Exploring the Adversarial Robustness of AI-generated Image Detectors

Thomas Lazzerini, Samuele Cappelletti, Martina D'Angelo
University of Trento

***Abstract*—**

I. Introduction

Synthetic images are now flooding the real world. From online dating sites to social media, fake profiles and scams are everywhere. The problem with synthetic images is that, while some of them are funny and harmless, others could be harmful, they could be exploited by malicious users [1]. In relation to this, in the image forensic field there is a continuous fight between *fake image detectors* and *adversarial attacks*. On one hand, the detectors try to distinguish fake images from real ones, while, on the other hand, the attacks try to trick the detectors by manipulating the images (both real and fake ones). In order to detect fake images, we can exploit the traces/artifacts that fake image generators leave on the generated images. To do so, we have two main types of techniques: the *low-level forensic techniques* and the *high-level forensic techniques*. To former focuses on the pixel-level artifacts, which are almost invisible to the human eye. The latter focuses on physical inconsistencies and on repeated and uniform patterns, both of which are mostly visible to the human eye. Examples of physical inconsistencies are lighting, shadows, reflections or vanishing points inconsistencies [2][3]. While, an example of repeated and uniform patterns, typical of GAN-based image generators, is the generation of the mouth, the nose and the eyes always in the same position [4]. In general, we prefer to rely on low-level artifacts since fake image generators are becoming smarter every day, thus they are learning to generate always more realistic images, with fewer physical inconsistencies.

II. Detectors

In this section we will briefly describe a couple of fake image detectors: one uses CLIP to extract the feature vectors from the images [5] and one identifies the low-level traces/artifacts by training a GAN and a Diffusion Model [6].

A. CLIP

B. PIZZA

III. Attacks

A. Mimicry

B. SD Laundering

C. White Black

D. Adversarial Robustness

IV. Experiment

V. Conclusions

References

[1] N. Carlini and H. Farid, "Evading deepfake-image detectors with white- and black-box attacks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 658–659.

[2] H. Farid, "Lighting (in) consistency of paint by text," *arXiv preprint arXiv:2207.13744*, 2022.

[3] ——, "Perspective (in) consistency of paint by text," *arXiv preprint arXiv:2206.14617*, 2022.

[4] S. Mundra, G. J. A. Porcile, S. Marvaniya, J. R. Verbus, and H. Farid, "Exposing gan-generated profile photos from compact embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 884–892.

[5] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the bar of ai-generated image detection with clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4356–4366.

[6] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.