Zachary Blake
Mobile App Dev II
Haim Levkowitz

Deepfakes and Public Misinformation

Abstract

The realm of machine learning has made many advancements in the past few years. The general use of machine learning is to predict a dependent variable, based on an input example data set. Based on this training data, machine learning algorithms can make predictions on whether an input has this dependent variable or not. For example, an algorithm trained on multiple species of Iris' can make predictions on what species a new input Iris is. Eventually, this evolved into machine learning algorithms that could generate new data based on training input data. This eventually resulted in "Deepfakes" which are generated images, video, and audio of real or fake (generated) people. These Deepfakes aim to create misleading or fake content from existing data on people. Deepfakes use powerful machine learning techniques such as generative adversarial networks (GANs) to create life-like recreations of real people. Deepfakes have recently been used in media, such as movies or propaganda.

What are Deepfakes?

The word Deepfake comes from a portmanteau of "Deep Learning" and fake. Deepfakes are generated, realistic depictions of people. The generated data can constitute a person's voice, facial expressions, and inflections (Westerlund, 2019). Deepfake videos use facial remapping technology to switch subject faces to create misleading video. The main use of a Deepfake is misinformation, where the Deepfake will try to mislead the watcher into believing the video is real. The term originates from a *Reddit* user named *deepfakes* who uploaded Deepfaked pornography of actors face-swapped with

celebrities, circa 2017 (Westerlund, 2019). While not all of the Deepfakes were explicit in content, it created an idea of misleading images, photos, and videos that had the potential to be made.

How are they used?

The avenues used are typically social media platforms, where the videos are shared and seen by many thousands of people (Westerlund, 2019). The videos are shared and gain public traction through the scope of social media. One example of this was a recently created DeepFake of Ukrainian President Volodymyr Zelenskyy (Allyn, 2022). The war between Russia and Ukraine consisted of a Russian advance into taking Ukrainian territory beginning in February 2022. The 2022 video depicts the Ukrainian President telling his country's soldiers to lay down their weapons and surrender to the Russian advance. The creator of the video is unknown, but is suspected to be a propaganda tool created by Russia to allow an easier advancement into Ukrainian territory. The vector of transmission was Facebook, YouTube, and Twitter, social media outlets where the video was shared thousands of times. Upon closer inspection, the Deepfake had issues, such as an imperfect accent, and odd facial movements / features (Allyn, 2022). The fake video was also propagated by hackers on Ukrainian television with a news ticker on the bottom.

Another, less insidious example would be the Deepfake recreation of Carrie Fisher in the movie *Rogue One: A Star Wars Story.* The recreation takes a generated version of Carrie, around the age of when she first filmed *Star Wars: Episode IV – A New Hope,* and swaps it with an actor in *Rogue One.* The Deepfake consists of a small cameo (about 15 seconds) at the end of the movie where Carrie Fisher's face is swapped with actor Ingvild Deila. Deila was chosen based on a similar height, body type, and profile to Carrie Fisher (Winick, 2020). Carrie Fisher was shown the result and approved of the Deepfake. While this Deepfake was created before Carrie Fisher died in 2016, other Deepfakes have been created posthumous of actors.

How do Deepfakes work?

Deepfakes utilize generative adversarial networks (GANs) which consist of two neural networks in a contest with each other. One neural network is selected as "the generator" which creates the data. Another network is selected as "the discriminator" evaluates the generated data and picks out what it thinks is fake (Westerlund, 2019). This model pushes the algorithm forward with each iteration, and is an *unsupervised* machine learning algorithm. GANs can scrub thousands of photos of a person, then generate a new image without being an exact copy of any of the training photos (Westerlund, 2019).

The adversarial part of GAN describes the direct relationship of the generator and discriminator. You can liken a generator to a counterfeiter producing fake currency (University of Montreal, 2015). As a result, the discriminator acts like the police, trying to expose the fake currency. The method auto-updates with each generated sample. Below is an example of an improving GAN.

