The Technical Implications of Deepfakes

(Technical Paper)

The Societal Impact of Deepfakes

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Ian Switzer

Fall, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Richard Jacques
Department of Engineering & Society

Introduction

The twenty-first century has brought countless technological developments, but one of the most influential has been the advancement of artificial intelligence (AI), which has made its way into everything from cars to mobile phones. Machine learning has become a viable strategy for generating media such as text, audio, images, and videos. AI-generated text, though impressive, hardly raises any ethical concerns. On the other hand, AI-generated audio clips, images, or videos, also known as deepfakes, can potentially have huge impacts on society as a whole. Research in the late 90s and early 2000s paved the way for the advanced deepfakes that exist today. Sophisticated AI can produce fake audio clips of someone talking that sound indistinguishable from a real person to the naked ear. That audio can be fed into another algorithm that generates a talking face that matches the words. This video could be a deepfake of anyone, including celebrities and politicians, which only worsens the already rampant issue of misinformation on the internet. Though deepfakes have some potentially beneficial uses, most deepfake technology has been used for nefarious purposes, such as to create explicit videos of celebrities or blackmail politicians. Advanced deepfakes have only been around for a few years, but the technology has already progressed to the point where anyone can create them with relative ease. This technical discussion will seek to analyze the different methods for creating deepfakes and determine how deepfakes can be recognized at a technical level. This STS discussion will attempt to determine the potential societal impacts of deepfakes and how online platforms can respond to sophisticated misinformation effectively.

Technical Discussion

Fake images and videos can be created manually using editing tools like Photoshop, but the term ‘deepfake’ refers to images and videos specifically created by machine learning

software. Deepfakes are created using deep learning, which is an advanced form of artificial intelligence. One common method for making deepfakes is using Generative Adversarial Networks, which involves two neural networks that work against one another. The first neural network attempts to generate an image or video based on a sample dataset, and the second neural network guesses whether the generated image or video is fake (Goodfellow et al., 2014). This process repeats until the second neural network can no longer determine whether the generated media is fake. At that point, the program decides that the output is good enough and presents it to the user. The sample dataset can contain photos or videos of celebrities, politicians, or anyone else, and the result will be a deepfake of that person. Another method for creating deepfakes uses autoencoders, which uses encoding and decoding algorithms to splice the face from one person onto another (Nguyen et al., 2019). This can be used to generate a video of one person speaking based on a video of someone else saying the same thing. Researchers at the University of Washington used neural networks to create a deepfake video of Obama based on an audio clip (Suwajanakorn et al., 2017). Their technique involved generating mouth shapes that fit the audio clip and then texturing it to fit the target face. These are just a few examples of methods for creating deepfakes. Machine learning is always evolving and offers a myriad of techniques for generating entire fake images and videos, splicing one face onto another, creating lip and mouth shapes from audio clips, etc. Advances in machine learning have allowed these algorithms to be run on any computer, regardless of hardware. Even an older laptop could run most of these algorithms, though it might take some time.

As deepfakes have grown more and more common, researchers have also focused on identification techniques to determine whether images and videos are real or fake. Early deepfakes that used face-swapping were easily detectable because the face either never blinked

or did not blink at a realistic frequency (Li et al., 2018). However, soon after this research was published, deepfake synthesis algorithms were improved to include blinking. Another technique, developed in 2019, uses machine learning to evaluate a person's consistent facial expressions based on known authentic videos and compares it to the facial structures in a questionable video to see if these facial structures change unexpectedly (Agarwal et al., 2019). This approach has several limitations, the most significant being the reliance on existing videos for training. This model would be impractical for anyone besides famous figures for whom there is ample video footage. The model also falls short when the subject is in varied contexts, such as looking at the camera versus looking at someone else while talking, though this limitation is largely based on the variation present in the dataset. A third method for detecting deepfakes involves training a neural network to analyze inconsistencies between frames in videos because the GAN method for creating deepfake videos is unaware of the last frame it created when generating the next one (Guera & Delp, 2018). This technique is much more robust because it relies only on the video in question itself for detection, rather than a large pre-existing dataset for a particular person. Deepfake detection is progressing rapidly, but it is nearly impossible to predict the advancements that the next generation of deepfake technology will bring. Governments and other influential entities are well-aware of the existence of deepfakes and recognize the importance for detection and prevention.

STS Discussion

There are countless applications for deepfakes that are indistinguishable from real images and videos, some good and some bad. Machine learning has been used for special effects in movies to simplify the generation of computer graphics (Miller, 2022). Deepfake technology also has the potential to be used to assist hearing-impaired people by presenting a real-time synthetic

video for lip-reading based on phone call audio (Suwajanakorn et al., 2017). However, the downsides of deepfakes are much more important to consider. Deepfakes can be used to imitate important political figures in order to spread misinformation or hate speech. On the other hand, important figures could attempt to discredit an incriminating video by calling it a deepfake. In Malaysia, someone leaked an explicit video of a senior Cabinet minister, which the Prime Minister said he believed was fake but was later authenticated by the Malaysian Cybersecurity team (McCoy, 2019).

