heig-vd

HASLERSTIFTUNG

CI4CB
Computational Intelligence for Computational Biology

HEALTH ENGINEERING ECONOMICS

Filière TIC

D-REX: Extraction de règles pour des réseaux de neurones profonds

Développer une méthode qui permette de comprendre, hiérarchiser et expliquer la connaissance acquise par un réseau de neurones profond afin de mieux comprendre son fonctionnement.

Projet de recherche soutenu par la Fondation Hasler

CONTEXTE

> Les réseaux de neurones artificiels, et en particulier les réseaux profonds (DNNs) sont des méthodes d'apprentissage automatique extrêmement efficientes. Ces réseaux ont récemment gagné en popularité, notamment dans le domaine de la classification d'image, surpassant les performances humaines.

Cependant, le principal désavantage de ces réseaux réside dans leur manque de transparence. Ainsi, on en sait très peu quant aux données et relations causales qui mènent aux prédictions observées.

Ce manque de transparence est notamment très problématique pour l'application de méthodes basées sur les DNNs dans des domaines qui nécessitent une validation des décisions, par exemple dans le diagnostique médical.

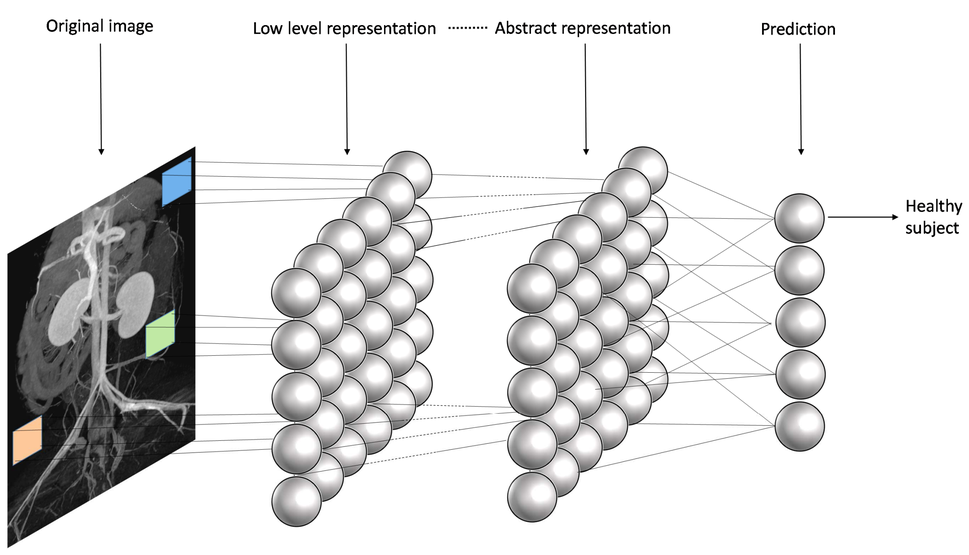
BUT

> Le but de ce projet est par conséquent de développer une méthode qui permette de comprendre comment l'information est assimilée, traitée puis utilisée par les DNNs. Cela permettra d'une part d'extraire une partie de la connaissance acquise par ces réseaux durant la phase d'apprentissage et, d'autre part, de fournir une explication justifiant chaque prédiction effectuée par le réseau.

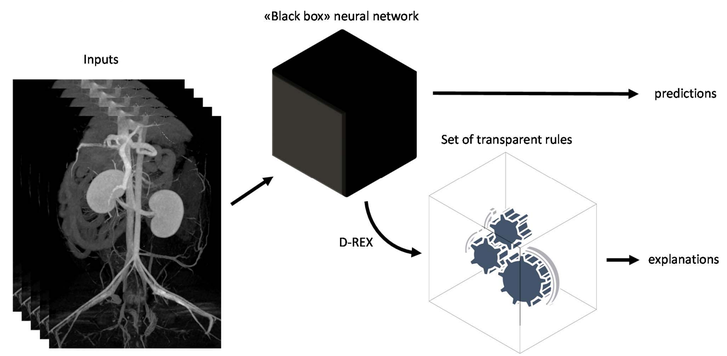
APPLICATIONS

> Une telle méthode a de nombreuses applications. Dans l'imagerie médicale par exemple, elle permettrait d'extraire de la connaissance de très grande quantité de données, permettant ainsi de découvrir potentiellement de nouvelles méthodes de dépistage.

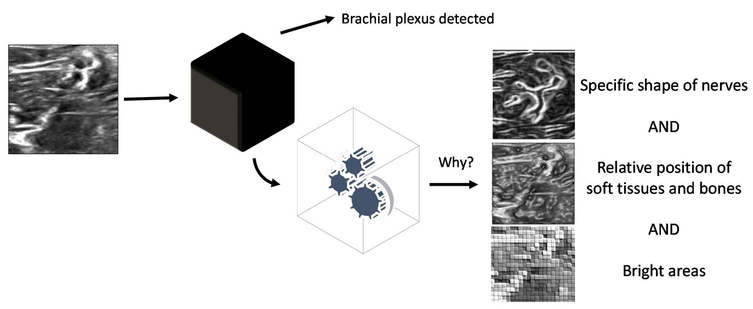
Représentation d'un DNN traitant des images - L'image d'entrée est analysée par plusieurs couches successives de neurones qui réagissent à certaines motifs spécifiques. Plus une couche est située profondément, plus elle sera sensible a des caractéristiques complexes



Concept de D-REX - le fonctionnement interne du réseau de neurones est expliqué par un ensemble de règles transparentes



Exemple d'un DNN détectant la présence du plexus brachial dans une image ultrason - Une règle permet d'expliquer la réponse du réseau



Contacts: Prof. Carlos Andrés Peña, IICT, TIC, carlos.pena@heig-vd.ch
Financement: Fondation Hasler

Hes·so
Haute Ecole Spécialisée de Suisse occidentale