Internship Offer:
Adversarial Attacks and Defenses in Machine Learning: Implementation and Comparison

For those who prefer to read in French, a French version of this offer is available below.

Institution:

Télécom Paris, Institut Polytechnique de Paris

Supervisors:

- Jose Antonio Lorencio Abril, PhD Student at Télécom Paris
- Mounira Msahli, Associate Professor at Télécom Paris
- Albert Bifet, Professor at Télécom Paris

Context:

Adversarial examples pose a significant challenge to the robustness and reliability of machine learning systems. Research in this area aims to understand and mitigate these vulnerabilities. Seminal works such as *Intriguing properties of neural networks* [1], *Explaining and Harnessing Adversarial Examples* [2], and *Towards Evaluating the Robustness of Neural Networks* [3] have laid the groundwork for this field.

Essentially, an adversarial example consists of slightly modifying a natural data sample that is correctly classified by a neural network so that the modified version is incorrectly classified but remains almost the same for a human observer. Formally, given a classifier f and an input x with true label y, an adversarial example $x^* = x + r$ can be found by solving:

$$\min_r \|r\| \quad \text{subject to } f(x+r) \neq y.$$

This minimization ensures that the perturbation $\|x^* - x\|$ is minimal while fooling the classifier.

To visualize what happens in input space, see Figure 1, where we show a 2D example of an original data sample being moved to the closest point on the class boundary.

In parallel, defense mechanisms have been developed to improve the robustness of models against adversarial attacks. Techniques such as adversarial training [4], gradient masking [5], and input preprocessing [6] have shown varying levels of success. However, the diversity of methodologies and evaluation strategies complicates the comparison of these defenses.

Objectives:

The objective of this internship is to create a comprehensive and fair benchmarking framework for adversarial attacks and defenses in machine learning. This includes:

- Implementing and evaluating state-of-the-art adversarial attacks and defense mechanisms.

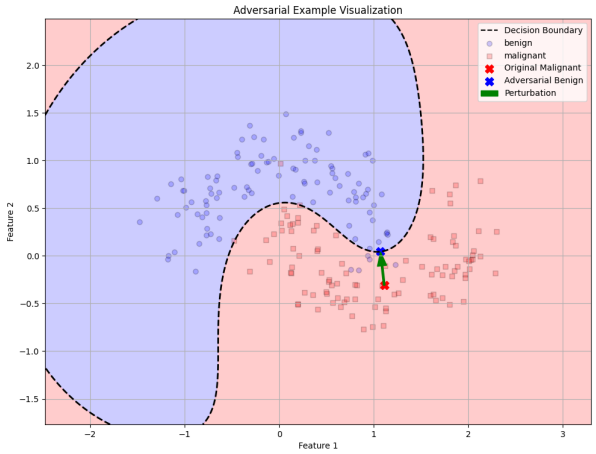

Figure 1: A visualization of adversarial examples in 2D.

- Defining and applying metrics to assess robustness.
- Providing insights and recommendations for future research in adversarial machine learning.

Candidate Profile:

- Enrollment in an M1 or M2 program in Machine Learning, Computer Science, Computer Vision, or Applied Mathematics.
- Strong programming skills in Python.
- Experience or strong interest in deep learning and AI robustness.
- A good level of English (at least B2).

Duration:

4-6 months

How to Apply:

Send your CV, academic transcripts, and a motivation letter to:

- jose.lorencioabril@telecom-paris.fr
- mounira.msahli@telecom-paris.fr
- albert.bifet@telecom-paris.fr

Offre de Stage:
Attaques et défenses adversariales en apprentissage automatique: Mise en Oeuvre et Comparaison

Institution :

Télécom Paris, Institut Polytechnique de Paris

Encadrants :

- Jose Antonio Lorencio Abril, Doctorant à Télécom Paris
- Mounira Msahli, Maître de Conférences à Télécom Paris
- Albert Bifet, Professeur à Télécom Paris

Contexte :

Les exemples adversariaux représentent un défi majeur pour la robustesse et la fiabilité des systèmes d'apprentissage automatique. Ce domaine de recherche cherche à comprendre et atténuer ces vulnérabilités. Des travaux fondateurs comme *Intriguing properties of neural networks* [1], *Explaining and Harnessing Adversarial Examples* [2], et *Towards Evaluating the Robustness of Neural Networks* [3] ont posé les bases de ce champ.

Essentiellement, un exemple adversarial consiste à modifier légèrement un échantillon de données naturel correctement classé par un réseau de neurones afin que la version modifiée soit mal classée, tout en restant presque identique pour un observateur humain. Formellement, pour trouver un $x^* = x + r$ qui trompe le classificateur f:

$$\min_r \|r\| \quad \text{subject to } f(x+r) \neq y.$$

Cette minimisation garantit que la perturbation $\|x^* - x\|$ est minimale tout en trompant le classificateur.

Pour visualiser ce qui se passe dans l'espace des entrées, voir la Figure 1, où un exemple 2D montre un échantillon de données original déplacé vers le point le plus proche de la frontière de classe.

En parallèle, des mécanismes de défense ont été développés pour améliorer la robustesse des modèles face aux attaques adversariales. Des techniques telles que l'entraînement adversarial [4], le masquage de gradients [5], et le prétraitement des entrées [6] ont montré des niveaux de succès variables. Cependant, la diversité des méthodologies et des stratégies d'évaluation complique la comparaison de ces défenses.

Objectifs :

L'objectif de ce stage est de développer un cadre d'évaluation complet et équitable pour les attaques et défenses adversariales en apprentissage automatique, incluant :

- La mise en œuvre et l'évaluation des attaques et mécanismes de défense à la pointe de la recherche.

- La définition et l'application de métriques pour mesurer la robustesse.
- La formulation de recommandations pour orienter les recherches futures.

Compétences Requises:

- Inscription en M1 ou M2 en apprentissage automatique, informatique, vision par ordinateur ou mathématiques appliquées.
- Compétences solides en programmation Python.
- Expérience ou fort intérêt pour l'apprentissage profond et la robustesse de l'IA.
- Un bon niveau d'anglais (au moins B2).

Durée :

4-6 mois

Comment postuler :

Envoyez votre CV, relevés de notes académiques et une lettre de motivation à :

- jose.lorencioabril@telecom-paris.fr
- mounira.msahli@telecom-paris.fr
- albert.bifet@telecom-paris.fr

References

[1] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

[3] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*.

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.

[5] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*.

[6] Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*.