Hal Stewart

31 October 2021

DMC 435

The Truth Behind Deepfakes

When most people are scrolling through social media whether it is Twitter, Facebook, Instagram, or whatever platform of choice, there is some hesitation on the 'accuracy' of what is posted. However, what if what people were reading was not real at all? What if the pictures or videos they saw claiming or showing something were not real? How would people be able to discern what to believe on the Internet? That is the issue with Deepfakes. Deepfakes are any kind of synthetic media where a person in an image or video is swapped with another's likeness (Somers). However, it is not just with images or videos now there is new technology now 'Deepfaking' Tweets and other text.

The technology behind Deepfakes is ever changing using different Machine Learning (ML) techniques. The original technology behind Deepfakes was used for entertainment purposes. One of the more prominent recent examples is Paul Walker in the *Fast and Furious* movies. The special effects team used Walker's brothers and CGI to place his likeness over theirs (Caufield). This technology was used back in 2015, and since then even more evolution has occurred. The combination of facial swapping and deep learning has proven to be the surge in the advancements in realistic Deepfakes. These advancements have led big tech companies like Facebook and Google and many national governments to try to crack the code on Deepfakes and prove which images are real or doctored.

After analyzing the technology and ML techniques behind Deepfakes, the question becomes: Is all the technology behind Deepfakes so dangerous that everyday people will be

affected by their usage? Is there really a way to 'crack' a Deepfake? Do the moral and legal implications of Deepfakes truly outway the benefits the technology Deepfakes could bring to the table? There are many ways to look at the issues and benefits of Deepfakes to provide some avenue to navigate these questions. Analyzing the progress on Deepfakes and the ML behind it shows that there is some real progress being made with ML that can be used in other fields, but there are also inherent dangers in Deepfaking. Therefore, there are both positive and negative effects into Deepfakes that need to be weighed against the moral and legal implications of the technology.

Unfortunately, pinpointing the exact start time of 'Deepfakes' is very difficult to do. The technology behind Deepfakes was originally used in the entertainment sector. Singh writes, "For animation films and influential science fiction films, deepfake can provide realistic output" (Singh 1). Back then, Deepfake technology had not progressed enough to make realistic images, so the only way to implement the concepts were in non realistic settings. ML techniques progressed rapidly throughout the decade with the implementation of deep learning algorithms like GANs and CNNs, but the internet had not fully dived into what the world would know to be 'Deepfakes'.

The rise of Deepfakes do have their origins on the internet, specifically the website Reddit. Somers states, "The term "deepfake" was first coined in late 2017 by a Reddit user of the same name. This user created a space on the online news and aggregation site, where they shared pornographic videos that used open source face-swapping technology" (Somers). From this point forward, the negative connotation of Deepfakes began to arise. A prime example of the harmfulness of Deepfakes is in the political field. Westerlund acknowledges, "In the political scene, a 2018 deepfake created by Hollywood filmmaker Jordan Peele featured former US

President Obama discussing the dangers of fake news and mocking the current president Trump" (Westerlund). Examples like the one above became prominent on the internet after the 2017 Reddit discussion.

Because of this negative connotation and other technology that is very similar many people now have shifted away from "Deepfakes" as the overall term for what the technology is. One of the more prominent terms being used instead of Deepfake is: *artificial intelligence-generated synthetic media* (Somers). This term includes everything that is synthetically made. Somers states these to be, "broad enough to include the original definition of deepfake, but also specific enough to omit things like computer generated images from movies, or photoshopped images — both of which are technically examples of something that's been modified" (Somers). The history and background of Deepfakes all have to do with machine learning, and the technology of Deepfakes is very accessible to anyone with a coding or mathematical background.

There are a few ways machine learning techniques have affected the growth of Deepfakes. Many of these techniques are from the concepts of deep learning. IBM states, "Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain" ... "allowing it to "learn" from large amounts of data" (IBM Cloud Education). Deep learning has allowed the growth of neural networks which is the backbone of Deepfakes. Some of the more popular neural networks to create Deepfakes are Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Deep Convolutional Generative Adversarial Networks (DCGANs), and Variational Autoencoders (VAEs). Though any neural network can make a

Deepfake, these are the more popular architecturally to make very accurate Deepfakes. The two most commonly used are GANs.

