Robustness and Generalization of Synthetic Images Detectors

Davide Alessandro Coccomini[1,2], Roberto Caldelli[3,4], Claudio Gennaro[1], Giuseppe Fiameni[5], Giuseppe Amato[1] and Fabrizio Falchi[1]

[1]ISTI-CNR, Pisa, Italy
[2]University of Pisa, Pisa, Italy
[3]CNIT, Florence, Italy
[4]Universitas Mercatorum, Rome, Italy
[5]NVIDIA AI Technology Center, Italy

Abstract

In recent times, the increasing spread of synthetic media, known as deepfakes has been made possible by the rapid progress in artificial intelligence technologies, especially deep learning algorithms. Growing worries about the increasing availability and believability of deepfakes have spurred researchers to concentrate on developing methods to detect them. In this field researchers at ISTI CNR's AIMH Lab, in collaboration with researchers from other organizations, have conducted research, investigations, and projects to contribute to combating this trend, exploring new solutions and threats. This article summarizes the most recent efforts made in this area by our researchers and in collaboration with other institutions and experts.

Keywords

Deepfake Detection, Deep Learning, Super Resolution

1. Introduction

Deepfakes and synthetic media are becoming more prevalent and realistic day by day, presenting society with an increasingly urgent challenge, learning to distinguish reality from fiction effectively. These fake content can and are continually being used to spread disinformation, create smear campaigns, and manipulate reality with potentially devastating impacts for anyone who may end up a victim. To contrast this phenomenon, research has been conducted in recent years creating detectors, often based on deep learning techniques, that can classify a piece of content (such as an image) as realistic or fake. Despite many efforts, this discrimination capability is still insufficient today with many open problems in the field of deepfake detection. One example above all is that of generalization[1, 2]. In fact, deepfake detectors, although particularly effective in detecting images generated or manipulated by the same methods they are trained on, fail when using different and novel techniques. Another problem is that of adversarial attacks, strategies of camouflaging traces, enhancing fake content or ad-hoc manipulations designed to fool the detector, which can be used to make detection even more complex. Deepfake detection models must therefore be designed so that they provide a high degree of robustness to possible adversarial attacks and also be able to effectively distinguish deepfakes without raising false alarms. For this reason, AIMH Lab at ISTI CNR has carried out numerous research attempts to explore new innovative techniques to advance this field but also to highlight possible hidden dangers that may damage the efforts made in previous research, representing dangers that detection systems may encounter in the real world. In particular, this paper summarizes the efforts made in [3] and [4].

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Davide Alessandro Coccomini
davidealessandro.coccomini@isti.cnr.it (D. A. Coccomini); roberto.caldelli@unifi.it (R. Caldelli); claudio.gennaro@isti.cnr.it (C. Gennaro); gfiameni@nvidia.com (G. Fiameni); giuseppe.amato@isti.cnr.it (G. Amato); fabrizio.falchi@isti.cnr.it (F. Falchi)
0000-0002-0755-6154 (D. A. Coccomini); 0000-0003-3471-1196 (R. Caldelli); 0000-0002-3715-149X (C. Gennaro); 0000-0001-8687-6609 (G. Fiameni); 0000-0003-0171-4315 (G. Amato); 0000-0001-6258-5313 (F. Falchi)

2. Research Works in Deepfake Detection

In this section, we present our most recent works in the field of Deepfake Detection, highlighting the contributions and discoveries made.

2.1. Super-Resolution as an Adversarial Attack for Deepfake Detection

Super-resolution (SR) algorithms are a set of techniques designed to improve the resolution of an image. Starting with a low-resolution one, through deep learning techniques, it is scaled up to a higher resolution. During this

Model	Forgery	Attacked	Accuracy ↑	FNR ↓	FPR ↓	AUC ↑
Swin	Deepfakes	✖	95.3	5.9	3.6	99.1
		✓	90.7	6.1	12.4	97.4
	NeuralTextures	✖	87.1	12.9	12.8	94.9
		✓	81.5	13.2	23.9	90.4
	Face2Face	✖	95.2	6.3	3.3	98.9
		✓	87.0	24.4	1.7	96.1
	FaceSwap	✖	95.2	4.9	4.6	98.6
		✓	86.4	21.9	5.3	93.9
	FaceShifter	✖	94.4	7.2	4.1	98.7
		✓	89.0	18.9	3.1	97.4
Resnet	Deepfakes	✖	95.6	5.5	3.2	99.2
		✓	95.0	6.9	10.1	98.9
	NeuralTextures	✖	87.1	12.9	12.8	94.9
		✓	81.5	13.2	23.9	90.4
	Face2Face	✖	95.9	5.0	3.2	98.9
		✓	91.2	14.4	4.7	97.6
	FaceSwap	✖	95.9	6.4	1.9	99.1
		✓	88.5	21.1	2.6	95.6
	FaceShifter	✖	95.3	6.1	3.4	98.8
		✓	85.9	24.8	3.3	95.2

Table 1
Evaluation on Faceforensics++[5] test set. Each set is composed of half pristine and half fake images. The Attacked column indicates if the SR-attack has been applied to the images. The attack is applied to fake and pristine images.

process, some aspects of the image may change. For example, some previously visible details may become more blurred, totally disappear, or, conversely, be emphasized and brought to light. Deepfake generation algorithms commonly tend to introduce some more or less visible artifacts. In the case of human faces, these artifacts can be, for example, anomalies in pupils, contours of lips, eyes or ears, or accessories. Typically, deepfake detection models learn to recognize the specific anomalies introduced in manipulated images and, because of them, can discriminate between pristine and fake content. In [4], we explored whether Super Resolution techniques can be used as an adversarial attack to camouflage artifacts introduced by deepfake generations approaches.

