# Développer l'explicabilité des réseaux de neurones convolutifs par application de masques générés par évolution artificielle

Anne Jeannin-Girardon, anne.jeannin@unistra.fr / Équipe CSTB
Co-encadrement : Romain Orhand

**Contexte**

Les processus de prise de décision par des algorithmes d'apprentissage automatique profonds posent aujourd'hui des défis important en matière de compréhension de ces processus, et en particulier concernant l'explication des décisions et prédictions de ces modèles. On fait là face à des problématiques critiques impliquant des questions d'éthique ou encore de droit. Au niveau Européen, le Règlement Général sur la Protection des Données (RGPD) stipule par exemple que, lors du traitement automatique de données personnelles, "La personne concernée devrait avoir le droit [...] d'obtenir une explication quant à la décision prise"[1]. L'application de ce droit est toutefois problématique compte tenu de la complexité des modèles de traitement de données pouvant être mis en œuvre. À plus large échelle, il s'agit aussi de permettre une meilleure compréhension des modèles d'apprentissage automatique, de façon à pouvoir (a) les améliorer (augmenter leur précision, corriger des erreurs de prédiction), (b) les contrôler (contrôler leur comportement, mieux évaluer leurs limites ou vulnérabilités) et (c) extraire de nouveaux faits potentiellement latents dans le modèle.

À l'heure actuelle, "l'explicabilité" peut se mettre en œuvre à trois niveaux distincts:

1. Comme module externe, par exemple avec un outil comme LIME [1]
2. En ajustant les sorties d'un modèle, avec par exemple dans le cas du traitement d'images en transformant celle-ci de façon à produire des cartes de saillance [2]
3. Par conception : des modèles moins complexes, comme des arbres de décisions, sont par conception plus explicable qu'un modèle type réseau de neurones profond.

Une revue à jour des problématiques, méthodologies et applications liées à "l'intelligence artificielle explicable" peut être consulté dans [3].

**Travail demandé**

Après une étude bibliographique permettant de mieux cerner les enjeux et les méthodes actuelles pour développer l'explicabilité des modèles d'apprentissage automatique profond, nous proposons de mettre en œuvre un module externe s'appliquant au domaine de la reconnaissance d'objets dans des images, et reposant sur l'utilisation d'algorithmes évolutionnaires. Le but est de faire évoluer une population de masques qui seront appliqués sur les images de façon à mettre en évidence les portions de l'image ayant donné lieu à une prédiction (similaire, dans le résultat, au fonctionnement de LIME). Selon les résultats obtenus, on pourra réfléchir à une généralisation de la méthode à d'autres types de données.

[1] https://eur-lex.europa.eu/eli/reg/2016/679/oj Article 71

**Références**

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining016.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller, Eds., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer International Publishing, 2019.