکل مليسيا ملاک

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

FACULTY OF INFORMATION TECHNOLOGY AND COMMUNICATION (FTMK)

BITI 1113 ARTIFICIAL INTELLIGENCE

SEMESTER 2 2019/2020

DEEPFAKES (FINAL REPORT)

LECTURER'S NAME: ASSOC. PROF. DR. AZAH KAMILAH DRAMAN @ MUDA

NAME	MATRIC NO
MOHAMAD ZAIRI BIN ABD GHANI	B031910029
NURUL ANISA SHAHIRAH BINTI MAT ROPI	B031910103
ANISAH BINTI KAMSIN	B031910156
JEYA JASSVINE A/P JEYABALA SUNDRAM	B031910136

1.0 INTRODUCTION TO DEEPFAKES

Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. While the act of faking content is not new, deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content with a highpotential to deceive. The main machine learning methods used to create deepfakes are based on deep learning and involve training generative neural network architectures, such as autoencoders or generative adversarial Networks.

Deepfakes have garnered widespread attention for their uses in celebrity pornographic videos, revenge porn, fake news, hoaxes, and financial fraud. This has elicited responses from both industry and government to detect and limit their use.

Photo manipulation was developed in the 19th century and soon applied to motion pictures. Technology steadily improved during the 20th century, and more quickly with digital video. Deepfake technology has been developed by researchers at academic institutions beginning in the 1990s, and later by amateurs in online communities. More recently the methods have been adopted by industry.

Academic research related to deepfakes lies predominantly within the field of computer vision, a subfield of computer science. An early landmark project was the Video Rewrite program, published in 1997, which modified existing video footage ofa person speaking to depict that person mouthing the words contained in a different audio track. It was the first system to fully automate this kind of facial reanimation, and it did so using machine learning techniques to make connections between the sounds produced by a video's subject and the shape of the subject's face.

Contemporary academic projects have focused on creating more realistic videos and on improving techniques. The “Synthesizing Obama” program, published in 2017, modifies video footage of former president Barack Obama to depict him mouthing the words contained in a separate audio track. The project lists as a main research contribution its photorealistic technique for synthesizing mouth shapes from audio. The Face2Face program, published in 2016, modifies video footage of a person's face to depict them mimicking the facial expressions of another person in real time. The project lists as a main research contribution the first method for re-enacting facial expressions in real time using a camera that does not capture depth, making it possible for the technique to be performed using common consumer cameras.

In August 2018, researchers at the University of California, Berkeley published a paper introducing a fake dancing app that can create the impression of masterful dancing ability using AI. This project expands the application of deepfakes to the entire body; previous works focused on the head or parts of the face.

A traditional example of what was considered a deepfake back then is an example of modifying lip motions of a person speaking using different audio clips. This is done by making connections between the sounds of the audio track and the shape of the subjects face. However things have rapidly evolved and it has now become increasingly easy to manipulate a real face of one person into an image or video. This is because of the accessibility to a huge number of public data, evolution of deep learning techniques. Several apps such as ZAO and FaceApp has been the key to create fake images and videos even when you have no expertise.

There are many types of manipulations in deepkfake but according to a leading survey, the top 4 manipulations are entire face synthesis, identity swap, attribute manipulation and expression swap.

For entire face synthesis, this manipulation usually creates entire non-existent face images through GAN. This technique results in very realistic and high quality facial images. This sort of techniques can highly benefit video gaming and 3D modelling industries.

For identity swap, this technique consists of replacing the face of one person in a video with the face of the other person. There are two different methods to this whereby the first one is classical computer graphics (eg : FaceSwap) or novel deep learning techniques known as Deepfakes (ZAO application). These could benefit movie sectors highly.

As for attribute manipulation, this is basically editing or retouching the face like the hair colour or skin colour, gender and age and so on. This manipulating process ussually uses GAN (eg : StarGAN).

And as for expression swap, the facial reenactment occurs consisting of modifying only facial expressions of a person. An example of this was the famous video of Mark Zuckerberg saying things he never said.

These types of manipulation have recieved the most attention in the last couple of years. There are certain techniques that are equally dangerous but isn't too famous. For example, face morphing.

2.0 PROBLEM BACKGROUND

Deepfake technology has understandably become notorious in the threat deepfakes seemingly pose to politics and others. However, the ability to generate realistic simulations using artificial intelligence will, on the whole, be only a positive for humanity.

