Deepfake detection by human crowds, machines, and machine-informed crowds

Elise Møllerstedt Gunnestad
Department of Informatics
University of Oslo
Oslo, Norway
elisegun@uio.no

Ida Klundseter Eneroth
Department of Informatics
University of Oslo
Oslo, Norway
idakene@uio.no

***Abstract*—Deepfake videos are a growing concern in the digital world as they pose a significant threat to the authenticity of visual media. Deepfakes have become so realistic that it is challenging to detect them. This extended abstracts will summarize the Deepfake detection by human cross, machines and machine-informed crowds by Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. The researchers conducted three experiments which explored detection accuracy of humans, machine learning algorithms and the combination of the two. The results suggest that combining the predictions of algorithms with human judgements can lead to a more accurate deepfake detection. The findings of this study can help the development of more effective deepfake detection technologies, thus reducing the threat of misinformation to society.**

I. Introduction

With the rise of deepfake technology, there has been a growing concern about the impact of these videos on society. Deepfake videos are realistic, but artificial, and can be used to spread false information, propaganda and fabricating evidence. Detecting whether a video is fake or not is challenging, and therefore, it is essential to explore different methods of detection to prevent the misuse of deepfakes.

II. Motivation and research questions

The main research question of this study is to explore the effectiveness of different methods of detecting deepfake videos, including human judgment, machine learning algorithms, and a combination of human and machine learning. The motivation of the study is to explore how we can improve detection software, to reduce the negative impact unidentified false videos may have.

III. Method

The study consists of three experiments.

A. Study 1

In the first experiment, the researchers wanted to compare the ability of humans and machines to detect deepfakes. They used 56 pairs of videos and asked the participants to identify which was the real one. There were 882 participants who saw at least 10 pairs of videos. Half of the videos were correctly identified by 83% of participants. There were 3 videos who were incorrectly identified by over 50%. The results did not show that participants improved their detection accuracy within their first 10 videos, but did show that for every additional 10 s spent on deciding, the accuracy decreased. 82% of the participants outperformed the model at 65% accuracy.

B. Study 2

In the second experiment, the researchers set up a website with a percentage slider. The participants could drag it from "100% confidence this is NOT a DeepFake" to "100% confidence this is a DeepFake", with "just as likely" in the middle. There were 9492 participants, where 9188 of these found the website organically. There were 50 videos, half were fake. 4 of these were of Kim Jung-un and Vladimir Putin. The participants accurately identified deepfakes in 57% of the attempts, compared to the leading machine learning model which identified 84% correctly. However, when it comes to identifying the real videos, the accuracy was the same at 75%. For the fake political videos, the participants outperform the leading model at a 60% accuracy, while the algorithm puts the probability of it being fake at 1-8%. For every video watched, the participants are more likely to report it as a deepfake. Overall, the average response of the participants per video, called the crowd mean, is about the same as the leading model at 80%.

C. Study 3

In the third experiment, the researchers wanted to test a new method for detecting deepfakes that combined human and machine learning. In this experiment they included the prediction from the leading model of revealing deepfake videos. Here, the participants were able to change their initial answer after seeing how the model predicted it. Results were given as confidence intervals, as in experiment two. The study showed that the participants updated their confidence in 24% of trials. The accuracy also increased from 66% to 73%. For 40 of the videos, where the model predicted correctly, the accuracy of prediction increased with 10.4%. For the remaining 10 videos, where the model had incorrect predictions, the results were 2.7% less accurate than before they saw the results from the AI model. In one extreme example, a model detected a deepfake video to be fake with 28% accuracy. This affected the participants, and the responses were on average 18% less accurate at identifying deepfakes. The deepfake videos

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

of Kim Jung-un and Vladimir Putin, were predicted by the model as 2% to 8% certainty that the videos were deepfakes. This prediction affected the participants from a 56% detection accuracy to 34% on the deepfake of Kim Jung-un, and 70% to 55% on the deepfake of Vladimir Putin.

IV. Conclution

In each experiment, the authors evaluated the accuracy of different detection methods, including human judgment, the leading machine learning algorithm, and a combination of human and machine learning. The results showed that combining the predictions of machine learning algorithms with human judgments can lead to even more accurate deepfake detection as they have different strengths. However, if the machine learning model is inaccurate, it decreases the accuracy of the humans. In videos which were especially blurry, grainy and dark, 8 out of 14 were correctly identified by the recruited participants, and 10 were correctly characterized by the model. This, along with other statistics on different disturbances in the video the results show that the model is better at detecting deepfakes than humans, with two exceptions. When there were either two people in the video, or a floating distraction the model was clearly worse at recognizing deepfakes. Respectively 21.9% and 11.3% decrease whereas the humans only had 7.6% and 3.5% decrease. Additionally, prior knowledge significantly improved the detection accuracy for both human and machine learning methods. Humans relied on faces to a large degree to determine whether the video was fake, and if the face was covered the detection rate dramatically decreased. However, the model used mostly disturbances in the background to deduce the realness of the video. Since humans and the model spot different queues the accuracy usually increased when combining their efforts. In conclusion, the study provides valuable insights into the detection of deepfake videos and highlights the importance of combining human and machine learning methods for effective deepfake detection, but only if the model has a high accuracy. The findings of this study can inform the development of more effective deepfake detection technologies and help reduce the threat of deepfakes to the integrity of visual media.

References

[1] Groh, Matthew et al., Deepfake detection by human crowds, machines, and machine-informed crowds, *Proceedings of the National Academy of Sciences*, **119**, e2110013119, 2022.