To do this, we proposed an SR-attack pipeline, whose purpose is to disguise artifacts present in deepfake images while still trying to keep the appearance as unaltered as possible. The pipeline begins with the detection of a face from the deepfake image that is subsequently downscaled by a factor K using interpolation techniques. The resulting image is restored to its initial resolution through a super-resolution approach. The resulting face can eventually be reinserted into the original image, resulting in the camouflaged image.

An example of the impact of the proposed SR-attack is shown in Figure 1. As the figure shows, the attack leads to the removal of artifacts introduced by deepfake generation techniques, such as noise around the mouth, and thus makes their detection extremely complex. In fact, to distinguish a counterfeit image from a pristine one we often rely on observing these artifacts and oddities that can be introduced by the manipulation process. The fact that super-resolution leads to their removal or attenuation qualifies it as a potentially effective attack against deepfake detectors but also against the human eye itself.

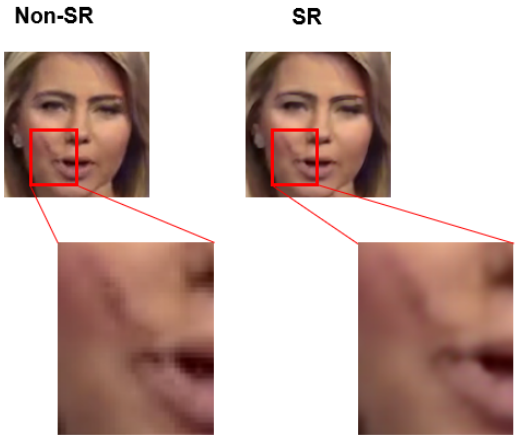

Figure 1: Example of the impact of SR attack on fake images. On the left, an example of manipulated face with a zoom on the artifacts around the mouth. On the right the same face but after the application of the SR-attack. From the zoom on the second one it can be seen how the artifacts are drastically smoothed.

The usage of the SR-attack conduct to a blurring effect on the artifacts introduced in the fake images and this makes them more difficult to detect. This is pretty evident in terms of performance; in fact, the use of the attack drastically degrades the performance of classifiers trained to do deepfake detection.

Table 1 shows the accuracies of a Swin Transformer[6] and a Resnet50[7] on images manipulated with different techniques, considering them before and after SR-attack. The dataset used is FaceForensics++[5] and the deepfake generation methods considered are Deepfakes[8], Face2Face[9], FaceSwap[10], FaceShifter[11] and NeuralTextures[12]. This allows us to evaluate the effect of the super-resolution attack on different types of manipulations, thus highlighting on which it is more or less effective. Both the models are trained on the FaceForensics++ training set considering for the construction of the fake class, the same deepfake generation method used for the test.

According to our experiments, for both the considered models, when images are attacked with the proposed approach, there is an increase in False Negative Rate, particularly on some methods namely Face2Face, FaceSwap and FaceShifter. Others, however, are found to be more robust to attack, namely Deepfakes and NeuralTextures on which, however, there is an increase in False Positives. This behavior indicates that in some cases, the use of super-resolution techniques could lead to the elevation of false alarms, leading models to identify legitimately enhanced images through these approaches as deepfakes.

The latter result highlights a problem that could prove crippling to traditional deepfake detectors and could prevent their deployment in the real world. Indeed, it is plausible to think that on social networks it will become increasingly common to improve the quality of one's photos through Super-Resolution techniques. If this were to happen and in parallel the deepfake detectors were unable to understand that these are legitimate images but instead end up mistaking them for deepfakes, the number of false alarms would be such as to prevent their effective deployment on a large scale. It is therefore necessary on the one hand to defend against the malicious use of super-resolution to disguise artifacts introduced by manipulation techniques but also to make detectors robust so that they are able to recognize legitimately augmented images.

In these experiments we used only EDSR[13] as the basis of our attack but the proposed attack can be conducted using different types of SR algorithms (such as BSRGAN[14]), and depending on the peculiarities of each, greater or lesser effectiveness can be achieved on each specific deepfake generation method. The choice of the K factor also has an impact in that as it increases, the detectors' errors increase but the quality of the image itself also deteriorates. Therefore, it is crucial to choose the SR algorithm and the value of the K factor appropriately in order to achieve maximum effectiveness from the attack.

2.2. Future Works

In this section we expose some of the future works we are working on either as extensions of previously presented works or as new applications and solutions for effective deepfake detection.