Increasingly, new uses are being found for deepfakes. Good uses. Whether recreating long-dead artists in museums or editing video without the need for a reshoot, deepfake technology will allow us to experience things that no longer exist, or that have never existed.

For example, digital face editing is something that is heavily employed in the movie industry. Digital de-aging, for instance, is particularly popular: Samuel L. Jackson and Clark Gregg have been digitally de-aged by approximately 25 years in Captain Marvel.

Aside from having numerous applications in entertainment and education, it's being increasingly used in healthcare and other areas.

For example, the development of deep generative models raises new possibilities in healthcare, where we are rightly concerned about protecting the privacy of patients in treatment and ongoing research. With large amounts of real, digital patient data, a single hospital with adequate computational power could create an entirely imaginary population of virtual patients, removing the need to share the data of real patients.

We would also like to see advances in AI lead to new and more efficient ways of diagnosing and treating illness in individuals and populations. The technology could enable researchers to generate true-to-life data to develop and test new ways of diagnosing or monitoring disease without risking breaches in real patient privacy.

Deepfake-related technology could also be used in the near future to dub actors in other languages, as it happened in the video in which David Beckham delivered his malaria awareness message in nine languages thanks to AI.

Both the gaming and fashion industries might also benefit from this, allowing players and customers to create custom avatars with their actual faces.

AI could also be used to help patients who are struggling to accept their body image, allowing them to see a version of their body they are more comfortable with.

There are so many positive ways in which deepfakes could be used to make our life better - and that is why this is a technology that needs to be understood, not just feared.

3.0 OBJECTIVES

Deepfakes are videos that have been constructed to make a person appear to say or do something that they never said or did. With Artificial Intelligence-based methods for creating deepfakes becoming increasingly sophisticated and accessible, deepfakes are raising a set of challenging policy, technology, and legal issues.

Deepfake content is created by using two competing AI algorithms. One is called the generator and the other is called the discriminator. The generator, which creates the phoney multimedia content, asks the discriminator to determine whether the content is real or artificial.

Deepfakes are also being used to place people in pornographic videos that they in fact had no part in filming. However, it would be just as easy to create a deepfake of an emergency alert warning an attack was imminent, or destroy someone's marriage with a fake sex video, or disrupt a close election by dropping a fake video or audio recording of one of the candidates days before voting starts.

While the act of faking content is not new, deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content with a high potential to deceive. The main machine learning methods used to create deepfakes are based on deep learning and involve training generative neural network architectures, such as autoencoders or generative adversarial networks (GANs).

The first step in establishing a GAN is to identify the desired output and create a training dataset for the generator. Once the generator begins creating an acceptable level of output, video clips can be fed to the discriminator.

As the generator gets better at creating fake video clips, the discriminator gets better at spotting them. Conversely, as the discriminator gets better at spotting fake video, the generator gets better at creating them.

4.0 TECHNIQUES USED IN DEEPFAKE

There are many techniques involved in creating a deepfake. Much of the Artificial Intelligence knowledge is needed to create a realistic deepfake. In this report, we will explain further regarding the technical background.

Firstly, we must understand what a neural network is. A neural network is a non linear model that is used for predicting or generating an output based on the input. This structure is made up of neurons and each layer is connected in sequence through synapses. The synapses are associated with weights, which affects how much of the forward propagation goes through the neural network. The weights can be altered when it goes through the back propagation. Neural networks are very important because it can be used to predict data of a new and unknown output. In simpler words, an image X is trained and when an image Y is the new input, the software can still detect it exactly the same way it detected image X to produce the output.

Next, to create a deepfake we must go through loss function technique. This is a way of evaluating how specific the algorithm models a specific data. The two most common types of loss functions are the L1 and L2 loss functions and the perceptual loss technique. The L1 is the sum of absolute differences between true value and predicted value. The L2 loss function is the squared differences between the true and predicted values. A formula is shown below :

$$L1=|x-xg|1andL2=|x-xg|2$$

However, when using this technique, you would need paired images. For example, image A and B with the same expressions. When you want to compare two images that aligned, we can use the perceptual loss technique. This function works by summing up all the squared errors between all the pixels and the taking the overall mean. Most soruces say that this technique is faster and generates a higher quality.

