**Deepfake Technology and Its Ramifications**

Avery Peiffer

Department of English and Composition, University of Pittsburgh

ENGCMP 0400: Written Professional Communication

Professor Karen Reardon

September 14, 2020

Abstract

This paper explores the technology behind deepfakes, computer generated photos and videos that are, ideally, indistinguishable from their real counterparts. Deepfake technology is built using a relatively simple machine learning model called an autoencoder, which will be outlined briefly. In digital media, deepfakes are desired for their ability to generate new faces or modify existing ones. As such, the autoencoder technology has beneficial uses in camera apps and filmmaking, where it can be necessary to change various aspects of an individual’s face. The face swap features on Snapchat and Instagram are powered by an autoencoder model, for example (Brown, 2019). However, deepfakes can incorporate existing faces into photos or videos without the subject’s consent, leading to justified fear about the technology’s frightening applications. Such applications, both positive and negative, will be detailed. This paper also explores industry efforts to detect deepfakes, as well as the unique phenomenon in which improvements in deepfake detection are accompanied by equal improvements in deepfake creation. This is a topic that interests me specifically, as I intend to pursue data science as a career path.

*Keywords*: autoencoder, deepfake, neural network, unsupervised learning

**Deepfake Technology and Its Ramifications**

On April 18, 2020, State Farm aired a commercial to preview ESPN’s upcoming documentary series *The Last Dance* (State Farm Insurance, 2020). The commercial uses footage from 1998, with the then-38-year-old ESPN anchor Kenny Mayne reporting on the Chicago Bulls’ recent NBA championship. After the report, Mayne unexpectedly begins speaking using present-day slang and even mentions the documentary series by name (Hsu, 2020). The footage is obviously doctored, but it is believable enough for the concept to land. It is no small feat to layer the 60-year-old Mayne’s mouth over video of his younger self.

This commercial was created using deepfake technology, a relatively youthful branch of artificial intelligence that seeks to generate images and videos that are extremely similar to the real thing. Deepfakes are most associated with replicating human faces due to their use in settings such as the State Farm commercial. This paper examines deepfake technology and the positive and negative ramifications it holds in a media landscape that tends to prioritize user engagement over the objective truth.

**Autoencoders**

**Neural Networks Overview**

The use of deepfakes requires the utilization of a neural network, which is a “computational system loosely inspired by the way in which the brain processes information” (Zucconi, 2019). Essentially, the goal of a neural network is to reproduce a task, which can be easily done by humans, by recognizing patterns within a set of data. This process, called training, involves supplying the network with sample inputs and desired outputs. As the network recognizes the patterns necessary to get from the input to the output, it “learns” how to perform the task.

**Autoencoders for Facial Recognition**

In traditional neural networks, there is a difference between the sample inputs and the desired outputs. For example, a network could learn how to recognize numbers and letters based on an individual’s handwriting, as has been documented by Oh and Suen (2002, p. 233). Autoencoders are a type of network that deviate from this structure, in that they are trained to compress an input, and then recreate it as accurately as possible (Ng et al., 2013). Because they must recreate the input from a compressed image, the network will recognize patterns related to the specific aspects that differ between images. When applied to pictures of faces, autoencoders will learn about the unique features that comprise faces, and how important each of those features are. In doing so, autoencoders can learn to generate new faces or modify existing ones. This can be seen on the website thispersondoesnotexist.com, based on research by Karras et al, which will generate a realistic image of a face that its autoencoder has never seen before (2019).

For the purposes of deepfakes, it is more relevant that an autoencoder can modify existing faces. This process involves taking in data about a specific person’s face and detecting the patterns that make up the features of that face. After it has learned about the face, a well-trained autoencoder can then “predict what that person’s face would look like when mimicking the expressions of someone else” (Lim, 2020). The massive amount of sample photographs needed to learn a person’s face means that, for now, deepfake technology is mostly applied to celebrities, of whom many photographs and videos already exist.