GANs are one of the more popular infrastructures. A GAN or Generative Adversarial Network is a network that pits two models: a discriminator and generator against each other to have the generator create a fake image or text and then the discriminator would try to see if the image or text is real or not. Spivak writes, "Discriminative algorithms try to classify input data; that is, given the features of a data instance, they predict a label or category to which that data belongs"(Spivak 342). He also writes, "...a generative model provides a way to *generate* data that looks like it came from the dataset. Instead of predicting a label given certain features, it attempts to predict features given a certain label"(Spivak 343). The breakdown of these two algorithms are the backbone for the GAN. The picture below demonstrates how GANs function.

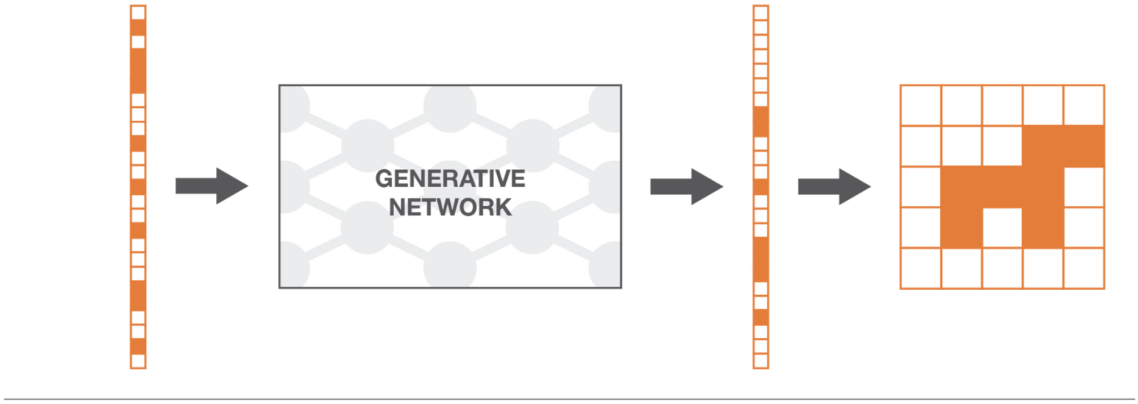

Illustration of the notion of generative models using neural networks. Obviously, the dimensionality we are really talking about are much higher than represented here.

Rocca, Joseph. “Understanding Generative Adversarial Networks (Gans).” *Medium*, Towards Data Science, 21 Mar. 2021,

Deepfakes use this process with images or videos. The generator will create an image or video and the discriminator will try to discern if it’s real or not. If the discriminator cannot, there is a good chance depending on how well the GAN is made that humans would not be able to either. Creating a powerful enough GAN takes a lot of processing power and time to properly train the network, but it is very doable.

As a result of the ease of making a Deepfake, the consequences of this has led to the emergence of threats using Deepfakes. Some of the more prevalent threats are blackmail, political, and cybersecurity issues for people and organizations. One of the more obvious forms of blackmail Deepfakes can use is revenge pornography. Unfortunately, many people who are targeted with this form of Deepfakes are women, and many of them do not even know that these images are out there of them. CBS News writes, “According to a 2019 report by the cybersecurity company Deeptrace, 96% of all deepfakes online are pornographic and the top five deepfake pornography websites exclusively target women”(Sherman). This means many women are the target of digital blackmail without them even knowing.

Another example of Deepfake threats are in the political sphere. An example of a political Deepfake is:

> “In a 2018 deepfake video, Donald Trump offered advice to the people of Belgium about climate change. The video was created by a Belgian political party “sp.a” in order to attract people to sign an online petition calling on the Belgian

> government to take more urgent climate action. The video provoked outrage about the American president meddling in a foreign country with Belgium's climate policy" (Westfield).

This example of a political Deepfake caused a lot of uproar, and it could become even worse using the likeness of other controversial political figures. The final more prevalent threat is cybersecurity. An example of these cybersecurity concerns comes in the forms of scams. In 2019, A CEO for an U.K. energy company sent over $250,000 dollars because of an audio deepfake (Somers). Another kind of cybersecurity threat combines both the political sphere and the everyday public. Johansen states, "In Gabon, a deepfake video led to an attempted military coup in the East African nation" (Johansen). This kind of Deepfake was able to trick an entire military which could have been disastrous knowing that the video was faked. These kinds of Deepfakes can scam not only companies but everyday people out of money, and the cybersecurity detection could not know otherwise.

