Difficulties in Deepfake Recognition

Daniel Matisoff

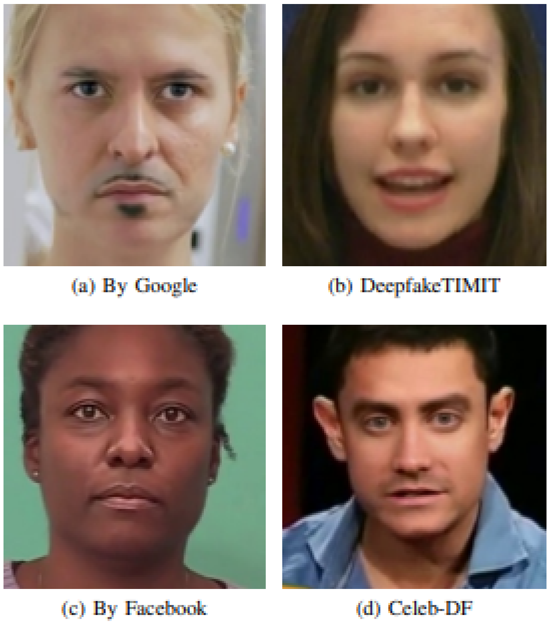


Fig. 1. Different Images from Libraries of Deepfaked photographs [4]

***Abstract*—Deepfakes are images, videos, audio, or another piece of media that was artificially generated using deep-learning models. Humans have difficulty telling advanced deep-faked media apart from genuine media. People often are confident in their ability to tell the genuineness of a piece of media, but are not any better at telling the fake media from the real media than most models. The difference is in how the models and humans tell faked images apart. Utilizing the way AI and other humans tell fake images from real images, training programs can be made to increase identification capabilities. Audio deepfaked detection is currently heavily understudied.**

I. Introduction

For a long time, videos, photos, and audio recordings were solid pieces of evidence when making a claim. These days, anyone can easily fabricate a video or photo to mislead someone. A deepfake is when someone alters or creates fake pieces of media, with the intent to seem authentic. Deepfakes are created with generative AI, hence the 'deep' in the name. See 1 for some examples of deepfaked faces.

Because they are made with AI, it is easy to confuse a deepfaked image with a real one. Occasionally, based on context, these deepfakes are obvious- Former President Barack Obama never played Fortnite with Joe Biden and Donald Trump. Often, though, it will be real people in realistic situations. These fake videos and photos have gotten people into a lot of trouble. Deepfakes have been used to get people into scandals [6]. In January 2024, Taylor Swift had a deepfaked image of her spread across Twitter. It caused a huge uproar and ended with Twitter banning searches of her for a day.

There also come the audio deepfakes. There are a lot of scams these days where people will scrape a voice from a publicly available recording, and use it to extort related individuals. They will call a relative or relation to the person whose voice they scraped, and ask them for money in a panic. It can be very difficult to tell whether the voice is synthetic or real. It's important for us to understand whether humans can detect deepfakes reliably, and if they can, ways we can train people to detect them better.

The research articles I will be discussing is a set of studies detailing humans analyzing deepfakes and attempting to tell whether a photo, video, or audio clip is faked or not. These studies may eventually help us find a way to definitively tell whether a human can determing if a piece of media is deepfaked or not.

II. Literature Survey

Deepfakes can be extremely difficult to recognize. But that begs the question, can humans recognize when a photo is fake- at all? Or is it impossible for anyone to accurately tell whether a photo is real or not? A lot of people definitely *think* they can tell the difference [2]. A few studies were done to see whether humans can detect deepfakes of faces. The researchers put up an online survey with around 280 participants, and had the participants guess whether the photo was a deepfake, or real. While a majority of the participants got certain images correct -some images were easier to label as fake than others- the results were very skewed. Total accuracy of detecting these fake images were between 30 and 85%. For one in five images, there was a less than 50% accuracy. That means that someone guessing randomly would have a better guess on 20% of the images. But unfortunately, all of the participants were very confident in their answers. In fact, the average confidence per-image never dropped below a five on a 10 point scale. See 2 for details on the confidence of the users compared to the accuracy.

So people are very confident in their ability to tell whether an image is deepfaked or not- but aren't actually able to deduce the difference all that well. The study also tested whether using advice and intervention methods to help with detection would work. It turns out that the use of reminders and advice did not make the participants more accurate with their assessments. They remained confident in their answers- no matter whether they were right or wrong. Overconfidence in your abilities to detect deepfakes can lead to letting down your guard. Especially on something like social media, where security is very relaxed. This coupled with the fact that multiple faces can be deepfaked to add credibility to a faked profile and the overconfidence can lead to information being stolen.

