Perturbations adversariales de réseaux de neurones

2024 - 2025
Michaël Clément - Rémi Giraud
michael.clement@enseirb-matmeca.fr - remi.giraud@enseirb-matmeca.fr

Mots-clés : Apprentissage profond; Perturbations adversariales; Détection d'objets; Vidéo

Contexte : En *deep learning*, les perturbations adversariales désignent des modifications intentionnelles appliquées aux données d'entrée, souvent imperceptibles à l'œil humain, mais capables d'induire en erreur les modèles d'apprentissage profond. Ces attaques adversariales exploitent les vulnérabilités des réseaux de neurones, les amenant à produire des prédictions incorrectes ou incohérentes. Ce phénomène met en lumière les faiblesses des modèles d'IA face à de telles attaques, soulignant ainsi la nécessité de renforcer la robustesse et la sécurité des systèmes utilisant ces modèles. Les modifications peuvent être de différentes natures, souvent imperceptibles et ajoutées numériquement à l'image de manière globale (voir Figure 1). Les perturbations peuvent aussi être liées à un motif ou *pattern* réel présent dans l'image capturée afin de tromper les capacités de détection d'objets ou de classification (voir Figures 2 et 3).

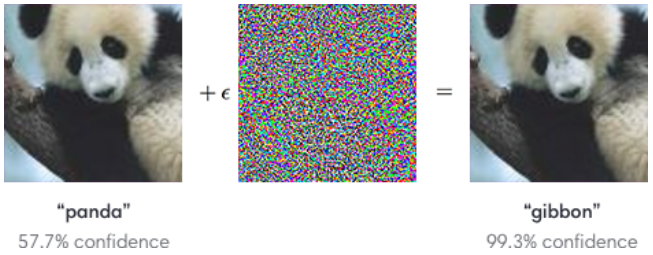

Figure 1 : Perturbation de classification par bruit additif [1]

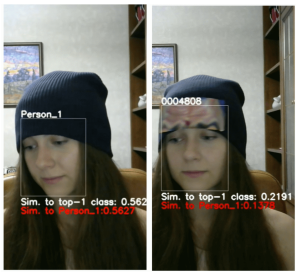

Figure 2 : Perturbation de détection par motifs réels [3]

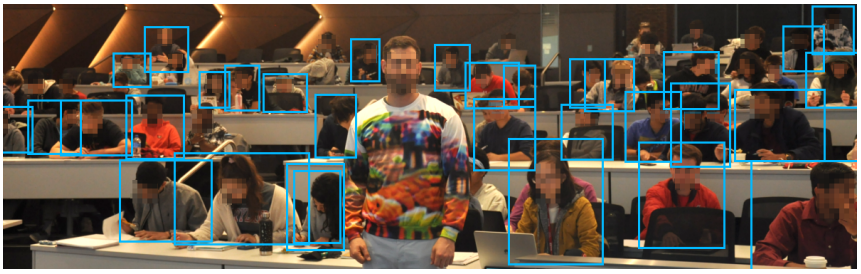

Figure 3 : Pull rendant "invisible" aux réseaux de détection de personnes [4].

Objectifs : L'objectif de ce projet est d'étudier les méthodes de génération de perturbations adversariales permettant de trouver des motifs trompant les architectures de réseaux de neurones actuelles.

Étapes du projet :

- État de l'art des méthodes de résolution de perturbations adversariales.
- Mise en oeuvre de méthodes permettant de générer :
  - Un bruit imperceptible pouvant perturber des modèles de classification d'images.
  - Des motifs réels pouvant perturber soit la détection, soit la classification de réseaux de reconnaissance d'objets en vidéo temps-réel (ex. YOLO [2]).
- Selon l'avancement du projet, les méthodes pourront être testées en conditions réelles avec impression des motifs et test depuis des captures via téléphone, ou caméra sur Raspberry Pi.

Références :

[1] Ian J Goodfellow. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.

[3] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th international conference on pattern recognition (ICPR)*, pages 819–826. IEEE, 2021.

[4] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 1–17. Springer, 2020.