2.2.1. Robustness of Deepfake Detectors

The fact that Deepfake Detectors are susceptible to the use of SR techniques on images, whether fake or pristine, exposes a serious problem in their use in the real world and therefore requires that more studies be conducted to make them robust to this kind of content. It is necessary to find an effective way to make the models robust to this attack for example by introducing super-resolution as data augmentation during training.

The attack itself can also be further explored and improved by going to identify the optimal K value and corresponding SR method as well as experimenting with different strategies for applying the attack, such as focusing on a frame rather than a detected face.

2.2.2. Deepfake Detection without Deepfakes

As stated before, one of the most stringent problems in the field of deepfake detection is that of generalization. Indeed, there is ample evidence that detectors tend to learn to effectively recognize deepfake content obtained through methods used to construct their training set, but fail when they need to classify content obtained through novel techniques. This leads to a total inadequacy of conventional deepfake detectors in being used in the real world. In fact, new deepfake generation techniques are continually being created, and it would be impractical to retrain the model each time to introduce every single possible method. In the context of synthetic images, this tendency of deepfake detectors stems from the fact that each generator introduces a specific fingerprint into the image[15, 16, 17]. It tends to be invisible to the human eye but involves the presence of structured patterns in the frequency domain (grids, symmetric peaks, halos, etc.).

From the observation of this phenomenon, as a future work we are exploring a new training technique for deepfake detectors that tries to stimulate the model to recognize the presence of structured patterns in the frequency domain and not to learn a specific fingerprint. The preliminary results of this approach can be found in [3].

We propose to reproduce prototype structured patterns inspired by what we observed from the fingerprints

actually introduced by generative patterns of various types. These patterns, in frequency, are injected in pristine images and considered as "fake" in the training phase. During the training, we show the model pristine images and others on which a pattern has been applied, indicating them to the model as fakes.

From our preliminary experiments, we demonstrated that models trained on this proto-task, are extremely effective at identifying synthetic images despite never really seeing one in the training phase.

The use of synthetic patterns may also be used to support traditional training of deepfake detectors, introducing them only occasionally and still maintaining deepfakes in the training set. In addition, it will be possible to experiment with a virtually infinite number of patterns by searching for the most effective ones and studying their impact.

3. Conclusions

Carrying out research in the field of deepfake detection is an increasingly pressing need because of the multitude of techniques that now make it possible to produce synthetic or manipulated content with an increasing degree of credibility. As shown in our recent research, it is critical to explore both innovative methods to try to improve the capability of deepfake detectors looking for new training approaches and techniques. This could be needed to overcome the pressing problem of generalization. On the other hand, it is important to find new solution to face the possible risks and unexpected situations that these models might encounter in the real-world that could invalidate their potential, such as the adversarial attacks.

3.0.1. Acknowledgments

This work was partially supported by the following projects: Tuscany Health Ecosystem (THE) (CUP B83C22003930001) and SERICS (PE00000014, MUR PNRR - NextGenerationEU), AI4Media (EC H2020 - n. 951911) and AI4Debunk (Horizon EU n. 101135757), FOSTERER (Italian MUR PRIN 2022). We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support.

References

[1] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, G. Amato, Cross-forgery analysis of vision transformers and CNNs for deepfake image detection, in: International Conference on Multimedia Retrieval Workshop, 2022. doi:10.1145/3512732.3533582.

[2] D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, On the generalization of deep learning models in video deepfake detection, Journal of Imaging (2023). doi:10.3390/jimaging9050089.

[3] D. A. Coccomini, R. Caldelli, C. Gennaro, G. Fiameni, G. Amato, F. Falchi, Deepfake detection without deepfakes: Generalization via synthetic frequency patterns injection, 2024. arXiv:2403.13479.

[4] D. A. Coccomini, R. Caldelli, G. Falchi, Amato, F. Falchi, C. Gennaro, Adversarial magnification to deceive deepfake detection through super resolution (to appear), in: Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2023.

[5] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. doi:10.1109/ICCV48922.2021.00986.

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. doi:10.1109/CVPR.2016.90.

[8] Deepfakes, 2018. URL: https://github.com/deepfakes/faceswap.

[9] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, Commun. ACM (2018). doi:10.1145/3292039.

[10] K. M., Faceswap, 2017. URL: https://github.com/MarekKowalski/FaceSwap/.

[11] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Faceshifter: Towards high fidelity and occlusion aware face swapping, 2020. arXiv:1912.13457.

[12] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: Image synthesis using neural textures (2019). doi:10.1145/3306346.3323035.

[13] B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee, Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.

[14] K. Zhang, J. Liang, L. Van Gool, R. Timofte, Designing a practical degradation model for deep blind image super-resolution, in: IEEE International Conference on Computer Vision, 2021.

[15] D. A. Coccomini, A. Esuli, F. Falchi, C. Gennaro, G. Amato, Detecting images generated by diffusers, 2023. doi:10.48550/arXiv.2303.05275.

[16] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, CNN-generated images are surprisingly easy to spot... for now, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020, pp. 8692–8701. doi:10.1109/CVPR42600.2020.00872.

[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, in: Int. Conf. on Acoustics, Speech and Signal Processing, 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10095167.