Abstract

Recent years have brought a significant advancement in artificial intelligence methods allowing for the automation of multiple repetitive tasks in many domains, including image recognition and natural language processing. This resulted in performance comparable or superior to humans. These achievements were largely possible owing to recent breakthroughs in computational devices, especially graphical processing units (GPUs), and machine learning techniques utilizing Deep Neural Networks (DNNs). However, despite their remarkable performance, DNNs involve multiple flaws. One major drawback is their black-box nature caused by the lack of explanations for their decisions. The inability to inspect the model's reasoning process is a concerning aspect, particularly in high-stakes decision fields such as medicine, where those decisions may significantly impact human lives.

In response to the lack of transparency of DNNs, various approaches have been developed to provide explanations for their decisions. These approaches may be divided into two categories: post-hoc and self-explainable methods. Post-hoc methods involve training a black-box model and subsequently developing an explainer model on top of it. This approach can be applied to already-trained models, but resulted explanations are frequently unreliable and imprecise. Nonetheless, self-explainable approaches incorporate interpretability as an intrinsic aspect of the model. This way they provide explanations alongside with their predictions. While self-explainable models offer higher quality explanations compared to explainers, they are more challenging to train, and their performance is moderately lower than the black-box models.

This doctoral thesis focuses on developing self-explanaible methods utilizing artificial neural networks, with particular focus on attention pooling mechanism and prototypical-parts methodology. The former is employed in weakly supervised learning, particularly in multi-instance learning, where a set of instances is assigned to a single label. The latter, prototypical-part models learn concepts from the training data during the learning process and compare those concepts to an input sample in the inference phase to obtain the prediction.

As part of this doctoral thesis, three new models were developed to address the limitations of the initial prototypical parts-based model ProtoPNet [11]. The first model, ProtoPShare [I] introduces prototype parts sharing across classes by combining those already trained. For this purpose a new metric was defined detecting semantic similarity between prototypical parts, even when they are far in the latent space. The next model, ProtoPool [II], is more advanced and shares the prototypical parts from scratch. It is possible thanks to regularization techniques based on the Gumbel-Softmax Trick. Moreover, ProtoPool introduces focal similarity, which enables more descriptive prototypical parts than those obtained from baseline methods. Finally, ProGReST [III], generalizes prototypical parts to regression problems of predicting molecular properties.

In addition, two methods have been introduced based on the attention pooling methodology. The first one, SA-AbMILP model [IV] uses a self-attention mechanism to learn dependencies between instances in a bag resulting in better performance in non-standard multiple instance learning assumptions, such as presence-based and threshold-based. This enables to understand which visual features influence the model's decision. The second model, ProtoMIL [V], aims to introduce global level explanation into the MIL classification problem by combining prototypical parts and attention pooling.

To summarize, this doctoral thesis focuses on interpretable deep learning based on prototypical parts-based models and attention mechanism. Five papers were published based on those research, two at A* conferences [I, II], and three at A conferences [III, IV, V] according to CORE ranking (in all of them the Ph.D. candidate is the first author). Additionally, the Ph.D. candidate was a principal investigator of the NCN Preludium grant and the grant from the Jagiellonian University Excellence Initiative. He also completed a research internship at the Computer Vision Center of the Autonomous University of Barcelona in the research group of Professor Joost van de Weijer and is a co-author of a patent application submitted to the European Patent Office.

Keywords: interpretability, explainable artificial intelligence, deep learning.