**Deepfake Applications**

The premise of deepfake technology is frightening to many, and rightfully so. Currently, the most common malicious application of deepfakes is in using celebrity faces for pornography; though this practice is horribly unjust to the celebrities who are not consenting to the use of their images, it is unlikely to have lasting impacts beyond that person. However, the use of deepfakes to spread misinformation could be the most concerning application of the technology. In a digital landscape where objective truth is less important than keeping an audience’s attention, it is not difficult to envision deepfakes being abused to feed into a narrative. For example, a harmful conspiracy theory could easily gain traction among a more mainstream audience if a government official agreed with it in a deepfaked video.

It must be acknowledged that deepfakes do have the potential to be useful in many different circumstances. For example, as the COVID-19 pandemic has made it difficult and more expensive to shoot video, companies have turned to deepfakes to produce their corporate training videos (Simonite, 2020). Additionally, production companies such as Disney have begun using deepfakes to portray characters when an older or younger version is needed, or when the actor is unavailable or deceased (Lim, 2020). These mundane applications, which seek to increase convenience and cost-effectiveness, will likely be the most common uses for deepfake technology once it has settled into an acceptable role in digital media.

**Combating Deepfakes**

**Industry Efforts**

Due to their potential to spread misinformation, many technology companies are taking steps to limit the effectiveness of deepfakes online. In January of 2020, Facebook explicitly banned deepfakes from its platform, citing their ability to deceive the average user (Room, 2020). However, this action was met from criticism from many who believed that Facebook was not doing enough to combat the spread of disinformation, prompting further discussion on Facebook’s role in moderating content. In December of 2019, several technology companies, including Facebook, announced a Deepfake Detection Challenge in which participants submitted deepfake detection algorithms for financial reward. A dataset of over 100,000 video clips was sourced and reviewed by Dolhansky et al. (2020). The winning algorithm was able to detect deepfakes at an average rate of 65.18%, highlighting the difficulty for a computer to reliably detect them (Lim, 2020).

Researchers have brought up issues with improving algorithms for deepfake detection. Currently, those who create deepfake algorithms test them by seeing if a detection algorithm will flag their content as suspicious. A deepfake that passes this test is considered satisfactory, as it cannot be labeled as fake using only current algorithms. When detection algorithms improve, deepfake creation algorithms essentially receive a higher quality network to test against (Vincent, 2019). The deepfakes that result from this testing are therefore ever-more difficult to distinguish from the real thing. A recent detection algorithm from Sabir et al. boasted a 97% accuracy rate in detecting deepfakes from a limited sample (2019). Unfortunately, this algorithm was not heralded as a savior for preventing online misinformation, because it can now be used to test deepfake generating algorithms. It is for this reason that Facebook is not making its own detection algorithms public (Lim, 2020).

**Recognizing Deepfakes**

Fortunately, internet users can learn how to spot deepfakes without the use of an algorithm. The MIT Media Lab recently published an article detailing several inconsistencies that are often present in deepfaked faces (Groh, 2020). If someone comes across a strange video that they believe could be doctored, they should examine whether various criteria of the subject are consistent with a normal human face. Deepfaked subjects often have incongruities in the agedness and texture of their skin; their forehead can appear more wrinkled than their cheeks, for example. The video can misrepresent the natural physics of lighting, meaning that shadows and glare appear in unexpected places. Various features, such as facial hair, moles, and lips, often fail to come across as fully natural in deepfaked videos. Finally, deepfaked subjects may blink at an irregular rate compared to normal people (Groh, 2020). While these inconsistencies make it easier to mitigate the misinformation spread by deepfakes, it remains difficult to eliminate them entirely.

**Future Career Plans**

I am not specifically interested in working to detecting deepfakes, as I do not feel well-suited for the difficulties of combating misinformation online. In any regard, they act as a useful case study on my intended career field of data science. Though it is a broad term, data science involves the use of algorithms and statistics to study and draw insights from data. Data science is unique in that it requires almost no startup cost; it is possible to be a practicing data scientist with little to no formal education or experience. While its low barrier to entry encourages any eager mind to contribute to data science, it also means that anyone can exploit data science tools for malicious intent, such as spreading misinformation with deepfakes.

