De nombreux travaux scientifiques traitent des vulnérabilités informatiques spécifiques aux réseaux de neurones, aussi bien au niveau attaque que défense. Parmi ces vulnérabilités, le leurrage est une des plus connues et des plus étudiées. Le leurrage consiste en l’introduction, par un attaquant, d’une perturbation dans la donnée fournie en entrée du réseau de neurones, afin de modifier la réponse du système.

Lorsque la perturbation a pour objectif d’être discrète, c’est-à-dire de ne pas modifier de façon importante la donnée d’entrée, on parle d’attaque de leurrage par exemple contradictoire (ou *adversarial example* en anglais), et donc de leurrage contradictoire: la donnée modifiée est en effet un exemple contradictoire, c’est-à-dire un échantillon proche de l’échantillon initial mais déclenchant une réponse radicalement différente (et fausse). Selon le type de données traitées, il peut s’agir de la modification de quelques pixels soit répartis dans l’ensemble de l’image soit dans une zone précise de l’image (on parle alors d’attaque par patch), d’une superposition de signal, d’ajout d’un léger bruit sur l’ensemble de la donnée, etc.

L’introduction de ces perturbations se fait le plus souvent de manière informatique. Dans certaines configurations toutefois, la perturbation peut aussi être déclenchée par une action physique, dans le monde réel, ce qui réduit fortement les possibilités de défense et de détection d’attaque : c’est par exemple le cas du leurrage par patch dans les images, illustré dans les images suivantes :



Illustrations de leurrage par patch. A gauche, un patch spécifique permet de fausser les décisions prises par un système de reconnaissance d’images, et ce quelle que soit l’image dans laquelle le patch est placé [1]. A droite, un patch porté par un l’individu de droite permet de leurrer un système de détection de personne [2].

On parle de leurrage en mode boîte blanche lorsque l’attaquant dispose d’une information totale sur le réseau de neurones, incluant son architecture et l’ensemble de ses paramètres. Quand l’attaquant n’a accès ni à l’architecture du réseau de neurones ni à ses paramètres, mais dispose d’une information partielle sur celui-ci (par exemple la réponse du système aux entrées présentées ou le type de modèle utilisé), on parle alors d’attaque par leurrage en mode boîte grise.

Le stage se propose de travailler sur une attaque de leurrage contradictoire en mode boite blanche, par patch, d’un système de détection de visage dans des images. L’objectif final est de transférer en physique les patchs de leurrage produits initialement dans le monde numérique, et donc de démontrer la faisabilité du leurrage du modèle de détection dans le monde réel. Le réseau de neurones visé par les attaques pourra soit être un modèle de reconnaissance de contenus dans des images (par exemple de détection de visages) produit en interne, soit un réseau de neurones disponible en open-source.

[1] T. B. Brown, D. Mané, A. Roy, M. Abadi, J. Gilmer (2017). *Adversarial patch.* <https://arxiv.org/pdf/1712.09665.pdf>

[2] S. Thys, W. Van Ranst, T. Goedemé (2019). *Fooling automated surveillance cameras: adversarial patches to attack person detection.* <https://arxiv.org/abs/1904.08653>