Mitigating Demeaning Manipulations of Deepfakes

Madison Doan

Dr Steven Maulden

ENGL 3003

November 16, 2023

Table of Contents

1. Introduction
2. History/Origin of Deepfakes
3. Machine Learning Science
4. Consequences of Malicious Deepfake Use
5. Mitigating and Detecting Synthetic Media
6. Moving Forward
7. Conclusion

Introduction

Deepfakes—or synthetic media types have evolved from the internet to introduce an unfamiliar threat to the public through malicious actors often seeking to exploit the humanhood of victims. Modernity has given us facial, voice, and other manipulations that allow ill-disposed operators to abuse these advancements and cause further deepening of the ugly twin that comes from the beauty of the world wide web. Originating from deep learning, the creation of these artificial media requires various network schematics and technical algorithms that may make it difficult to challenge the essence thereof. There are developing techniques that are increasingly accurately sorting out between artificial instances but there is an upper hand from social media exploitation due to the ability to change lives or numerous stances within seconds. While science is growingly astounding, we must prevail in cyber security practice by encouraging technological development in addition to deterring and mitigating artificial intelligence abuse, to classify real from fraud, and protect the authenticity of humankind's interactions.

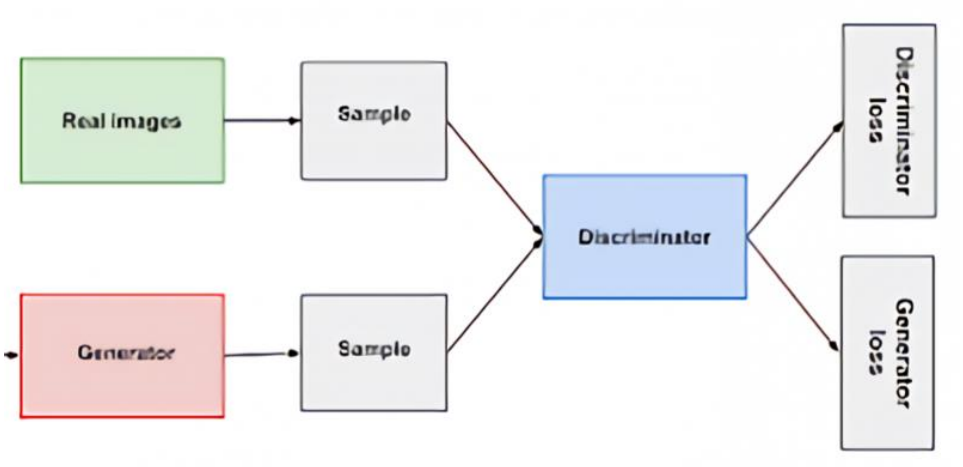
To understand the consequences and implications of human manipulation, it is essential to have a keen knowledge of synthetic media applications. There will be a brief history of facial manipulation and what purpose it served to benefit the interests of scientists. Additionally, scientific explanation will serve a great purpose in completely understanding the threats this brings to society. There are several advantages one may utilize deepfake technology for and such consequences can be proven detrimental to promote mitigation methods. Lastly, we must move forward and discuss risk assessments to all deepfake possibilities and regain security in such a central place in many people's lives by outsourcing offensive intelligence and challenging unethical disposition.

History

The use of synthetic media developed from digital media forensics and aided corporations to ability to portray nonexistent situations such as a passed celebrity's presence or marketing techniques to attract targeted crowds. Once this technology was released, it fell into the hands of several malicious actors that increased deepfake's popularity. The origin of the term "deepfake" appeared in 2017 from a user that took advantage of open-sourced code of face swapping technologies to insert a false likelihood of a known facial image into pornography. Since then, synthetic humanhood has ranged from entire artificial documentaries to marketing schemes and have been weaponized several times socially and politically such as the first appearance of deepfakes, being a dubbed video of former President, Barack Obama. The fun of these technologies was soon overlooked when the potential for detrimental harm of impersonating the US President became apparent. While the artificial intelligence of deepfakes may be beneficial to specific people it does not outweigh the historical implications and remains a sizeable threat to the internet.

Science

Machine learning for artificial facial manipulations utilizes several networking techniques to generate accuracy among generations of facial productions. One of the most common forms of such algorithms is a Generative Adversarial Network (GAN) pictured below. Simply put, a GAN is a pairing of two Artificial Neural Networks (ANN) where one focuses on generating fake samples of one data domain, and the other focuses on detecting whether the generated sample is real or fake. More specifically, a GAN demonstrates a system in which samples are generated and encoded to be standardized by a discriminator decoder with each failure discarded to create optimal definitiveness in machine learning.