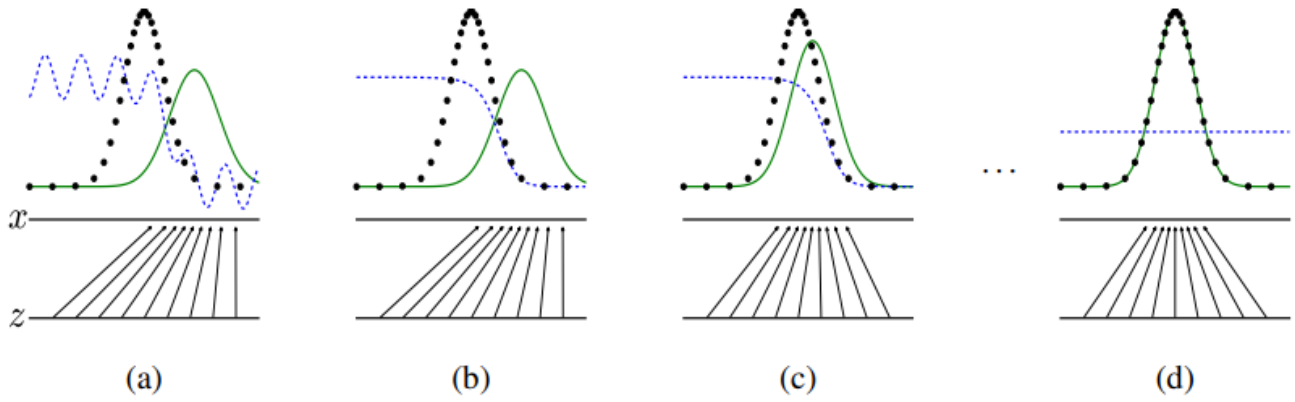

Figure 1: GANs trained by updating the discriminative distribution (blue, dashed line), the data generating distribution (black, dotted line), and the generative distribution (green, solid). It is run from (a) to (d) until it cannot be improved any more.

The algorithm is run until the discriminator and generator have enough capacity, at which point either cannot be improved further (University of Montreal, 2015). The unsupervised tendency of GANs lend itself to a generative format (Deng, Yu, 2014).

Conclusion and the Future of Deepfakes

In conclusion, Deepfakes are a fascinating use of machine learning algorithms designed to generate fake data. The Deepfakes themselves are not limited to people and can generate things like animals, or objects from training images. Deepfaked art exists in this capacity already, with paintings being generated by neural networks.

The utilization of machine learning algorithms have produced Deepfakes and could create other generative, unsupervised media. Could a Deepfake trained on thousands of movies create a coherent movie itself? What's clear is the environment of machine learning is creating emergent technologies

The realistic depictions of people grow stronger with each passing day, making it harder and harder to bust what is and isn't real. With the example of Zelenskyy (Allyn, 2022), Deepfakes can and are being used to control the public's perception using realistic generated media. A Deepfake released with an increased urgency (say, a world leader declaring war), could create conflict in the future. The continuing development of Deepfakes creates an environment where media is a weapon itself.

Bibtex References

1.) @article {1282,
title = {The Emergence of Deepfake Technology: A Review},
journal = {Technology Innovation Management Review},
volume = {9},
year = {2019},
month = {11/2019},
pages = {40-53},
publisher = {Talent First Network},
chapter = {40},
address = {Ottawa},
keywords = {artificial intelligence, cybersecurity, deep learning, Deepfake, fake news},
issn = {1927-0321},
doi = {http://doi.org/10.22215/timreview/1282},
url = {timreview.ca/article/1282},
author = {Westerlund, Mika}
}

2.) @misc{winick_2020,
title={How acting as Carrie Fisher's puppet made a career for Rogue One's princess leia},
url={https://www.technologyreview.com/2018/10/16/139739/how-acting-as-carrie-fishers-puppet-made-a-career-for-rogue-ones-princess-leia/},
journal={MIT Technology Review},
publisher={MIT Technology Review},
author={Winick, Erin},
year={2020},
month={Apr}}

3.) @misc{allyn_2022,
title={Deepfake video of Zelenskyy could be 'tip of the iceberg' in Info War, experts warn},
url={https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia},
journal={NPR},
publisher={NPR},
author={Allyn, Bobby},
year={2022},
month={Mar}}

4.) @misc{deng_yu_2014,
title={Deep Learning: Methods and Applications},
url={https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DeepLearning-NowPublishing-Vol7-SIG-039.pdf},
journal={Foundations and Trends in Signal Processing},
author={Deng, L; Yu, D},
year={2014},
month={Oct}}

5.) @misc{pouget_mirza_Xu_warde_ozair_courville_bengio_yu_2014,
title={Generative Adversarial Nets},
url={https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf},
journal={Departement d’informatique et de recherche op ´ erationnelle University of Montreal },
author={Pouget-Abadie, Jean; Goodfellow, Ian; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua},
year={2015}