Master Thesis / Internship:
Toward Interpretable Adversarial Attacks

Supervisor: Jordan Frecon-Deloire (`https://jordan-frecon.com/`)
Mail: jordan.frecon.deloire@univ-st-etienne.fr
Location: Laboratoire Hubert Curien, Saint-Etienne, France
Team: Data Intelligence
Level: Master 2 / 3rd year of engineering school
Gratuity: ≃ 540 euros/month

Keywords: Deep learning; Adversarial attack; Matrix factorization; Multi-task

Description Adversarial attacks are almost imperceptible transformations aiming to modify an example well classified by a deep neural network (DNN) into a new example, called adversarial, which is itself wrongly classified [1, 2]. Their existence along with the lack of interpretability and explainability of DNNs prevent the massive deployment of DNNs in sensible domains. This subject proposal embraces the original viewpoint of seeking for interpretability in DNNs through the prism of their weaknesses, namely the adversarial attacks. The first part of the internship will be devoted to the adaptation of matrix factorization techniques [3], which made it possible to discover latent structures in natural images, to adversarial attacks. An emphasis will be put on the *a posteriori* study of the patterns present in those attacks in order to interpret how to fool the decision process of DNNs. The second part will investigate a multi-task extension, taking into account both the class to fool and all the other possible classes to imitate.

Expected results

- Bibliographical study on *Adversarial attacks*, *Matrix factorization* and *Multi-task learning*
- Conception of a matrix factorization framework for learning interpretable adversarial attacks
- PyTorch implementation of an efficient solver
- Experimental study on benchmark image datasets and state-of-the-art DNNs
- A publication in a leading journal or conference could be considered depending on the results

Candidate profile The candidate should be proficient in both Python and PyTorch.

References

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing Properties of Neural Networks. arXiv preprint arXiv:1312.6199.

[2] Frecon, J., Gasso, G., Canu, S. (2022). Semi-Universal Adversarial Perturbations. *IEEE TPAMI - TechRxiv Preprint.*

[3] Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1).