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.

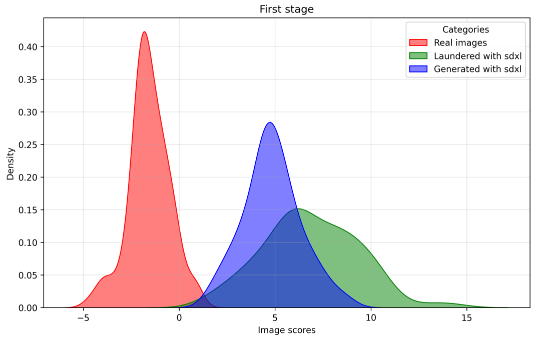

Fig. 1. Esempio di immagine.

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

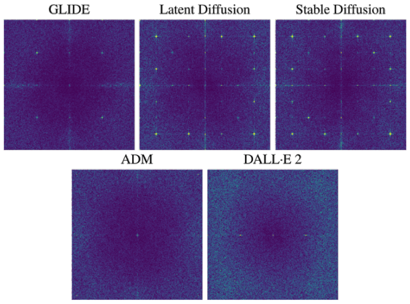

Fig. 2. Fourier transform of of the fingerprint of some DM architectures (*GLIDE* [2], *Latent Diffusion* [3], *Stable Diffusion* [4], *ADM* [5], *DALL·E 2* [6]) presented in [1]

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 [7]). 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 [8], may help the generalization over similar architectures, despite not providing reliable results on out-of-training artifacts.

III. Attacks

Despite the powerful detectors at our disposal, there exists many users that aim at attacking such detectors, in order to hide traces of their forgeries or also to introduce traces typical of generated images, to hide disguise content as fake. In the following chapter, some newly developed attacking techniques are discussed, to provide a general overview of the attacker-side.

A. Mimicry attack against image splicing forensic

As stated in [9], this *mimicry adversarial attack* can be used to hide image manipulation while forcing the detector to detect arbitrary ones by applying a gradient based optimization approach. Applied at large scale, this would cause high false-alarms, producing an effect similar to *DoS* attacks while undermining the reliability of the target detector.

The attack strategy proposed in [9] involves splitting the image in uniform patches and use these to compute a target representation for both the *pristine patch* t_p, computed from the pristine patches, and the *forged patch* t_f, computed from the forged patches. The function used for computing such target representations needs to be defined for each detector for the attack to be effective, this due to the fact that different detectors exploit different features. Once the targets have been computed, a gradient-based iterative approach is applied the each patch of the manipulated image, in order to make the patch feature representation more similar to the respective target's feature representation. A visualization of such iterative approach can be seen in Fig. 3

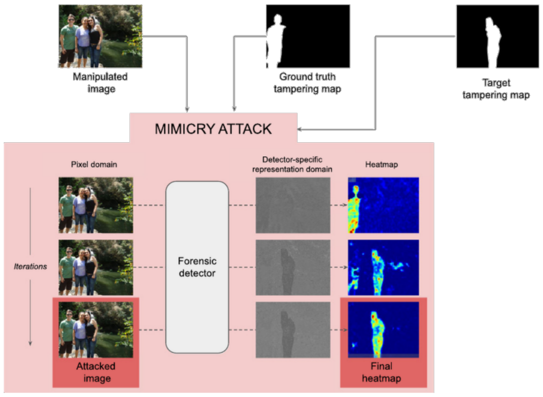

Fig. 3. Visualization of the mimicry attack strategy and its effects proposed in [9]. The "*ground truth*" tampering map represents the real forgery, while the "*target*" tampering map represent the arbitrary forgery the attacker wants the detector to output

The evaluation results reported in [9] suggests this attack is highly effective, both in hiding the real forgery and also highlighting a "decoy" forgery arbitrarily introduced. Two image detectors were tested, *Noiseprint* [10] and *EXIF-SC* [11], over two different datasets, *Columbia* [12] and *DSO-1* [13]. Several threshold-based and threshold-less metrics have been tested, with the latter being more important from the attacker point of view since the threshold values are unknown to him.

Another interesting result presented in [9] regards the *cross-detector* scenario, in which the attack is performed targeting a specific detector but then another is used in the evaluation. Also, *stacked attacks* are considered, in which an image is sequentially attacked against different detectors. An evaluation in these regards reveal mixed results: a misaligned attack in not effective, while the performances of a stacked attack are highly dependant on both the order of the attacks and the detector used in the evaluation. Nevertheless, this is an interesting scenario open for further studies.

B. Image Laundering with Stable Diffusion

Differently from "classic" diffusion models, like *Latent Diffusion*, *Stable Diffusion* models allow the users to provide an initial image as input [14] [15] [16] [17]. This image will be superimposed with noise and modified by the model according to the textual prompt. The weight of such modifications can be set via a dedicated strength parameter in the range $[0, 1]$.

Processing images in such pipeline using a strength parameter equal to 0 produces outputs with the maximum similarity to the inputs: the image is encoded and decoded right away, without any denoising step. As suggested in [18], this process could be exploited by malicious users in order to mask real content as synthetic. In fact, the encoding/decoding is sufficient to introduce enough artifacts into the real image to make it synthetic in the eyes of numerous detectors. This practice is known as *image laundering*.

The study in [18], proposes a two-step architecture, visualized in Fig. 4, as solution to efficiently discriminate between real, fully synthetic and laundered images. Such architecture is inspired by [19], in which the image is split into multiple random patches, a score is assigned to each one and the average o the highest scores provide the global score of the image: a positive score suggest a synthetic image, while a negative score a real one. Despite the good results, this backbone architecture alone is unfit for the laundered image detection task, hence the introduction of the 2 steps: the first step discriminate real from synthetic images, while the second step discriminate fully synthetic from laundered images.

