Projet de TER : Interprétation des réseaux de neurones profonds

Description du projet

Au cours des dernières années, les réseaux de neurones profonds (Deep Learning) ont fait avancer l'état de l'art dans de nombreux domaines comme par exemple, la détection d'objets dans les image et la reconnaissance vocale. Malgré ce succès, les réseaux de neurones sont considérés comme des boîtes noires. Leur fonctionnement interne reste obscure et on ne peut pas savoir s'ils basent leurs prédictions sur les bons critères ou sur des artefacts contenus dans les données.
Dans la tentative de comprendre le fonctionnement des réseaux de neurones, plusieurs méthodes ont été proposées dans la litérature [2] [3] [4] comme : Saliency, Deconvnet, Guided Backpropagation, Intergrated Gradients, Layerwise Relevent Propagation, PatternNet, etc. Les auteurs de [1] proposent une bibliothèque nomée "iNNvestigate" qui analyse le fonctionnement des réseaux de neurones en fournissant une interface commune et une implémentation pour de nombreuses méthodes d'analyse. Cette bibliothèque est particulièrement adaptées aux entrées sous forme d'images.
L'objectif du projet est d'adapter cette bibliothèque aux données vectorielles et particulièrement aux données d'expression de gènes. L'évaluation des résultats fournis par les différentes approches sera également étudiée.

Contact

Farida Zehraoui
email : farida.zehraoui@univ-evry.fr

Références

[1] Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks! *CoRR*, abs/1808.04260, 2018.

[2] Marco Ancona, Enea Ceolini, Cengiz Oztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*, 2018.

[3] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani B. Srivastava, Alun D. Preece, Simon J. Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. Interpretability of deep learning models : A survey of results. *2017 IEEE SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*, pages 1–6, 2017.

[4] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2017.