Next technique involves Generative Neural Network. In this main technique there will be many sub-techniques as there are many types of neural networks but in deepfakes the common few used are these six neural networks.

Firstly, the Encoder-Decoder Networks (ED). This network consists of at least two networks which are the Encoder and Decoder. The Encoder will compress images into small amount of data and Decoder will decompress the data back into the image.These two parts combine to form a autoencoder. In deepfakes, two autoencoders are used. One would be used on the face of the target and one would be for the face of the actor. The software will swap the inputs and outputs of the two autoencoders to transfer the facial movements and features of the actor onto the target.

Next is the Convulational Neural Network (CNN). CNN is a class of deep neural networks, commonly used for imagery. CNN works pretty similar to the traditional neural network (with weights) . Whilst this process, the weights will be learned during the training process and this will help the network learn what type of features to be extracted from the input. The CNN works with filters, pooling and up-sampling. Filters are sets of weights which are learned using the backward propagation technique. Pooling will reduce network dimensions and network gets deeper while up-sampling does the opposite.

Next, Generative Adversarial Networks (GAN). A GAN is a very interesting concept whereby it consists of two neural networks competing against each other, just like in a two player video game. For example, generator A and discriminator B. A will create fake samples to fool B but B will try to figure out between the real and fake product. This will cause A and B to go through training and eventually A will learn how to create fake samples that are not distinguishable by B. And once A is able to do so, B will be thrown away and A is used to produce content. When this is applied in real life, like in imagery, the output is a very realistic output that cannot be distinguished.

An improved version of GNN has also led to Image-to-Image translation(pix2pix). This network is made up of 2 parts which is the Generator and the Discriminator. The Generator will transform the input image to get an output. The discriminator will measure between the similarity of an input image to an unknown image (could be target image or input image from generator) and tries to guess if it was produced by the generator.

So basically: Discriminator = (x , input image) and/or (generator image, input image)

Next, is CycleGAN which is an improved version of the pix2pix. This enables image translations through unpaired trainings. The models are trained in an unsupervised manner using collection of images from the source and target domain that are not related in any way.

The Recurrent Neural Network (RNN) is a type of neural network that can handle sequential and variable length. RNN remembers the past and its decisions. It is different from other networks because, although the other networks “remember” things too, they don’t remember the past they only remember during the training while generating outputs. RNN is deepfakes handle mostly audios and sometimes videos.

After the network techniques another important technique used in deepfakes are the feature representations. To carry this out, there are many ways.

a) Facial action coding system - measure each of the face units

b) Monocular reconstruction - obtain a 3D morphable model of the head from a 2D image, where the pose and expressions are extracted and calculated by a set of vectors and matrices.

c) Image segmentation - help network seperate different concepts/parts (eg ; eyes, nose, mouth). The most common representations are defined positions on the face or body that can be easily tracked. These landmarks are presented to the layer in 2D with points at each landmark. Then the network will identify and associate them. However for audio(speech), the common approach is to measure the segments of the audio, the dominant voice frequencies.

As a summary, let's look into the deepfake creation basics. This will give a clear understanding since we have already explored all the techniques used in detail.

The first step to creating a deepfake is the video collection where there would be two types of video needed. One is the source video and the other is the target video (with the actor). The next step would be to carry out frame extraction of the video. An example of one-minute video can compile up to 3000 picture frames. Next, facial extraction. From the picture frames accumulated earlier, we have to extract the facial features and expressions (for both videos). Next we would carry out masking, or region of interest (Only on TARGET video). In lay mans terms, this is the process where we get rid of the victims specific facial features and replace with the actors facial features. Next, it is the training process and followed by compiling process to produce a realistic deepfake video.
This is a very simplified version of how we can create a deepfake and every single one of these steps require detailed techniques related to Machine Learning which we have just discussed about.

5.0 TEAM MEMBER'S ROLE

- Leader: Mohamad Zairi bin Abd Ghani
- Co-leader: Nurul Anisa Shahirah binti Mat Ropi
- Editor: Anisah binti Kamsin
- Co-Editor: Jeya Jassvine A/P Jeyabala Sundram

6.0 LINK TO PROJECT VIDEO IN YOUTUBE

- https://youtu.be/SdJQPJCQmLQ