BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

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

Fake face detection in digital images

Bc. Tomas Lapsansky*

Abstract

Deepfakes are encountered more and more in today's world. They can help us in many areas of life and make life easier, but they also come with a lot of risks. It is therefore always necessary to know how to identify a deepfake so that it cannot be used for the wrong purposes, so detecting them is key. Convolutional neural networks have proven to be one of the best methods for solving such complex tasks. In our work we have therefore tried to make an overview of the available models and try to improve the best of them. The achieved results offer a large number of questions for the future that need to be further solved, such as how to evaluate such models not only in experiment, but eventually in the real world where we need to use them.

*xlapsa00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

In recent years, in modern technology, we can observe the rise and rapid development of neural networks and artificial intelligence, including deepfake photos and videos. This technology can generate very realistic audio-visual recordings which can be effectively applied in the entertainment and film industry, as well as in the artistic sphere or education and attract a higher interest from a broader spectrum of people. As it is usual, every technology has not only a light but also a dark side.

2. How is it done?

With the help of this technology, it is possible to carry out several types of attacks in a wide population spectrum, whether it is online harassment, political spectrum and influencing crowds of people by spreading popular and false news, but also false accusations and influencing the judiciary, or hiding criminal identities when verifying identity documents. Several of these attacks vectors [1] and also the real events related to them where these attacks were used are described in our work.

What makes this technology dangerous is also its availability. We can create deepfakes of audio-visual content or images using several pictures of a person and our mobile phone or computer. Several freely available applications can do this or a large number of neural network models [2, 3] that can help a more experienced user. What makes them dangerous is that nowadays, they are not even demanding computing power.

3. Detection

Since we pointed out the importance of preventing the use of this technology on the wrong target, it is necessary to know how to defend against these attacks effectively. Modern neural networks that we have researched can nowadays generate such a realistic image or video that it is often difficult to detect these attacks using human access alone. We, therefore, focused on convolutional neural networks and mapped state-of-the-art solutions of these networks that we tried to fit for the task we performed. We explored several models, of which EfficientNet [4] came out best. We then finetuned this network for the deepfakes detection task and tried to arrive at the best combination of hyperparameters.

For efficient evaluation of the visual record, we decided to create a complete pipeline for deepfake video (or image) processing. In the case of a video, it is necessary to split the video into individual frames and evaluate each frame separately. In the case of a picture, we evaluate only the image itself. Next, we perform face detection over each frame in the image using the pre-trained MTCNN model. We then cut the detected face to the dimensions required by our model and perform face centring on the centre of the input image. If the face is located on the edge of the image, it is necessary to fill the missing space with a constant colour.

We then tried to improve this architecture with several approaches. The first way we managed to improve the architecture is the so-called EfficientDet [5]. It is a model designed for object detection containing block and class predictions, which we decided to use. This approach increased the ROC (Receiver Operating Characteristic[1]) value on an unknown dataset by several percent. Due to the complexity of the task we decided to get as close as possible to the real-world scenario, so we did not compare the results only with the dataset on which they were trained (in our case FaceForensics [6]), but the main guiding component of the evaluation was the data from a completely different dataset Celeb-DF [7]. With the same number of training epochs, we were able to improve the overall results from 0.77 to 0.82 AUC with our modified architecture. By examining different combinations of freezing the prediction weights, we concluded that the best results are obtained when freezing the first 3 blocks of the network (the network contains 7 blocks followed by the prediction blocks from EfficientDet). For example, this behaviour can be caused by the fact that the network may contain a feature extraction in the first three blocks of the network needed for this task. We cannot confirm this theory in the current state, and a deeper investigation of the individual EfficientNet backbone blocks would be needed. The results of this model can be seen in Figure 1.

We then extended the model with the u-net architecture, which adds a fragmented mask of locations in case of detection, which assumes that the record is modified and therefore offers a better overview of the task in case of its evaluation. Although this approach can bring us closer to the behaviour of individual neural network blocks, at the current stage, we have not been able to use this architecture to improve the accuracy of the network. However, it has offered us a deeper insight into the functioning of the individual parts of the network.

4. Future work

Although we have succeeded in creating an architecture that can detect deepfake even over completely new data, this solution raises several other questions for the future that need to be further addressed.

The first issue is evaluating the solution's consistency and subsequent comparison with other existing detectors. Since there is no unified procedure for evaluating such a complex bag, each paper handles the evaluation in its own way. Even though uniform metrics are used for the calculation and can be used to evaluate the model, each solution uses different data for training. Thus, a model can be specifically trained for a dataset that is used to validate the model, and thus its results are better than those of models that have never seen this dataset. Our solution used the FaceForensics dataset for the training process and the Celeb-DF dataset for a more comprehensive evaluation.

Another question is how to control the way the model is trained effectively. When evaluating the model, we encountered a factor when after several epochs of fine-tuning, the model achieved great results on the validation dataset at the level of 95 % + accuracy. When evaluating the model after individual epochs over a new dataset, the results were not so good and rather deteriorated over time. In some cases, there was no convergence to a uniform result.

Due to these factors, we can argue that there is a need for a deeper investigation and improvement of the neural network evaluation process for such complex tasks and a unification of the evaluation process for individual complex tasks that need to be solved.

[1] https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

References

[1] Tina Brooks and etc. Increasing threat of deepfake identities.

[2] Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. Ghost—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022.

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, 2017.

[4] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[5] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection, 2020.

[6] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 2019.

[7] Yuezun Li. Celeb deepfake forensics. https://github.com/yuezunli/celeb-deepfakeforensics, 2023.