MILES
Machine Intelligence and Learning Systems

Adversarial attacks in the light of stability of ResNets

— **Keywords :** Deep learning, Adversarial Attack, Dynamical systems
— **Duration :** 4 to 6 months
— **Supervisors :**
Alexandre Allauzen : alexandre.allauzen@dauphine.fr
Laurent Meunier : laurent.meunier@dauphine.eu
— **Place :** MILES, LAMSADE, Université Paris Dauphine.

Context : Deep learning algorithms have shown their vulnerability to adversarial attacks [3], small and imperceptible perturbations of the inputs that maliciously fools neural networks. Since this discovery, building adversarial attacks and defenses against them have been an active and hot research topic. However, the understanding of this phenomenon remains in the early stages and their is still a gap to come up with to better understand adversarial attacks.

Goals : Recent papers have explored the interaction between machine learning and numerical methods, opening new perspectives for many questions. For instance, the paper [1] proposes an interpretation of deep learning as a parameter estimation problem of nonlinear dynamical systems. This formulation allows us to analyze the stability and well-posedness of deep neural network architectures like ResNet [2].

During this internship, we will study how the notion of stability inherited from dynamical systems can be used in the context of adversarial attacks : can the robustness be improved by introducing some stability properties in the ResNet architectures ? The intuition would suppose that a "stable" architecture would be more robust to adversarial attacks.

Organisation : A literature review of recent work in the field will provide guidelines to build the experimental setup. Starting from a standard image classification task and after training an efficient ResNet-like models, the experiments will attempt to improve the robustness of the model by adding constraints or using structured matrices. These constraints can be introduced in the loss function or as peculiar matrix forms. The experimental framework will use the python library PYTORCH.

Références

[1] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1) :014004, Dec 2017.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.