Deepfake detection improves when using algorithms that are more aware of demographic diversity

April 16 2024, by Siwei Lyu, Yan Ju

Credit: Markus Winkler from Pexels

Deepfakes—essentially putting words in someone else's mouth in a very believable way—are becoming more sophisticated by the day and increasingly hard to spot. Recent examples of deepfakes include Taylor Swift nude images, an audio recording of President Joe Biden telling New Hampshire residents not to vote, and a video of Ukrainian President Volodymyr Zelenskyy calling on his troops to lay down their arms.

Although companies have created detectors to help spot deepfakes, studies have found that biases in the data used to train these tools can lead to certain demographic groups being unfairly targeted.

My team and I discovered new methods that improve both the fairness and the accuracy of the algorithms used to detect deepfakes.

To do so, we used a large dataset of facial forgeries that lets researchers like us train our deep-learning approaches. We built our work around the state-of-the-art Xception detection algorithm, which is a widely used foundation for deepfake detection systems and can detect deepfakes with an accuracy of 91.5%.

We created two separate deepfake detection methods intended to encourage fairness.

One was focused on making the algorithm more aware of demographic diversity by labeling datasets by gender and race to minimize errors among underrepresented groups.

The other aimed to improve fairness without relying on demographic labels by focusing instead on features not visible to the human eye.

It turns out the first method worked best. It increased accuracy rates from the 91.5% baseline to 94.17%, which was a bigger increase than our second method as well as several others we tested. Moreover, it increased accuracy while enhancing fairness, which was our main focus.

We believe fairness and accuracy are crucial if the public is to accept artificial intelligence technology. When large language models like ChatGPT "hallucinate," they can perpetuate erroneous information. This affects public trust and safety.

Likewise, deepfake images and videos can undermine the adoption of AI if they cannot be quickly and accurately detected. Improving the fairness of these detection algorithms so that certain demographic groups aren't disproportionately harmed by them is a key aspect to this.

Our research addresses deepfake detection algorithms' fairness, rather than just attempting to balance the data. It offers a new approach to algorithm design that considers demographic fairness as a core aspect.

Provided by The Conversation