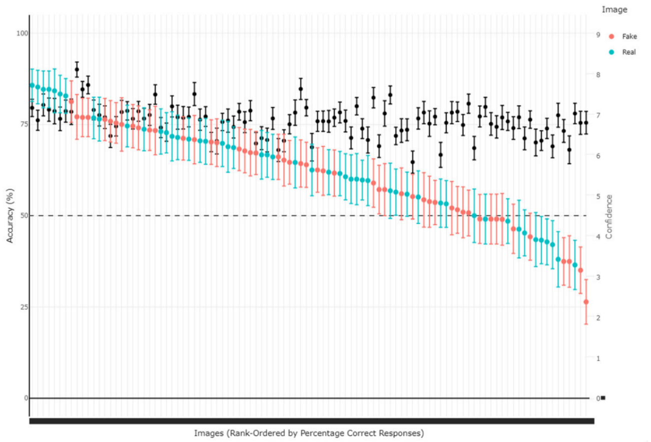
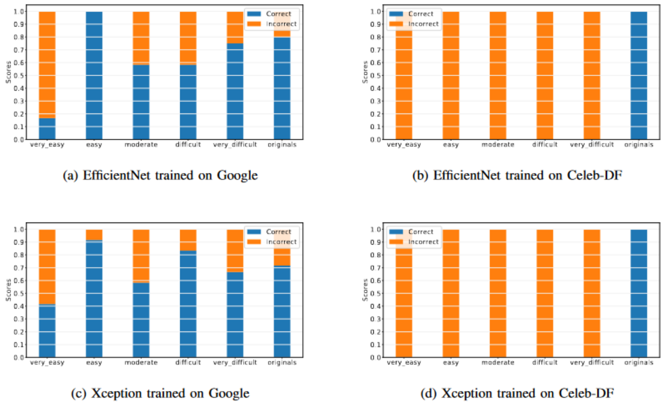


Fig. 2. Graph of confidence compared to accuracy [2]

The fact that there are little-to-no barriers for creating these deepfaked images and other forms of media means that its possible for almost anyone to create a fake account that seems extremely believable. Lots of the code is open-source and free-to-download online. If there were barriers to entry, such as a credential check in order to access the source code, there would be a lot less security risks.

Videos are a very common medium to deepfake. Due to the internet's reliance on visual and video formatting with the popularity of YouTube and tiktok, videos are one of the most commonly consumed pieces of media [1]. In 2021, YouTube was the most used platform on the internet. With the prevalence of video media, there is the worry of video-based deepfakes as well. A few studies have been done attempting to discover which is better at detecting deepfakes: humans, or machines? Some interesting results have popped up in these studies. For instance, in one study done in 2021, they tested three cases: Humans alone, Machines alone, and Humans with machines [3]. They performed two online studies with over 15,000 participants, asking participants to differentiate between deepfaked and real videos. They discovered that the humans and machines measured similarly in terms of accuracy. When it comes to the human-machine cooperative results though, they found that working together meant a more accurate result. The problem comes when the model is faulty. An inaccurate model leads to the machine deceiving the human, and therefore a much lower accuracy rating.

The study also noted that if a person's face was obscured, that humans have a harder time telling if it is deepfaked. Another study focused specifically on this recognition of faces in videos. This study was performed on 19 humans and 2 different algorithmic deepfake detection methods [4]. Each of the detection models were trained on two separate datasets, creating four total test models. The models' performance was then compared to the humans' performance. In the end, humans were tricked by high-quality deepfakes 75.5% of the time. While the models had an easy time detecting these deepfakes, they had difficulty detecting ones that to humans were clearly fake. Of course, this depends on the training data and model. as you can see from this graph 3, the training data for the model can largely change its results. The Google training data worked really well for the videos given for the test. The Caleb-DF training data most likely trained it to look for different things, causing it to fail the tests. It is also important to note that compared to the other studies, their human sample size was much smaller, and may have skewed the data.



(a) EfficientNet trained on Google

(b) EfficientNet trained on Celeb-DF

(c) Xception trained on Google

(d) Xception trained on Celeb-DF

Fig. 3. The results from the four different algorithms [4]

So compared to machines, humans do marginally worse at detecting deepfakes. So, is there anything we can do as humans to better train ourselves to detect deepfakes? A study done in 2021 offers a solution. They had 95 participants attempt to identify deepfaked videos, and tried to find trends between their conclusions [7]. While they did this, they tracked their eyesight as well: where they were looking when they were deciding whether it was a fake or not. They compiled this data into a heat map, which they analyzed to attempt to discover the process that humans use to deduce faked videos. Using this data, alongside a few AI models trained on deepfake datasets, they made a training process to try and increase the humans' success rate.

The training sessions did a few things. First, they walked through some easy examples for detecting deepfakes. They showcased points of interest and analysis strategies to help the person see different ways that deepfakes can be identified. They also used a real video to showcase the difference. They also pointed out the similarities, showing that there are certain features and points in the videos that tend to stay consistent. They also used the heat-map to point out specific points of interest. The eyes, for instance, was a big one. The AI trained models often focused on the chin and jawline to detect deepfakes, so they were included in the training sequence.

After running a test with a trained group and a control group, the results were conclusive: the group using their training system increased their overall accuracy. When a person is adequately trained to detect deepfakes, their accuracy increases. We can train humans to become better at detecting deepfakes.

