DETECTING DEEPFAKES: LAYERED ARCHITECTURES AND GENERATED ARTIFACTS

DEEPFAKES' DEEP SCARS ON DEMOCRACY

An Undergraduate Thesis Portfolio
Presented to the Faculty of the
School of Engineering and Applied Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Nathan Williams

May 12, 2023

SOCIOTECHNICAL SYNTHESIS

While originally meant for visual effects in the film industry, recently deepfake technology has been abused to sow confusion and spread misinformation throughout society. With current deepfake detection classifiers averaging around 70% accuracy on black box trials, there are no surefire methods for identifying deepfaked content. The state-of-the-art technical research report seeks to improve the accuracy of deepfake detection classifiers by combining two leading techniques in deepfake detection: stacking neural networks and generating temporal training artifacts. The STS research paper explores how different countries attempt to mitigate the dangers of deepfaked misinformation on social media platforms. This comparative case study research paper uses Actor Network Theory to analyze the strengths and weaknesses of unique actor networks and proposes a new system for regulating synthetic media. By exploring both societal and technological avenues for combatting malicious deepfakes, this thesis seeks to expand the field of artificial intelligence and contribute to protecting society from misinformation.

If technology existed that could detect deepfaked media without error, then any attempt to spread misinformation with synthetic media could be identified and removed with ease. Therefore, it is imperative to explore new methods to improve the accuracy of deepfake detection classifiers. The technical portion of this capstone reviewed studies of state-of-the-art methods used in deepfake detection classifiers and any accompanying open-source code. The eventual goal of this research being to propose a new method for deepfake classification or propose a way to improve on existing methods.

The result of this research is the proposal of a deepfake detection model that combines neural network stacking and temporal artifact generation to improve accuracy. By using

specialized recurrent neural networks to create temporal training artifacts, classifiers can analyze videos based on a sequence of frames rather than individual frame data providing a more in-depth evaluation of the input data. Furthermore, stacking multiple sub-models in a random weighted ensemble can improve accuracy by giving the model more forms of data analysis to choose from. While it is predicted that the proposed model would improve detection accuracy, its implementation would prove difficult due to lack of sufficient training data and its requirements for time and computational resources.

Since deepfake detection technology is not accurate enough to be deployed commercially, my STS research report seeks to answer how societal methods can be used to mitigate the dangers of deepfakes. The STS portion of this capstone suggests that to regulate synthetic media effectively and ethically, a system where social media corporations are held legally liable for their platforms' content should be implemented. To support this proposal, this report analyzes the misinformation management actor networks for the United States, China, and Australia considering their unique strengths and weaknesses.

When looking at the United States' approach to regulating deepfaked misinformation, research showed that reliance on outdated legislation that holds no actor accountable does nothing to protect media consumers from malicious actors. In contrast, while China's actor network has effective and rapid enforcement against malicious actors the removal of 3rd party fact checkers supports unethical government control of social media. Finally, Australia's system shows the most promise in combatting deepfaked misinformation with social media corporations opting to adhere to a code of practice that emphasizes removal of misinformation and periodic transparency reports.

The consequences of malicious deepfakes can be devastating not only to individuals but to entire societies. Left unchecked misinformation and confusion could run rampant as a result. The use of Actor Network theory can be immensely helpful in systematically mitigating the wounds of synthetic media, and as the accuracy of deepfake detection technology improves Actor Network Theory can incorporate this technology into an even better system.

TABLE OF CONTENTS