The creation of explicit photos and videos using this technology has been happening for years, but with the rapid advancements in machine learning, explicit deepfakes have started popping up more and more. These pictures or videos could be created to harass an individual or organization, for personal enjoyment, or for political gain. Back in 2016, during the presidential election, it would have been unheard of for a candidate to claim that a recording of them was a fake. Nowadays, as Galston puts it, "a well-timed forgery could tip an election" (Galston, 2020). In order to combat deepfake technology, the Defense Advanced Research Projects Agency (DARPA) office of the U.S. Department of Defense created a Semantic Forensics (SemaFor) program to analyze the semantic features of deepfakes (Defense, 2019). This program focuses on semantic errors such as mismatched earrings that existing detection algorithms would not notice.

Governments and social media platforms need to be aware of potential deepfakes because fake news can spread faster than the truth. It can be difficult for people to recognize deepfakes when they are not aware that they exist. To raise awareness and encourage innovation, Facebook announced a deepfake detection competition in 2019 with $10 million in prizes and grants and released a dataset of faces from consenting adults (Cole, 2019). Facebook has faced a lot of scrutiny in the past for their lack of content management policies and their unauthorized usage of

user data. This competition is a step in the right direction for the company. Platforms like Facebook allow misinformation to run rampant, especially because most users are not technologically savvy. The ones that are tech savvy have the potential to wreak havoc on communities and governments.

Most new technologies begin in a phase where only experts can use them and then slowly progress towards a version that anyone can use. Some examples include: computers, which were initially massive machines used by companies and researchers; mobile phones, which started out as big and bulky and difficult to transport; and the internet itself, which began as ARPANET, an academic research network funded by what is now DARPA (Lee, 2014). Machine learning recently transitioned into the phase where anyone can sit down at a computer and start building a deep learning model. Now, anyone can create simple deepfakes with relative ease. Sophisticated deepfakes require more advanced models but there are countless websites and apps that offer powerful deepfake functionality for cheap or free (Khan, 2022). Someone could go onto their cell phone or computer, upload a photo or video, and have a brand new deepfake in minutes. People tend to trust pictures and videos that look completely realistic to the naked eye, which means that social media platforms need to be hyper-vigilant going forward because deepfakes will only become more prevalent and more advanced.

Conclusion

The growing threat of deepfake technology has spurred some research into detection and prevention, but there is still a lot of work to be done to mitigate future problems. Methods for detecting fake pictures and videos have progressed quickly but have been constantly outpaced by developments in deepfaking techniques. Social media giants are aware of deepfakes and have taken steps to prevent relevant misinformation but the fight against deepfakes is an uphill battle.

Media forensics teams are forced to be reactive rather than proactive, reacting to new deepfake techniques as they appear. It is impossible to prevent deepfakes from being created or new techniques from being developed. The best course of action for governments and corporations alike is to remain vigilant and encourage people to question what they see on the internet. Social media giants are in the unique position where they can influence what people see on a global scale, so they are the ones that people will blame if false images or videos make their way across the internet. Unfortunately, it is becoming increasingly easy for the average person to create deepfakes, whether by using an app or website or constructing their own machine learning model. In the coming years, deepfakes will slowly break down public trust of videos on the internet, whether for good or bad.

References

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting world leaders against deep fakes. *IEEE International Conference on Computer Vision and Pattern Recognition 2019 Workshop on Media Forensics*. Retrieved October 28, 2022, from http://www.hao-li.com/publications/papers/cvpr2019workshopsPWLADF.pdf.

Cole, S. (2019, September 5). *Facebook just announced $10 million 'Deepfakes Detection Challenge'*. VICE. Retrieved October 28, 2022, from https://www.vice.com/en/article/8xwqp3/facebook-deepfake-detection-challenge-dataset.

Defense Advanced Research Projects Agency, & Corvey, W., DARPA (2019). Retrieved October 28, 2022, from https://www.darpa.mil/program/semantic-forensics.

Galston, W. A. (2022, March 9). *Is seeing still believing? the deepfake challenge to truth in Politics*. Brookings. Retrieved October 29, 2022, from https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. *Arxiv*. https://doi.org/1406.2661.

Guera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. https://doi.org/10.1109/avss.2018.8639163.

Khan, I. H. (2022, February 13). *Top deepfake apps and websites you can try*. LinkedIn. Retrieved October 30, 2022, from https://www.linkedin.com/pulse/top-deepfake-apps-websites-you-can-try-imran-hussain-khan/.

Lee, T. B. (2014, June 16). *The internet, explained*. Vox. Retrieved October 29, 2022, from https://www.vox.com/2014/6/16/18076282/the-internet.

Li, Y., Chang, M.-C., & Lyu, S. (2018). In ICTU oculi: Exposing AI created fake videos by detecting eye blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. https://doi.org/10.1109/wifs.2018.8630787.

McCoy, P. (2019, June 13). *I believe sex video is fake, says Mahathir*. The Straits Times. Retrieved October 27, 2022, from https://www.straitstimes.com/asia/se-asia/i-believe-sex-video-is-fake-says-mahathir.

Miller, T. (2022, March 21). *How deepfake technology is changing the movie industry*. Seat42F. Retrieved September 30, 2022, from https://seat42f.com/how-deepfake-technology-is-changing-the-movie-industry.

Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep Learning for Deepfakes Creation and Detection. https://doi.org/1909.11573.

Suwajanakorn, S., Seitz, S. M., Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning Lip Sync from Audio. https://dx.doi.org/10.1145/3072959.3073640.