Master Internship 2023

Toward efficient then explainable models: Choquetization of a Neural Net

Supervisors: Michèle Sebag, Johanne Cohen, Christophe Labreuche, Roman Bresson

Lab: TAU − CNRS − INRIA − LISN, Univ. Paris-Saclay
Lab: Thales Research Technology, Palaiseau

Context and Motivation:
Learning explainable models is mandatory in critical and sensitive domains, where the human operator must stay in control of the eventual decision and/or assess its compliance w.r.t. ethical regulations. Explainability is also most desirable in order to debug a learned model: understanding the model' mistakes enables one to e.g., augment the training data as appropriate to correct the misclassifications.

Basically, the main approaches developed in the literature to learn explainable models are based on: i) explaining efficiently trained models, e.g. by finding a linear approximation of the trained model in the neighborhood of the case/decision to explain [1]; ii) learning models that are explainable by design, e.g. Choquet integrals, developed in the Multi-Criteria Decision Aid community to formalize and aggregate the expert decision criteria [2,3].
Both approaches incur different and complementary limitations:
The first approach yields a case-based explanation, formalized through the weights of the input coordinates. In particular, in the case where the input representation is of low granularity, e.g. for computer vision, ad hoc approaches are needed to extract the salient region in the image, responsible for the class estimation [4].
The second approach delivers understandable models by design, under the assumption that the decision elements, at the proper level of granularity (e.g., not pixels) are available as input representation. Further, the simplicity of the decision architecture may entail a loss of performance of the trained model compared to a standard neural net.

Goal of the internship:
A third approach will be investigated in the internship, aimed to getting the best of both worlds. Formally:

- Phase 1: A trained black box model (neural net X input $\rightarrow$ $Z = f(X)$ hidden (last) layer $\rightarrow$ output $\hat{Y} = g(Z)$) will be considered in the first step. The nodes Z, referred to as latent concepts, expectedly contain all information in order to predict when linearly combined, but they are not understandable;

- Phase 2: The Choquetization of this neural net proceeds by: i) initializing $Z = f(X)$; ii) searching a hierarchical Choquet integral architecture (HCIA) g' such that $g'(Z)$ fits Y.
  Formally, the HCIA g' is recursively defined from two operators: the scoring operator maps a real-valued latent Z_i onto a simple score (monotone or peak or valley-shaped); along a tree-structure, the aggregation operator associates each parent node with a score aggregating the children node scores.
  The HCIA g' would thus support the explanation of the final decision *if* the latent concepts Z were intelligible.

The second phase learns g' and possibly learns f' (revising f if the back-propagation operates on the whole neural net). Let Z' denote the revised latent concepts ($Z' = f'(X)$).

- Phase 3: This phase aims to understand (give a name to) the Z'_is. In the context of the internship, the *name giver* is the expert in the loop and the proposed methodology can proceed as follows. To each Z'_i is associated a binary classification problem (samples with high/low score on Z_i). Representative samples of the two classes are presented to the expert, along an interactive elicitation process [4].
  The interaction will allow to refine and validate these names, that will be reinjected in the process, e.g. augmenting the representation of the data.

The internship requires excellent creativity, as well as theoretical and algorithmic skills (programming environment in Python).

Bibliography

[1] *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. KDD 2016.
[2] *Neural Representation and Learning of Hierarchical 2-additive Choquet Integrals*. Roman Bresson, Johanne Cohen, Eyke Hüllermeier, Christophe Labreuche, Michèle Sebag. IJCAI 2020
[3] *On the Identifiability of Hierarchical Decision Models*, Roman Bresson, Johanne Cohen, Eyke Hüllermeier, Christophe Labreuche, Michèle Sebag. KR 2021.
[4] *Learning Saliency Maps for Object Categorization*, Frank Moosmann, Diane Larlus, Frederic Jurie. MLJ 2006.
[4] *Interactive Intent Modeling for Exploratory Search*, Tuukka Ruotsalo, Jaakko Peltonen, Manuel J. A. Eugster, Dorota Glowacka, Patrik Floréen, Petri Myllymäki, Giulio Jacucci, Samuel Kaski. ACM Trans. Inf. Syst. 36(4): 44:1-44:46 2018.