Previously, I thought that data science tools were only used to make predictions about existing information, as seen in artificial intelligence models that can make accurate weather forecasts (*How Data Science Can Enhance Weather Forecasting*, 2020). Researching deepfakes and unsupervised learning has taught me about the ability of data science to generate completely new information. My biggest takeaway from researching this subject is the responsibility of data scientists to consider the possible consequences of the tools they build. Nearly every aspect of modern life is driven by data science models, so it is important to understand that any exploitation of those models is likely to have far-reaching effects.

**Conclusion**

In conclusion, deepfakes present a serious challenge to the integrity of information found online. They are both effective and hard to eliminate, as shown by their resiliency against improvements in detection algorithms. For this reason, experts have turned to improved education as the most effective defense against deepfakes. If a video or photo seems bizarre, is of low quality, and has some strange visual effects, there is a high probability that it is a deepfake (Lim, 2020). Though they have niche uses, their potential to seriously harm online discourse has alarmed many in the artificial intelligence community, and considerable effort is being taken to ensure that they do not further corrupt the spread of information online.

**References**

Brown, D. (2019, December 9). *A more realistic Bitmoji? Snapchat is rolling out a tool called Cameo that uses deepfake technology*. USA Today. <https://www.usatoday.com/story/tech/2019/12/08/snapchat-cameo-deepfake-tool-social-media-app/4376790002/>.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. (2020, June 12). *The DeepFake Detection Challenge Dataset*. ArXiv.Org. <https://arxiv.org/abs/2006.07397>

Groh, M. (2020, April). *Project Overview ‹ Detect DeepFakes: How to counteract misinformation created by AI –*. MIT Media Lab. <https://www.media.mit.edu/projects/detect-fakes/overview/>

*How data science can enhance weather forecasting*. (2020, July 6). The European Business Review. <https://www.europeanbusinessreview.com/how-data-science-can-enhance-weather-forecasting/>

Hsu, T. (2020, April 22). *An ESPN Commercial Hints at Advertising's Deepfake Future*. <https://www.nytimes.com/2020/04/22/business/media/espn-kenny-mayne-state-farm-commercial.html>.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2019, December 3). *Analyzing and Improving the Image Quality of StyleGAN.* ArXiv.Org. <https://arxiv.org/abs/1912.04958>

Lim, H. (2020, July 14). *A Look at Deepfakes in 2020*. Lionbridge AI. <https://lionbridge.ai/articles/a-look-at-deepfakes-in-2020/>.

Ng, A., Tandon, S., Wang, T., Huval, B., Hannun, A., Maas, A., … Ngiam, J. (2013, August 28). *Autoencoders*. Unsupervised Feature Learning and Deep Learning. <http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/>.

Oh, I.-S., & Suen, C. Y. (2002). A class-modular feedforward neural network for handwriting recognition. *Pattern Recognition*, *35*(1), 229–244. <https://doi.org/10.1016/s0031-3203(00)00181-3>

Romm, Tony. (2020, January 8). *Facebook bans deepfakes, but new policy may not cover controversial Pelosi video*. The Washington Post. <https://www.washingtonpost.com/technology/2020/01/06/facebook-ban-deepfakes-sources-say-new-policy-may-not-cover-controversial-pelosi-video/>.

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P. (2019, May 16). *Recurrent Convolutional Strategies for Face Manipulation Detection in Videos*. ArXiv.Org. <https://arxiv.org/pdf/1905.00582.pdf>

Simonite, T. (2020). *Deepfakes Are Becoming the Hot New Corporate Training Tool*. Wired. <https://www.wired.com/story/covid-drives-real-businesses-deepfake-technology/>.

State Farm Insurance. (2020, April 20). *Predictions | State Farm + ESPN Commercial* *(Featuring Kenny Mayne)* [Video]. YouTube. <https://www.youtube.com/watch?v=FzOVqClci_s>

Vincent, J. (2019, June 27). *Deepfake detection algorithms will never be enough*. The Verge. <https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work>

Zucconi, A. (2019, August 16). *An Introduction to Neural Networks and Autoencoders*. Alan Zucconi. <https://www.alanzucconi.com/2018/03/14/an-introduction-to-autoencoders/>.