IMT Nord Europe
École Mines-Télécom
IMT-Université de Lille

Investigating Deepfake Watermarking

M2 Internship for 5 months

Vincent Itier, IMT Nord Europe: `vincent.itier@imt-nord-europe.fr`
Jérémie Boulanger, Univ. Lille: `jeremie.boulanger@univ-lille.fr`

1 Context

Deepfakes are becoming more present and easier to generate and their malicious use may be a vector for false information. In this project, we focus on a specific kind of deepfakes, those which generate fully artificial images. This can be done using Variational Auto Encoder (VAE) [1], Generative Adversarial Networks (GAN) [2] or diffusion model [3]. This abundance of artificial content must be accompanied by evolutions in verification techniques.

We propose to evaluate and to implement active methods *i.e.* methods that actively change the content by embedding a mark in it. The goal of such methods is to be able to assess if an image has been generated by a given generator. For instance, in Stable diffusion [4], authors add the watermark "StableDiffusionV1" just before saving the generated image, however this is just classical image watermarking. Recently, OpenAi has been reported working on "statistically watermaking the outputs of a text"[5].

Message embedding during image generation has been proposed [6]. In this method, the generator (GAN) must produce an image with the additional constraint of inserting the secret message. This constraint is enforced by adding the difference between the extracted message and the inserted one, to the generator loss function. Therefore, minimizing the loss function implies to maximize the quality of the watermark embedding (while enforcing the quality of the reconstructed image).

2 Goals and Challenges

- First, the intern will study diffusion models and implement a diffusion model on a toy exemple. Based on this, he/she will investigate the impact of watermarking the input noise on the generated output and watermark retrieval. The goal of watermarking is to have generated image not be degraded by the watermark while the watermark should be robust to some transformations (compression, rotation, cropping, ...). As the generated images depends upon the input noise, the marks initially inserted should still be present at the image level. The analysis of this classical watermarking trade-off will provide some insight about the quality of the proposed method.

- Another possible challenge is to implement the method of [6] in a VAE. After extensive analysis of the method, we want to extend it by inserting the watermark directly in the latent space of the model. Some work has been conducted in this way [7].

3 Candidate profile

The candidate should have background in machine learning, a strong motivation toward research, as well as coding skills in Python (Pytorch, ...). Knowledge in multimedia security is not mandatory but will be appreciated. You will receive 5 month internship grant.

References

[1] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[3] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *arXiv preprint arXiv:2209.00796*, 2022.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2021.

[5] S. Aaronson. `https://scottaaronson.blog/?p=6823`.

[6] J. Fei, Z. Xia, B. Tondi, and M. Barni, "Supervised GAN watermarking for intellectual property protection," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, IEEE, 2022.

[7] P. Fernandez, A. Sablayrolles, T. Furon, H. Jégou, and M. Douze, "Watermarking images in self-supervised latent spaces," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3054–3058, IEEE, 2022.