Category: Artificial Intelligence, Machine Learning, Deep Learning, Security

Job title: Detection of adversarial examples in natural image databases

Contract: Post-doctorat

Contract duration: 12 months

Subject:
Artificial neural networks (ANNs) have become fundamental tools for processing data that is now massively exchanged. Their deployment in sensitive technologies (autonomous cars, smart grids, content filtering, pre-processing of personal data on the peripheral internet, etc.) is accelerating. However, ANNs have a major security default: they can be attacked by adversarial examples. Adversarial examples are data created maliciously to lead ANNs to make erroneous predictions. They pose a critical problem today and are therefore a hot topic in machine learning.

A bio-inspired model has already been built to detect adversarial examples, which implements reinjections similar to those proposed in some computational models of human memory. The hybrid model implements both auto-encoder and classifier functions, gets of excellent performance on small-scale databases. A patent is being filed that will protect the system for the detection of anomalies in a broad sense. A scientific paper is also being written, which presents the model for the specific case of the detection of adversarial examples (which can be considered a particular case of anomalies), a particularly difficult to detec).

The postdoctoral researcher will be required to generalize the model's performance to larger-scale natural image databases. In particular, it will be necessary to:

- Propose a new type of hybrid architecture to move from classic autoencoders to convolutional autoencoders (which would no longer require a pre-extraction of features);

- Test integration of the generative model system (which "stabilizes" the latent space), in particular variational autoencoders;

- Optimize the existing model parameters.

On the other hand, we would like to extend the topic of adverse case detection to the robustness of models against adversarial examples (i. e., the ability of models to properly classify adversarial examples). Preliminary results from smaller databases suggest that the reinjection of an adversarial example might help to identify its class of origin or to denoise it, a phenomenon that has yet to be quantified and generalized to larger databases.

Applicant Profile
The candidate should have completed a PhD in Computer Science, Machine Learning.

Knowledges and experiences in some or all of the following fields will be an asset during the position:

- Adversarial Machine Learning
- Security (attacks, protections, evaluation)
- Applied mathematics (probability / statistics)

Good programming practice in Python (Tensorflow, with some basic GPU environment knowledges). Applicants should master written and spoken English.

A brief description of the PhD thesis, a publication list and some recommendations should be included to your application.

Job location
France, Grenoble

Position start date
31/10/2021

Contact
Marina REYBOZ, **marina.reyboz@cea.fr**, 04.38.78.27.68
Pierre-Alain MOELLIC, **pierre-alain.moellic@cea.fr**, 04.42.61.67.38