DETECTING AND COMBATING DEEP FAKES

Date: November 24th, 2020

Disclaimer: This briefing note contains the encapsulation of views presented by the speaker and does not exclusively represent the views of the Canadian Association for Security and Intelligence Studies.

KEY EVENTS

On November 24, 2020, Dr. Hany Farid presented *Detecting and Combating Deep Fakes* at the 2020 CASIS West Coast Security Conference. The presentation was followed by a group panel for questions and answers. Main discussion topics included deep fakes, their contribution to misinformation, and the challenges in detecting them.

NATURE OF DISCUSSION

Presentation

Dr. Hany Farid discussed the different kinds of deep fakes, how they are created, and how they are being used for misinformation, fraud, and to damage people's reputation. The speaker also examined some of the techniques being used to detect deep fakes and the biggest challenge presented as a result of the spread of deep fakes and misinformation.

Question Period

During the question and answer session, the speaker discussed the constantly evolving nature of deep fakes techniques, the problem with the Good Samaritan law, and the inconspicuous problem of the liar's dividend.

BACKGROUND

Presentation

The manipulation of media goes back as far as the early days of photography when the ability to manipulate photographs was only in the hands of the few, but now in the digital age, the average person can easily manipulate images. People are selling these images to create false personas on the internet. Fake images like

the ones presented in thispersondoesnotexist.com are 100% synthesized by a machine learning algorithm, and they are highly realistic images encompassing different genders, races, ages, facial hair, and glasses.

Machine learning, computing power, and large datasets have furthered misinformation. Recently, a 17-year-old high school student created a fictional character to run for Congress, and it was accredited by the election board and the Twitter account also got the coveted blue checkmark. There have been other times where fake personas are created to try to give more credibility to a story.

These images are created by slapping down many pixels and modifying them until they look like a person. The images are the result of a back and forth process that is repeated millions of times between a generator and a discriminator supported by a deep neural network, which is the underlying computational machinery of this so-called generative adversarial network (GAN). This system can create anything; they can be faces, cars, houses, cats, airplanes, etc.

This GAN structure also works with videos, where you can change the identity of one person to another one. After enough iterations, one face can be replaced with another one. You start to see the power of this type of technology when you realize that you can change somebody's identity and get them to say whatever you want. A second type of deep fake is the so-called puppet master, where a single static image of a person can get animated by another person moving and talking in front of a webcam. This person is the puppet master, and the static image becomes the puppet. The third type of deep fake video is a lip sync debate, in which only the mouth is modified to be consistent with a new audio track.

A commonality between all these three deep fakes is that they are making people say and do things that they never did. The most common application of deep fake technology is non-consensual pornography, which is used as a weapon to damage people’s reputation. Misinformation and deep fakes form a very complex landscape, where you cannot trust what you read, see, or hear. When you cannot trust images, videos, or audios, it is very difficult to make thoughtful, rational, and reasoned decisions.

The use of deep fakes to commit fraud is already starting to surface, and unfortunately, by releasing false information with very compelling video and audio, this kind of technology has the potential to create detrimental results in the stock market. It would take a very short time to move millions of dollars in the market, while figuring out that it is a deep fake would take much longer.

One technique to detect deep fakes is soft biometrics, which is not like DNA or fingerprinting to distinguish someone from the other seven billion people, but it is enough to distinguish someone from someone else trying to impersonate the other person. This works by analyzing hours of video and then extracting behavioral mannerisms to see if the person is moving the way he or she is expected to move. A second technique identifies structural mistakes in lip sync deep fakes. It looks at the relationship between phonemes and visemes. In authentic videos the viseme and the phoneme match (e.g., pronouncing the letter M while also closing the mouth).

Every three months, there are advances in the creation of deep fakes, but the problem is not only the creation of deep fakes, but also the unprecedented scale and speed to distribute that information through social media platforms. These platforms seem eager to promote outrageous, hateful, divisive, and conspiratorial material only because it is good for business. Furthermore, these days, the public is particularly polarized, but the problem is not just the creation, distribution, or polarization, it's the combination of the three.

One of the biggest challenges that we will be facing is the so-called liar's dividend. Once we enter a world where any new story, image, video, or audio can be fake, nothing will have to be real, and this is going to be very dangerous because we will go into our own little ecosystem where we will only believe what we want. We are dealing with an arms race because as we develop new technology to detect deep fakes, people develop better fake technology. We need to develop better technology to detect this type of fake information, and we have to start thinking about how we educate the next generation of digital citizens so that they know how to get trusted information online.

Question Period

- Deep fakes techniques are constantly evolving and have a shelf life of usually a couple of years. We get some years out of a technology until it catches up. The spam filters of ten years ago, for example, don't work anymore, but we now have better spam filters and virus detection. New techniques are always being developed, and although academics publish their scientific findings, they do not release the data or the code; the problem is not eliminated but it creates some road bumps.

- There should be provisions on Section 230 (the Good Samaritan law) of the Communications Decency Act, which right now gives the technology sector an almost unprecedented protection from lawsuits for either removing

content or for not removing content. An example of how problematic this can be is the case of the Backpage website, which was knowingly trafficking under aged girls, but because of the Good Samaritan law, they got protection from the courts. Unfortunately, the whole discussion of Section 230 and the tech sector has become highly politicized, but we need to start having a sensible discussion about how we can regulate the technology sector in a way that we can keep the things that we like about it and at the same time deal with some of the harmful aspects of it.

- A perceptual study at UC Berkeley has shown that the average person has now a better chance to detect fake static images, although for videos we are not there yet; there are still some challenges. It is very easy to create fake videos from largely stationary videos, which do not have a lot of complexity from other movements. But whether the technology is really there or not, it doesn't matter if you're going to use the liar's dividend. For example, when George Floyd was killed and caught on camera, many people chose to believe that the video was a deep fake.

KEY POINTS OF DISCUSSION

Presentation

- Advances of the digital age have allowed the average person to manipulate images as well as systems to create images of people who don't exist.
- The generative adversarial network not only can create fake images, but by manipulating pixels, it can also change the identity of a person, create a puppet out of a static image, and create fake lip sync debates; all of it in videos.
- Soft biometrics can help distinguish the identity of a person in a video, and the analysis of lip sync deep fakes can also identify structural mistakes between phonemes and visemes.

Question Period

- Deep fakes and the techniques to detect them are constantly evolving; we may not eliminate the problem, but as we continue getting better, we also create some setbacks for those trying to do some harm.
- Section 230 of the Communications Decency Act should have some provisions in place to allow us to keep the things we like about technology and to take action against the harmful aspects of it.

- It is becoming easier for the average person to distinguish real images from synthesized images; however, regardless of how far behind technology may be in creating something, some people will still choose to believe it is all fake.

Published by the Journal of Intelligence, Conflict, and Warfare and Simon Fraser University

Available from: https://jicw.org/