Video deepfakes are important to study because they are both easy to find datasets for and people are getting much more aware of them. Audio deepfakes are just as important to pay attention to. Many people have been scammed through the use of audio deepfakes through the phone, and so learning to detect them is just as important, perhaps moreso. A study done to test the difference between AI and Human detection was done in 2022 [5]. In it, they published a survey in the form of a game online, for participants to play against an AI. 410 people participated, giving a large sample size plus demographic information. The study discovered that, in a vast difference to visual AI detection, Humans and AI performed relatively the same, and performed similarly towards specific modifications to the sound. Of course, that isn't including the extremely small artifacts in audio. A person most likely can't tell when a very small, artificial jump occurs in the waveform, but the AI can. However in realistic scenarios, the AI and the Humans performed relatively similar.

Familiarity with computers and IT skill level had no impact on the performance of an individual. This status just made the user confident. As we saw with the first research survey, overconfidence means nothing when it comes to deepfake analysis [2]. Your confidence in your answer does not matter when it comes to more advanced deepfakes. Two important factors did alter their performance rates though: age and familiarity with the language being spoken [5]. Participants who did not use English as their first language performed worse on average, if only slightly. Age, on the other hand, massively impacted test results. On a moderately linear trend, the older you got, the worse you performed. Those from thirty to forty on average performed at 75% accuracy. Meanwhile, people around ninety or a hundred were only at around a 65% accuracy score. The older you get, the worse you perform.

III. Future Direction

The future of deepfake detection is important. Whether it is increasing human ability to detect these deepfakes or improving AI models' abilities, it is imperative that we advance this as soon as possible.

Humans are very capable of detecting deepfakes and artificial elements when it comes to images and videos of faces. AI, however, are much better in broad strokes when detecting deepfaked images and videos. In the future, there are a few avenues we should be exploring. First is how to better integrate AI into a persons everyday video consumption. Development of an extension that scans a video as it's playing and pings the browser to let the user know that it's a fake can help with mitigating any forms of propaganda and malicious videos published on the internet. Similarly, use of AI on platforms like Twitter and Instagram could help with human detection.

Other ways to incorporate AI into everyday deepfake detection should also be researched. With more modern augmented reality capabilities, such as the Apple Vision Pro, we could incorporate this detection software into real life- in real time. Walking along the street, you see an ad playing before you and your glasses tell you that it's actually a fake video. Production of real-time augmented reality detection is a useful avenue. This can also be done on modern phones, with a specialized app.

Along these lines, research should be done on the difference between controlled-environment detection that is done in these labs, where we know the right answer, and unknown variables. If this research is to be implemented, we need to know it will work out in the real world. More research into whether humans and AI models working together has a positive or negative impact should be done, to see if these implementations would do more harm than good.

Research into what makes an AI training dataset useful is also important. A large amount of issues with the AI models were the datasets that they were trained on. We should get some research into what makes a deepfake dataset a good dataset. Obviously, diversity is an incredibly important factor, but we need specifics. With the black-box nature of deep learning, we can't truly know how the system works at the field's current state. What we can learn is how each piece of information we feed it alters it. If we learn that certain kinds of deepfakes do more damage to the overall system, we can purge those for a better result. Plus, with AI dataset poisoning programs like Nightshade now available, we need to be incredibly careful with what we incorporate into our datasets.

We need to create a publicly-available training program. With more research into deepfake detection training programs, specifically ones that go through the process of analysis and that showcase parts of a video that stay the same, we will be able to have the public at large more accurately identifying deepfaked images and videos. Add onto this a public information campaign on the dangers of deepfaked content and the public will be much better equipped to handle the modern internet.

There needs to be a lot more research on audio deepfakes. Currently, the state of AI audio deepfake detection is minimal. There is a lot of room for AI to get better at telling faked audio. Once these advances are made, we can utilize the information in similar ways as we have utilized the image-detecting AI models to better train humans to detect audio deepfakes. Audio deepfake detection technology can also be utilized in real time. An app on your phone that scans a call and checks to see if it is faked audio can help majorly decrease the number of people impacted by deepfaked audio scams.

Finally, utilizing audio deepfake detection and video deepfake detection might improve overall detection of deepfakes. Scanning both the audio and the visual input to check for artificial artifacts and other deepfake signs might improve or worsen detection, depending on if they conflict. More research should be done on the interaction between these two fields.

References

[1] Anderson, B., and Monica, A. Social media use in 2021, Apr. 2021.

[2] Bray, S. D., Johnson, S. D., and Kleinberg, B. Testing human ability to detect 'deepfake'images of human faces. *Journal of Cybersecurity 9*, 1 (2023), tyad011.

[3] Groh, M., Epstein, Z., Firestone, C., and Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences 119*, 1 (2022), e2110013119.

[4] Korshunov, P., and Marcel, S. Deepfake detection: humans vs. machines, 2020.

[5] Müller, N. M., Pizzi, K., and Williams, J. Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* (2022), pp. 85–91.

[6] SANER, E. Inside the taylor swift deepfake scandal” ”it’s men telling a powerful woman to get back in her box”, Jan 2024.

[7] TAHIR, R., BATOOL, B., JAMSHED, H., JAMEEL, M., ANWAR, M., AHMED, F., ZAFFAR, M. A., AND ZAFFAR, M. F. Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–16.