Reg. No: 2025/57/N/ST6/04389; Principal Investigator: lic. Piotr Borycki

Explanation of Pre-Trained deep learning models via Prototypes

With the increasing use of artificial intelligence methods in practical applications, the need to understand how each decision is made is also growing. Many shallow models, such as decision trees, are designed from the beggining with interpretability in mind. However, this is not the case for many complex models, such as deep neural networks. As a result, in recent years, so-called explainable artificial intelligence (XAI) methods have been gaining importance, aiming to provide insight into the decisions made by such systems.

The existing XAI methods are split into two main approaches. The first are post-hoc methods, which attempt to explain the decisions of an already trained system that may be used in practice. Methods in this category typically rely on analyzing traces left by the decision—for example, in the case of images, identifying which parts of the image had the greatest influence on the decision made. One of the major problems with this type of method is that they often provide results that are too general or difficult to interpret.

The second approach includes ante-hoc methods, which are artificial intelligence algorithms that are designed to be understandable from the very beginning. One of the most common types of methods in this category are those based on the concept of prototypes. Prototypes are examples that the AI algorithm has previously seen, and the justification for a decision based on them consists of reasoning such "I chose this option because it is similar to a previously seen case." This is much closer to the reasoning done by humans. In this case, however, the main issue is that creating such systems is difficult, costly, and often cannot directly be applied to already existing algorithms.

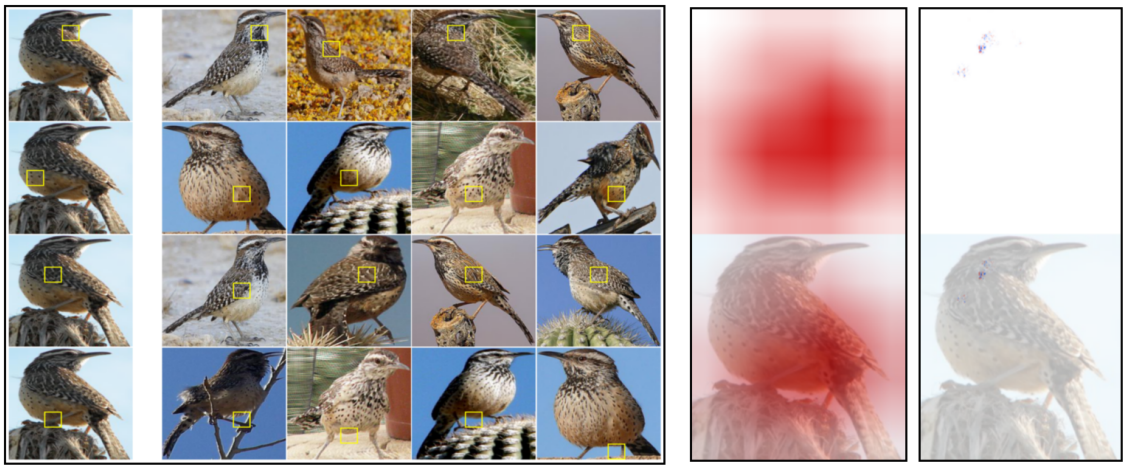

Figure 1: On the left the explanations produced by prototype inspired method can be seen. On the right explanations created by existing post-hoc methods.

The goal of the proposed research project is to combine both approaches, which means creating novel methods, compatible with currently employed AI systems, while offering more intuitive and understandable explanations, similar to those produced by prototype based methods.