BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Deepfakes in facial recognition

Bc. Milan Šalko

Abstract

Deepfakes, media generated by deep machine learning that are indistinguishable to humans from real ones, have experienced a huge boom in recent years. Several dozen papers have already been written about their ability to fool people. Equally, if not more, serious may be the problem of the extent to which facial and voice recognition systems are vulnerable to them. The misuse of deepfakes against automated facial recognition systems can threaten many areas of our lives, such as finances and access to buildings. This topic is essentially an unexplored problem. The goal of this thesis is to investigate the technical feasibility of an attack on face recognition. The experiments described in the thesis show that this attack is not only feasible, but moreover, the attacker does not need many resources for the attack. The scope of this problem is also described in the work. At the end, there are also described several proposed solutions to this problem which may not be at all ordered for implementation.

xsalko02@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Deepfakes, synthetic media created by neural networks, have become a widespread phenomenon in recent years [1]. The speed at which this branch of artificial intelligence research is developing is breathtaking. Deepfake generation tools can create or edit content in a fraction of the time or cost compared to traditional video or image editing. Thus, anyone can create a deepfake in an online tool without deep knowledge of neural networks. This is evidenced by how deepfakes have proliferated on social networks, where they are typically used for entertainment. Or the fact that warnings about their potential misuse also appear in the mainstream media.

Deepfakes, despite their rather bad reputation, can also have a number of positive uses, whether it is in artificial intelligence news, where a synthetic image of a presenter 24/7 brings viewers up to date with the latest events, they can also help in the creation of films, school materials or in healthcare [2], where they give a voice back to people who have lost it due to illness. However, these uses are still in their infancy.

The realism of deepfakes, which many times makes it impossible for people to distinguish deepfakes from real photos, brings new risks [3]. Combined with their widespread availability to all and malicious intent, they can have a large number of negative uses. Deepfakes thus become a threat to individuals, where they can be used in new attacks. Examples are new forms of phishing, various forms of spoofing or the creation of videos and photographs to discredit a person. However, deepfakes also pose a threat to society as a whole, for example, their use in the creation of fake-news or in the general undermining of the credibility of public authorities in the eyes of citizens [4].

An equally, if not more, serious issue may be not only whether deepfakes can deceive people, but also the extent to which facial and voice recognition systems are vulnerable to them. Particularly when we consider that biometric facial recognition systems have become a normal part of our daily lives, partially displacing, for example, fingerprint-based login on mobile phones. Logging into apps and online banking are good examples that could be targeted by an attacker who decides to use deepfakes against these systems.

Recently, several tools have been developed and are available on the Internet, such as various freely available research papers or paid tools, which provide the possibility to create a fairly plausible deepfake. Currently, there is not much work that focuses on verifying the technical feasibility of attacking facial recognition systems. The goal of this work is to answer questions related to the use of deepfakes and

their use against facial biometrics. Based on these facts and the knowledge gained in the initial stages of this work, several attack vectors have been proposed. These should answer the fundamental question arising from this work, namely: how vulnerable are commonly available commercial facial biometric systems to attacks using deepfakes?

2. Experiments

The first attack vector is to verify the technical feasibility of an attack on facial recognition using deepfakes. This creation is resource and knowledge intensive so it is true that a good quality deepfake can be created using a phone and one does not need to have any greater knowledge of how deepfakes are generated. Next, we looked at commercial solutions on which we planned to investigate how vulnerable biometric systems are to deepfakes. These systems have applications in various domains such as online banking or identity authentication in applications.

Using freely available deepfakes generators and following a few guidelines, we created deepfakes that were of sufficient quality to be considered as valid input by the recognition system. It should be added that the naive approach of transferring a face to any other person proved to be flawed. When generating, care must be taken to ensure that the actor who serves as a basis also has similar facial features. This is a slight complication of the attack but not an unsolvable problem. In such cases, it is easy to hire an actor or use the many videos available on the Internet. When these principles are followed, the system identifies the media presented by the attacker as the selected victim. We can use these findings to design further experiments to reveal how large the problem is. Another interesting but expected finding is that one-shot systems cannot cope with synthesis in the case that it is necessary to turn the head 90° during verification, or when the person has to pass his/her hand in front of his/her face. This finding may serve as a possible protection against this type of attack.

In the following attack, we verified these findings. Since there is no dataset that can be used to check the robustness of biometric systems to deepfakes. We decided to take an already existing deepfake dataset from which we selected photos that meet the ICAO standard. Then we started with the measurements over the biometrics themselves. From the collected data and using statistical tests, we found that face recognition biometric systems can be fooled by deepfakes. Since deepfakes scores are not statistically similar to impostor scores. This means that biometrics cannot unambiguously reject deepfakes. This problem is magnified by the fact that the dataset used is already a few years old and the deepfakes in it are far behind the quality of today's deepfakes. It is also clear from the measured data that in the case of higher quality data, the similarity shifts more towards the genuine score.

3. Conclusion

Experiments have shown that modern commercial solutions are vulnerable to deepfakes. A solution to this problem may be to use multiple photos from a video instead of one to verify the identity of a person. It may also help if we require the user to perform actions that today's deepfakes generators still have a problem with, such as waving in front of the face or turning the head.

References

[1] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *CoRR*, abs/2004.11138, 2020.

[2] Younglawyerssection. Deepfakes: A balancing act, Sep 2019.

[3] Pavel Korshunov and Sébastien Marcel. Deepfake detection: humans vs. machines, 2020.

[4] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9:40–53, 11/2019 2019.