When two GANs are used in series it becomes possible to map one image, real or fake, to another, allowing for a morph between two images to occur. A particular pairing of these GANs is a cyclical pairing where the domain of one image is changed to adapt to the domain of another. Moreover, the CycleGAN allows periodical propagation of generations that contribute to such machine learning amongst several other neural networks, which are the backbone of facial swapping mechanisms that precisely carry features between the two parties. The cyclic connection ensures facial swaps without a loss of features with this method. These all contribute to the novelty of deepfakes that make them so convincing often with an accuracy in the high nineties. Further advancements of this technology such as Convolution Neural Networks, which receive input from a matrix that capacitates layers of given images to detect humanhood likeliness, elevate the current condition and accuracy of deepfakes.

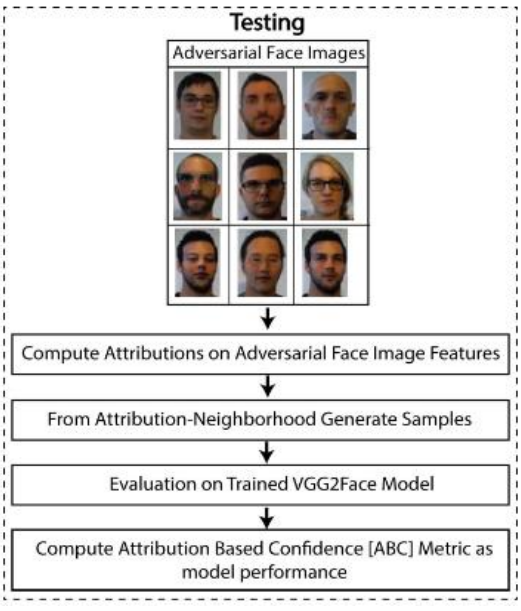
Threats

There are several current and hypothetical threats that keep users at risk. While deepfakes themselves are a threat to humanity, artificial presence of an individual can forfeit one's identity and ability to authenticate themselves. The digital age has brought a cry for authenticity between internet users and tangible harm may arise from these threats. Malicious utilization of this

technology has surfaced deep videos to spread misinformation such as Nancy Pelosi's doctored video that caused an uproar in 2019. More significant past instances include easy access to deepfake pornography techniques which is a severe violation to one's integrity. A large sum of these threats may also be categorized as social engineering, which if successful, can open a plethora of dangers to internet users, such as identity theft solely through a profile photo. These threats have the potential to strike up violence in populations, damage individuals, affect nation states, and question the overall depreciation or acceleration of democracy that the internet facilitates.

Since synthetic media requires the use of machine learning, it introduces a whole new entanglement of threats that arise when considering processing and stored data. While cutting edge technology is moving forward from real-human examples to its own artificially generated faces for deep learning, thousands of users are at risk solely from the original storage of private and facial information. Residual risk for artificial intelligence, specifically machine learning challenges the integrity and authenticity of human intellectual property in addition to its own threats.

