Exploring the Adversarial Robustness of AI-generated Image Detectors

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

***Abstract*—Semi-supervised image classification is a machine-learning task in which a model is trained using a combination of labeled and unlabeled data. This paper consists of a high-level survey of semi-supervised image classification literature and expands on its main theoretical and practical challenges, providing a taxonomy of the most popular semi-supervised learning algorithms. ciao come va**

I. Introduction

Semi-supervised image classification is nowadays a hot research topic. The objective is to tackle the major issue of Supervised Learning: the scarcity of labeled data...

As discussed in [1], semi-supervised learning is an important area of research.

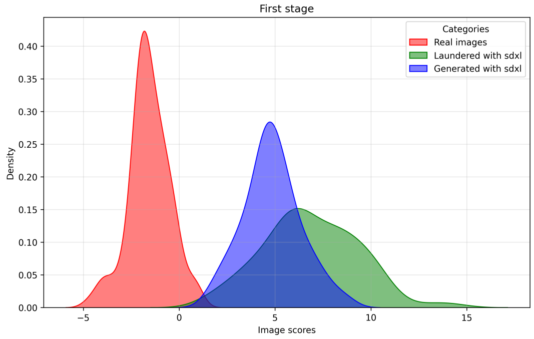

Fig. 1. Esempio di immagine.

II. Detectors

A. CLIP

B. Detection of Images by Diffusion Models

Lately, *Diffusion Models* gained the spotlight in the image generation community, allowing for unmatched test-to-image photorealism and diversity. These new powerful tools are a new asset in the hands of malicious users, posing new challenges to the forensic community.

Most SoTA detectors exploits low-level artifacts, not visible by a human eye, introduced during the generation phase by GAN generators. The study in [1] suggests that, as can be seen in Fig. 2, similar traces can be found also in DM-generated images

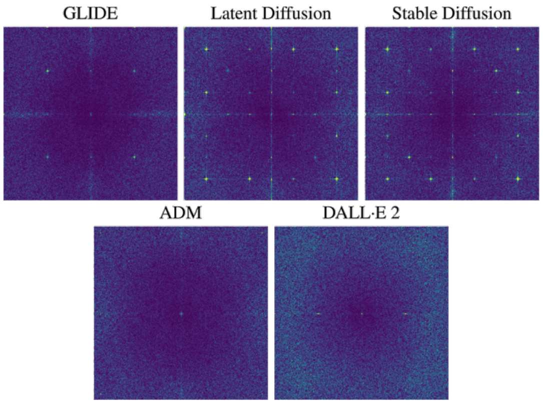

Fig. 2. Fourier transform of of the fingerprint of some DM presented in [1]

The study in [1] also provides interesting evaluation results, comparing the performances of several SoTA detectors over different GAN and DM generators both in ideal case (uncompressed images) and real case (compressed and resized using the guidelines in [2]). These evaluations highlight how performances vary significantly between the models, due to the differences in their artifacts, therefore suggesting generalization difficulties (for example, in classifying a DM images with a GAN training and vice versa). Despite these difficulties, the inclusion of DM during training and performing an careful calibration procedure, like the one suggested by [3], may help the generalization over similar architectures, despite not providing reliable results on out-of-training artifacts.

III. Attacks

A. Mimicry

B. SD Laundering

C. White Black

D. Adversarial Robustness

IV. Experiment

V. Conclusions

References

[1] 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.

[2] R. Corvi, D. Cozzolino, K. Nagano, and L. Verdoliva, “IEEE Video and Image Processing Cup,” https://grip-unina.github.io/vipcup2022/, 2022.

[3] J. Platt, “Probabilistic outputs for support vector machines and comparison to regularized likelihood methods,” *Advances in Large Margin Classifiers*, 1999.