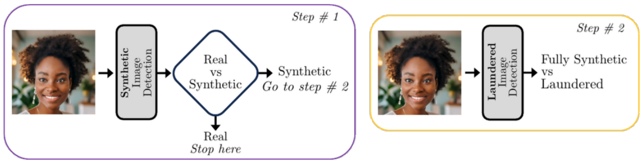
Step # 1
Synthetic Image Detection
Real vs Synthetic
Synthetic
Go to step # 2
Real
Stop here
Step # 2
Laundered Image Detection
Fully Synthetic vs Laundered

Fig. 4. Visualization of the 2 step architecture proposed in [18] for the laundered image classification task. The first step discriminate real images from synthetic (both fully and laundered) ones, while the second step discriminate fully synthetic images from laundered ones

The evaluation provided in [19] about such 2 step pipeline are good: the first step reach a good separability between real and synthetic image, while the second step reach almost perfect results over multiple models and multiple metrics, with only minor decreases in performances when post-processing operations like JPEG compression and resize are applied.

C. Adversarial Robustness

IV. Experiment

In this first phase of the project, our team executed a preliminary experiment, to asses the capability of the laundering attack, from the section III-B, on the dataset *TrueFake* provided by the *MMLAB* team.

The first phase of the experiment consisted in recovering 25 real images and, given their large size, extract 4 patches of size 1024×1024 from each of them, for a total of 100 real patches. Next, such patches were laundered, with a *denoising* parameter of 0, using the model *sd_xl_base_1.0* [20]. Lastly, a total of 100 fully synthetic images, generated by *Stable Diffusion XL*, were collected from the *TrueFake* dataset in equal quantity from each category available.

This small dataset was submitted to the 2 step pipeline from [18], visualized in Fig. 5, obtaining interesting results. The first step, as can be seen in Fig. 5, yielded similar results as [18], with a good separability around threshold 0. On the other hand, the second step, in Fig. 6, yielded results slightly different from [18], in particular the fully synthetic images had a average score of about 1, where [18] reported good separability at threshold 0.

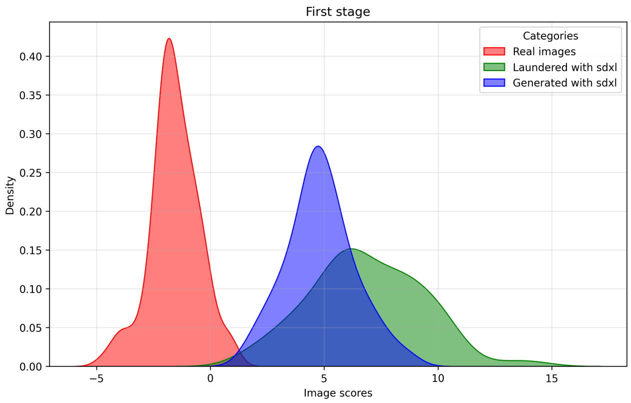

Fig. 5. Results of the first step of the pipeline from [18] using images from *MMLAB TrueFake* dataset

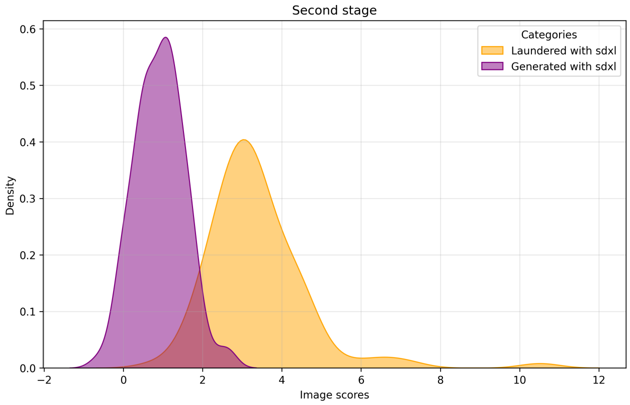

Fig. 6. Results of the second step of the pipeline from [18] using images from *MMLAB TrueFake* dataset

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] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "Stable diffusion," https://github.com/CompVis/stable-diffusion, 2022.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125v1*, 2022.

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

[8] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *Advances in Large Margin Classiers*, 1999.

[9] G. Boato, F. G. De Natale, G. De Stefano, C. Pasquini, and F. Roli, "Adversarial mimicry attacks against image splicing forensics: An approach for jointly hiding manipulations and creating false detections," *Pattern Recognition Letters*, vol. 179, pp. 73–79, 2024.

[10] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2019.

[11] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 101–117.

[12] T.-T. Ng, S.-F. Chang, and Q. Sun, "A data set of authentic and spliced image blocks," *Columbia University, ADVENT Technical Report*, vol. 4, 2004.

[13] T. Carvalho, F. A. Faria, H. Pedrini, R. d. S. Torres, and A. Rocha, "Illuminant-based transformed spaces for image forensics," *IEEE transactions on information forensics and security*, vol. 11, no. 4, pp. 720–733, 2015.

[14] Computer Vision and Learning LMU Munich, *Stable Diffusion*, 2022 (accessed June 20, 2024), https://github.com/CompVis/stable-diffusion.

[15] S. AI, *Stable Diffusion Version 2*, 2022, https://github.com/Stability-AI/stablediffusion.

[16] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[17] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.

[18] S. Mandelli, P. Bestagini, and S. Tubaro, "When synthetic traces hide real content: Analysis of stable diffusion image laundering," *arXiv preprint arXiv:2407.10736*, 2024.

[19] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Detecting gan-generated images by orthogonal training of multiple cnns," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3091–3095.

[20] stabilityai, *stable-diffusion-xl-base-1.0*, 2023 (accessed November 25, 2024), https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0.