Interpretable and sustainable artificial intelligence with intuitive explanations

In our daily lives, more and more activities, such as shopping or car navigation, are being assisted by Artificial Intelligence (AI) algorithms. Moreover, even in areas such as medicine, where the fate of human life is decided, AI methods are utilized. This trend shows that digital solutions will support more and more aspects of life and the economy.

The widespread use of AI solutions is associated with an increased demand for electric energy required both by the AI training process and exploitation in a production environment, as well as a lack of confidence in the decisions they make. Therefore, this project addresses the following two research questions. How to design AI methods to fully utilize the knowledge already obtained, thus saving resources in training the model? And how to present their decisions to users, such as physicians, so they trust AI results? Those questions are especially important in the case of deep neural networks, which obtain high accuracy in numerous applications, but are considered black-box models.

Accordingly, within the framework of this grant, we plan to design and develop AI models that use deep neural networks but simultaneously can return interpretations of their predictions and present the knowledge they have acquired. To this end, we will consider various evaluation techniques of model interpretability to analyze how comprehensible or confusing these interpretations are to the end user. To reduce the costs of user studies, we will work on introducing numerical metrics that evaluate the properties of interpretable models without involving human testers.

In the next step, we plan to work on the precise visualizations of the explanations to the user using cognitive science theories, visualization methods, and models that generate an image description. In addition, we will work on enforcing the model to perceive similar semantics as humans do. For example, in a bird species recognition system, the model should consider the characteristics of sparrows as more similar to each other than those of albatrosses.

On the matter of sustainable use of computing resources, and therefore electric energy, methods will be developed to prevent catastrophic forgetting so that gaining new knowledge does not result in forgetting knowledge possessed so far. Similarly, the possibility of learning a model without having huge labeled data sets, which can be costly, especially in biomedical data, is a desirable feature of AI methods. Therefore, we plan to develop continuous learning methods for deep interpretable models, as well as use contrastive learning to identify relevant features of a dataset, which can then be presented to the user to explain its prediction.

The final step of the project is to apply the developed techniques to problems that are not commonly used in evaluating AI methods, such as life sciences. It will allow us to show that applying interpretable AI methods in other fields is as effective as for natural images. In particular, an important step is to apply these methods to medical image analysis and exhaustively test the returned interpretations to determine what type of information the model extracts and whether they can increase the confidence of specific users such as medical doctors.

In summary, the project aims to carry out research on artificial intelligence that explains its decisions and enables the ecological and sustainable development of this discipline.