Explaining black-box deep neural models' predictions, behaviors, and performances through the unsupervised mining of their inner knowledge

Candidate: Francesco Ventura
Supervisor: Prof. Tania Cerquitelli

Politecnico Di Torino

Abstract

Artificial Intelligence applications are experiencing a disruptive expansion in many relevant aspects of our life thanks to the development of even better performing Deep Neural Networks (DNNs). However, along with higher performance, AI models are characterized by high complexity and opaqueness, i.e., they do not allow understanding the reasoning behind their automatic decision-making process. This widely limits their applicability and opens to a wide range of problems in many sensible contexts, e.g. health, transportation, security, and law. From one side, it is very hard to interpret the decision-making process of AI models both at the local and global levels. Furthermore, it is even harder to assess the reliability of their predictions over-time, e.g. because of the presence of concept-drift. Ignoring just one of these aspects may have very harsh consequences in real-life settings where it is supposed that users, both expert, and non-expert, should trust the decisions taken by "smart" platforms and devices. The evident complexity and relevance of these issues have led to the birth of a new branch of research, namely Explainable Artificial Intelligence (xAI).

In the literature, these challenges are faced separately and often without taking into account the latent knowledge contained in the deeper layers of the DNNs. Instead, we claim that these issues are part of the same big challenge: the Model Reliability Management.

For these reasons, we aim to address these challenges by proposing (i) a unified model-aware strategy for the explanation of deep neural networks at prediction-local and model-global level, and (ii) a unified model-aware assessment framework for the management of models' performance degradation over-time.

First, the prediction-local and model-global explainability has been addressed by introducing a new explanation framework tailored to Deep Neural Networks. We introduce new unsupervised mining strategies to extract and analyze the inner knowledge contained in multiple deep layers of different models employed in different contexts, i.e. image and textual Machine Learning tasks. The explanations produced by the proposed framework consist of innovative quantitative indicators and ad-hoc qualitative visual/textual information. These explanations are developed to be easily understandable by humans enabling both expert and non-expert users to better understand and to dig into the black-box decision-making process. Also, the proposed explanation framework has been developed to adapt to as many alternative models as possible going from Convolutional Networks to complex Natural Language Models like BERT. The proposed local and global explanations have been validated on diverse image and textual classification tasks and exploiting very different state-of-the-art deep architectures, i.e. VGG16, VGG19, InceptionV3, InceptionResNetV2, BERT, and LSTM. We show with these experiments that the proposed framework can be easily adapted to several use-cases and that we are able to effectively identify misleading patterns in the prediction process. Also, we provide a complete comparison of our methodology w.r.t. the current state-of-the-art and we quantitatively validate our solution collecting feedbacks from human users through an online survey. Our explanations have been considered more interpretable concerning the compared methods in 75% of the cases.

Then, assessing and managing the quality of the model's outcomes over-time requires dealing with the management of concept drift. To address this challenge we introduce a new Concept Drift Management framework that allows to continuously monitor the presence of drifting distributions of data that may affect the performance of the predictive models deployed in production systems. With this framework we aim to introduce a new level of model interpretability, addressing the current limitations of explanation frameworks that do not take into account over-time performance degradation. The framework is based on an unsupervised and scalable strategy that measures the drift of

newly incoming data concerning the knowledge of the model at the training time. As it is unsupervised it does not require any prior knowledge about the ground-truth of newly collected data. Also, we introduced a new definition of model degradation that quantifies the per-class amount of performance reduction. Furthermore, the proposed methodology is very general and it can be applied to several different use-cases and predictive models: The proposed framework has been tested on very diverse deep models, e.g. Doc2Vec, VGG16, and BERT, tailored to unstructured data domains, i.e. images and texts. We show that our methodology is able to detect the presence of drift already when just 10% of data is drifting in the window of analysis. Finally, we demonstrate that our methodology can be horizontally distributed and it can linearly scale managing millions of input samples in the order of few minutes.