Improving interpretability properties of prototypical parts-based deep neural networks.

In last years a vast development in deep learning methodologies and their applications to many domains can be observed. Deep neural networks revolutionized fields like computer vision or natural language processing obtaining results on par or better than humans.

Even if a model obtains good performance, the lack of prediction interpretability is often pointed out as its biggest drawback, like in Google's system for detecting lung cancer on CT scans. Moreover, the current EU law enforces the recognition systems to be self-explanatory. Currently, research on explaining the results of neural networks is focused on two ways:

- explainable artificial intelligence (XAI), which build another model to explain the results of deep learning model,
- interpretable deep learning models, which has built-in interpretability in its architecture.

The goal of the project is to generalize the prototypical-part-based model to share the prototypical parts between data classes from the very beginning of the training to reduce the complexity of the explanation as well as present the spatial arrangement between prototypes as a part of the prediction interpretation.

We will be using publicly available datasets like LIDC-IDRi or Stanford Cars as well as data collected and published as a result of our collaboration with the group of prof. Monika Brzychczy-Włoch, e.g. DIFaS dataset. Based on those datasets the model will be trained and the analysis of interpretability will be conducted. Moreover, we will investigate if this type of rationale behind the prediction can be used to find the differences between classes.

Our approach can potentially be integrated into software systems like computer-assisted diagnosis, computer decision support, aerial image analysis, natural language processing, or autonomous driving.