Evaluation d'une attaque adverse sur des réseaux de neurones profonds et développement de contre-attaques

Solène Bernard, Patrick Bas, John Klein

Les CNN sont beaucoup utilisés dans des tâches de classification d'images. Seulement, comme montré dans [3], nous pouvons facilement les tromper en générant des *adversarial examples*, ou exemples antagonistes, qui cherchent à tromper un classifieur en modifiant une image de manière imperceptible . Nous pouvons trouver par exemple dans la référence [4] un état de l'art des méthodes de défenses et d'attaque des exemples antagonistes.

Motivé par le contexte de la stéganographie [1], un protocole itératif a été proposé par Bernard *et al.* [1, 2] pour générer des contenus stéganographiques antagonistes de plus en plus sécurisés au fur et à mesure des itérations.

L'idée de ce projet de recherche est d'appliquer ce protocole à des tâches de classification comme la reconnaissance de chiffres. Par exemple, l'une des méthodes de la littérature destinée à rendre la classification plus sûre face aux exemples antagonistes consiste à ajouter des exemples antagonistes à la base d'entraînement. L'objectif de ce projet sera d'évaluer si l'entraînement sur une base antagoniste générée par notre protocole permettrait d'améliorer la sécurité d'un classifieur, et ce en le comparant à d'autres contre-attaques de la littérature.

Ce projet nécessitera dans un premier temps un travail d'étude bibliographique sur les méthodes de défenses et d'attaque par exemples antagonistes.

Le travail d'implémentation sera développé dans l'environnement Tensorflow en attaquant un classifieur léger travaillant sur la base MNIST. Des tests sur des bases d'images plus importantes pourront être envisagés via l'utilisation d'un GPU.

Contacts :
Patrick.Bas@centralelille.fr (Chercheur à CRIStAL, Bâtiment ESPRIT)
Solene.Bernard@centrale.centralelille.fr (Doctorante à CRIStAL, Bâtiment ESPRIT)

Références

[1] Solène Bernard, Tomáš Pevný, Patrick Bas, and John Klein. Exploiting Adversarial Embeddings for Better Steganography. In *IH-MMSec*, IH&MMSec'19 Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Paris, France, July 2019.

[2] Solène Bernard, Tomas Pevny, Patrick Bas, and John Klein. Utilisation d'insertions adverses pour améliorer la stéganographie. In *GRETSI*, Lille, France, August 2019.

[3] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018.

[4] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples : Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019.

1. La stéganographie consiste en la modification imperceptible d'un contenu afin d'insérer un message secret dans celui-ci.