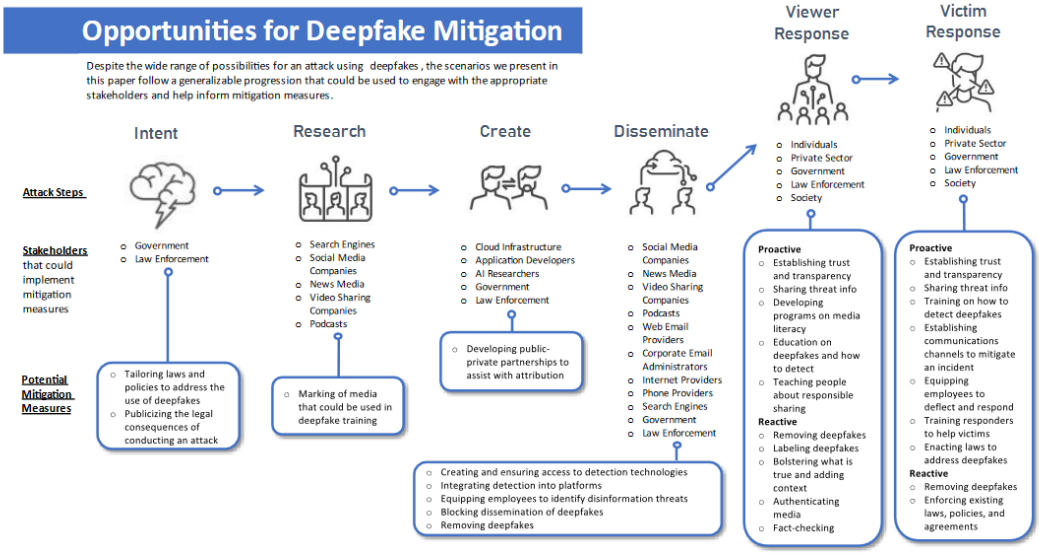
Mitigation



Mitigating synthetic media abuse comes from several areas and aims to protect, detect, and audit. At the scientific level researchers are declaring certainty in deepfakes using neural networks to analyze and score videos. Using algorithms such as the Attribute-Based

Confidence metric has allowed computing in a multi-step process based off high-attribution features to assure model-based performance. Volunteered faces create the standardized pool for this detection process to provide references to the evaluated subjects. The ABC system retains 96% accuracy while still using controlled samples to provide comparative generations and neural facial detections as a solution to the deepfake issue. This proposed solution attempts to mitigate the root of deep learning.

Mitigation at a surface level requires accurate risk assessment to determine variable paths and worse-case scenarios. Corporations are encouraged to numerate relative assets to theorize what risks are inherent to them. Furthermore, stakeholders receive mitigation by civil control such as laws and policies, intensive cyber training, and authentication mechanisms such as Two-Factor Authentication and digital signatures/certification. This organizational threat modeling will soften the likelihood of malicious actors attacking and deters implementation of artificial media. This model focuses on subject variables as well as measures to be taken to ensure infrastructural and governance security on attack surfaces to combat synthetic media misuse by a defensive architecture style.



Moving Forward

Technological advancements rapidly alter the cyber world, so it is imperative to stay onto developing deepfake technology to complete future mitigation. As each tool is being updated, machine learning requires trial and error and full commitment to secure the past, present and future of synthetic media types. To accompany the current detection and mitigation techniques cyber forces must additionally acquire control to not only deep learning, but the internet itself. Deep learning and neural networks are a significant improvement to technology that aid the usage of self-driving cars, mitigating racial bias, and other useful mechanisms that welcome artificial intelligence in trade for ease. It is important for all users to understand the nuances that arrive at the use of synthetic media and that mitigation can be acquired not only scientifically but socially as well with perseverance. The future of cyber security relies on tenacity between the good and evil in intelligence and overall transparency to the public to carry out ethical use of something so innovative such as machine learning.

Conclusion

Manipulations of human intelligence counters modern science, but with proper evaluation, we can continue to use deep learning technology to further greaten science while looking out for humankind. Cyber forces have a civic duty to society that is proven to be a pivotal matter to an individual in any case. There will remain the threat of deepfakes that malicious users have the advantage of, but the future of artificial intelligence relies on the balances of reality and the cyber world that need to be reinforced additionally to the general combat of malicious users. With proper research, threat assessment, and detection techniques we

can preserve the internet and outsource malicious users in their own game. The future of machine learning provides its' own world as its oyster, and with diligence we can keep growing while introducing society to the beauty of cutting-edge technology.

References

Abhari, Julian, and Ashwin Ashok. *Mitigating Racial Biases for Machine Learning Based Skin Cancer Detection*. 16 Oct. 2023, https://doi.org/10.1145/3565287.3617639. Accessed 22 Nov. 2023.

Dan, Yabo, et al. "Generative Adversarial Networks (GAN) Based Efficient Sampling of Chemical Composition Space for Inverse Design of Inorganic Materials." *Npj Computational Materials*, vol. 6, no. 1, 26 June 2020, pp. 1–7, www.nature.com/articles/s41524-020-00352-0, https://doi.org/10.1038/s41524-020-00352-0.

DEEP Increasing Threat of AKE Identities.

Fernandes, Steven, et al. *Detecting Deepfake Videos Using Attribution-Based Confidence Metric*. 2020.

Mirsky, Yisroel, and Wenke Lee. "The Creation and Detection of Deepfakes." *ACM Computing Surveys*, vol. 54, no. 1, 2 Jan. 2021, pp. 1–41, https://doi.org/10.1145/3425780.

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

Tolosana, Ruben, et al. "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection." *Information Fusion*, vol. 64, Dec. 2020, pp. 131–148, arxiv.org/pdf/2001.00179.pdf, https://doi.org/10.1016/j.inffus.2020.06.014. Accessed 10 Aug. 2020.