Despite these Deepfake threats on entities and everyday citizens, there are a few ways to combat Deepfakes. One of the major ways is with technology. There are currently some technological breakthroughs in detecting Deepfakes. Right now one of the best ways to detect Deepfakes is analyzing the image. Westerlund states, "Media forensic experts have suggested subtle indicators to detect deepfakes, including a range of imperfections such as face wobble, shimmer and distortion; waviness in a person's movements..." (Westerlund). These imperfections are a telltale sign that the image or video could be a Deepfake. However, technology to fix these imperfections is getting better and better. Some of the other technological improvements are adding digital watermarks into images, videos, and audio to authenticate the medium from a

Deepfake. Spivak writes, “For example, Canon's Original Data Security Kit ‘enhances security by providing image data encryption and decryption features in addition to a verification function that authenticates image originality.’” (Spivak 353). This is a good way to try to combat the manipulation of new media, but it would not be the best way for creating a Deepfake from preexisting media. Another big way to combat Deepfakes is with legislation. Some states in the US have already passed some laws to protect citizens who could be at risk of Deepfake attacks. Johansen states, “Virginia was the first state to impose criminal penalties on nonconsensual deepfake pornography” (Johansen). Since there are some laws put into place maybe cyber criminals will think next time about committing crimes. There are some of the few ways Deepfakes are being combated.

Nevertheless, there are some benefits to Deepfakes. The film industry is one of the many industries that uses the technologies in Deepfakes all the time. For any movie where actors need to be ‘de-aged’, that is Deepfake technology. Another industry that is using Deepfake technology is the medical field. Westerlund writes, “Scientists are also exploring the use of GANs to detect abnormalities in X-rays and their potential in creating virtual chemical molecules to speed up materials science and medical discoveries” (Westerlund). GANs are the backbone of many Deepfakes and without the research into that technology there would be very little way to help machine learning techniques grow. There are also some equity and accessibility benefits of Deepfake technology. Jaimen states, “VOCALiD leverages voicebank and proprietary voice blending technology to create unique vocal personas for any device that turns text into speech for those with speech and hearing difficulties" (Jaimen). There are many ways that machine learning technology can be very beneficial.

Overall, Deepfakes are quite the powerful machine learning tool to use. Deepfakes may be a relatively new concept in machine learning, but it has proven to be quite the powerful tool. The machine learning technology behind Deepfakes is relatively easy to code with a basic understanding of Tensorflow and Python libraries. GANs and other techniques just require huge amounts of processing power to make believable Deepfakes. However, there are many dangers behind Deepfakes. Deepfakes have caused a lot of harm to people, and until better detection techniques or laws are put into place Deepfakes may be more of a problem. With that being said there are many benefits to the technology behind Deepfakes. Stopping machine learning and AI from growing over one bad technique would prevent the growth of the importance of both of those fields. Therefore, there are both positives and negatives to Deepfakes, but the importance of machine learning and the technology behind Deepfakes can really benefit many people.

Works Cited

Caulfield, AJ. “The Truth about Recreating Paul Walker for Fast and the Furious - Exclusive.” *Looper.com*, Looper, 11 Mar. 2020, www.looper.com/184468/the-truth-about-recreating-paul-walker-for-fast-and-the-furious/

IBM Cloud Education. “What Is Deep Learning?” *IBM*, 1 May 2020, www.ibm.com/cloud/learn/deep-learning.

Jaiman, Ashish. “Positive Use Cases of Deepfakes.” *Medium*, Towards Data Science, 5 Oct. 2021, towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387.

Johansen , Alison. “Deepfakes: What They Are and Why They're Threatening.” *NortonLifeLock*, 2020, us.norton.com/internetsecurity-emerging-threats-what-are-deepfakes.html.

Rocca, Joseph. “Understanding Generative Adversarial Networks (Gans).” *Medium*, Towards Data Science, 21 Mar. 2021, towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29.

Sherman, Justin. “‘Completely Horrifying, Dehumanizing, Degrading’: One Woman's Fight against Deepfake Porn.” *CBS News*, CBS Interactive, 14 Oct. 2021, www.cbsnews.com/news/deepfake-porn-woman-fights-online-abuse-cbsn-originals/.

Singh, Harshpreet. “Deepfake Detection Research Project - NORMA@NCI Library.” *Norma* , 2020, norma.ncirl.ie/4473/1/harshpreetsingh.pdf.

Somers, Meredith. “Deepfakes, Explained.” *MIT Sloan*, 21 July 2020, mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained.

Spivak, Russell. ""Deepfakes": The Newest Way to Commit One of the Oldest Crimes." *Georgetown Law Technology Review,* vol. 3, no. 2, Spring 2019, p. 339-401. *HeinOnline*, https://heinonline.org/HOL/P?h=hein.journals/gtltr3&i=354.

Westerlund, Mika. “The Emergence of Deepfake Technology: A Review.” *Technology Innovation Management Review*, Nov. 2019, timreview.ca/article/1282.