Research Project Proposal: Explanations for deepfake detection

Samuele Pino, samuele.pino@mail.polimi.it

March 30, 2020

1. Introduction to the problem

1.1. Forgery detection

Media manipulation exists since the times of the first analog photographs but the switch to digital media, besides bringing great technical benefits, made the manipulation process easier. Although today we can enjoy entire movies synthesized by computer graphics, at the same time threats linked to the misuse of such technologies are frighteningly increasing.

It is of crucial importance to always be able to tell apart real images and videos from synthesized ones, but sometimes it can be a hard task. Fake media are computer generated or manipulated images or videos with the precise goal to fool human eye: for this reasons sophisticated computational techniques like Deep Learning exist, and are still being studied, to perform this task with a higher accuracy than humans.

In the recent years new image and video manipulation techniques popped up on the internet, seriously challenging the classic manual manipulation detection methods. Such techniques take advantage of deep neural networks to skip the manual editing phase and automatically generate results so realistic to be almost indistinguishable from real ones to the naked eye.

We are going to focus on face identity swap, i.e. when the original face of a video is replaced with the face of someone else, but keeping the expression and movements of the original one. The most popular face swapping technique in videos is called DeepFake: it emulates the face of a source individual in different light and pose conditions, putting it on top of the face of a different target individual. Deepfakes are generated through deep neural networks (hence the name) specially trained on datasets of video representing the source individual.

The reasons why the research exploded are obvious and are mainly about privacy, reputation, politics and public security. In one possible scenario of high tension between two countries: what would happen if these techniques were used to show one of the two leaders declaring a war action against the opponent?

1.2. The explainability problem

Similarly to the generative techniques, some of the most effective methods to detect video forgeries rely on deep learning models. Despite the unpaired accuracy that such models present, they often tend to sacrifice the property of explainability that instead characterizes classic algorithms. Sometimes the deep learning model has a structure so complex that we do not know what are the precise reasons why it predicted that an input was real or fake.

An highly desirable characteristic of a deep learning classifier is that it gives "correct prediction for the correct reason". In order to understand this, it would be extremely useful to dig deeper in the understanding of the predictions, in other worlds to explain what the model is thinking when a specific input is given.

Explainability is an important property in many deep learning applications. It is essential in safety critical applications like industrial robots working together with humans or in self-driving cars. In our case, explaining the exact reason why a video is predicted as deepfake would increase the trust in the model and simplify the possible manual inspection.

If the police wanted to feed a fake detection model with a video suspected to be forged, maybe a simple "yes/no" output would not be sufficient as an evidence, while providing a reason why it is fake would result in a

stronger evidence. As another example, additional information on the prediction would help moderators of big web platforms in the hard task of filtering fake content.

2. Main related works

Deepfake detection field has seen a high number of contributions in the past two years. While some approaches tend to manually select a set of specific features that can help to discriminate forged videos from real ones [1, 2, 3, 4, 5, 6, 7], others prefer to let the model detect and learn the features by itself, in a supervised environment [8, 9, 10, 11, 12, 13, 14]. For the latter case the most used technology is CNN, sometimes combined with RNN to take advantages of the temporal information present in videos [1, 9, 2, 6, 13].

Although some CNN deepfake detectors output also a mask locating the predicted fake areas of the image [5, 11, 12, 7], most of them do not provide any mask or other explicit explanation for their predictions. The problem of explaining the behaviour of deep neural networks has been tackled by several works, both in a white box and black box approach. Buhrmester et al. widely talk about the state of the art black box explainers in their survey [15].

To the best of our knowledge, there is no work addressing explainability and its importance in the field of deepfake detection.

3. Research plan

The goal of the research will be to investigate the explainability of deepfake detection models. We will address the following questions:

- Is it possible to explain deepfake predictions in an interpretable fashion?

  Although there are some deepfake detectors that produce separate location information together with the prediction, to explain which parts of the image are predicted to be forged, to the best of our knowledge nobody has tried black box explanation models on deepfake detectors. The reason to investigate this aspect is the wide applicability of the same explainer over different architectures of deep neural networks.

- Do apparently similar deep learning models use actually different feature to predict if a video is a deepfake?

  Deep neural networks for detection appear in a multiplicity of architectures, but it is hard to understand how a change in the architecture will affect the reasoning of the network. Since different models also present different accuracies in detection, it is likely that they focus on different features, or on the same features but in different ways. Applying explainers to different deepfake detectors will let us explain what are the different features that each model values for its prediction and, most importantly, why certain models perform better then other.

- Are black box techniques as effective as model-specific or by-design techniques for generating explanations?

  Another question is whether black box explanation can compete with other types of explanation, like white box or model specific ones. As we said, some models already produce a graphical explanation for their predictions, usually in the form of a gray scale mask. Our goal will be also to compare the two types of explanation and understand if black box option can reach an effectiveness similar to explanations provided by design by the detection model.

- Can we apply explanation techniques to video inputs, taking advantage of the time information?

  Most of the explainers work on the image/frame level, while some deepfake detectors work on the video level using also time as an input. For this reason we would like to investigate the possibility of using the temporal information encoded into videos to extract a more coherent and sound result from explanations.

Step	Description	February	March	April	May	June	July	August	September
0	State of the art								
1	Detectors implementation								
2	Explainers implementation								
3	Results collection								
4	Result evaluation								
5	Writing thesis								

Figure 1: Gantt diagram of the planned work.

- How can we use this further knowledge to improve the model or the dataset?

  Explanations provide interesting insights on the reasoning of classification models. We will investigate the possibility of using these insights either indirectly, by building better models, or directly, by using them in the training phase.

- How detailed can our explanation be? How intuitive can they be, while keeping a certain degree of completeness?

  There is a trade off between completeness and intuitivity of an explanation. Deep learning models contain millions of parameters, therefore a complete explanation, in terms of all these parameters, would be useless since impossible to understand. On the other hand, a too intuitive explanation may not carry enough information for our purposes. During the work, we will choose an appropriate compromise between these two properties.

The research type will be experimental: we will contribute by addressing a not yet investigated aspect of the field, we will perform experiments and compare them with the current state of the art.

The research plan will be characterized by the following steps:

1. Implementation and execution of the baseline detectors

   We will choose the state of the art techniques for deepfake detection and we will implement them, training them where necessary. We will execute them on public benchmarks and asses their performance.

2. Implementation and execution of known explanation algorithms

   We will execute known algorithms for black box model explanation on the deepfake classifier. This step will allow us to get some better insights on why different algorithms have possibly different benchmark scores.

3. Collecting results

   Several explanation techniques will be tested and compared, addressing accuracy of the explanation, understandability, and effectiveness in improving the model.

4. Evaluation design

   A metric for objective evaluation of the results will be designed. It may be based on public benchmark tests or on a survey performed on human subjects about the performances.

5. Writing the thesis

A Gantt diagram of the working plan is shown in Figure 1.

References

[1] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

[2] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[3] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2018.

[4] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.

[5] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, Jan 2019. doi: 10.1109/WACVW.2019.00020.

[6] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[7] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection, 2019.

[8] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.

[9] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, Nov 2018. doi: 10.1109/AVSS.2018.8639163.

[10] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images, 01 2019.

[11] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos, 2019.

[12] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation, 2019.

[13] E. Sabir, K. Cheng, A. Jaiswal, W. Abdalmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Workshop on Applications of Computer Vision and Pattern Recognition to Media Forensics with CVPR*, 2019.

[14] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.

[15] Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey, 2019.