Detecting and overcoming bias in data with explainable artificial intelligence

Deep learning is now the fastest-growing field of artificial intelligence (AI). The rapid development of deep neural networks is considered by some experts as the AI revolution, while artificial intelligence itself is increasingly referred to as "new electricity". The most popular models used in deep learning in computer vision are convolutional neural networks, which usually have complex architectures with many layers, and are described by millions of parameters. These parameters are automatically adjusted during the learning process. Artificial neural networks learn from input data (what will be analyzed by the system, e.g. a picture of a skin lesion) and output data (the result of the desired expertise, e.g. a statement that this is skin cancer).

Currently, in contrast to classical shallow models exploited in the past, most deep learning systems extract features automatically, and to do that, they tend to rely on a huge number of labeled data. Whereas the quality of dataset used to train neural networks has a huge impact on the model's performance, those datasets are often noisy, biased and sometimes even contain incorrectly labeled samples.

One of the often discussed problems in deep learning is the presence of bias in the data – bias means that data is not representative of the population or phenomenon of study, hence analyzing it may cause logical errors. Such a problem arose in the prediction of possible complications after pneumonia in 1990 in Pittsburgh, where the intelligent system concluded that asthma patients gained additional immunity thanks to the disease. This typical example of misconception happened because asthmatics with high risk of complications were observed by doctors in hospital, so they rarely had any complications. Bias in the data can also take other forms, for example, in one of the artificial intelligence systems, the neural network learned to recognize horse races only by the signatures in all the pictures, while completely ignoring horses and raiders.

And yet, still, fragile black-box deep neural network models are used to solve very sensitive and critical tasks. Therefore, the demand for clear reasoning and correct decision is very high, especially when deep black box systems are used in transportation (autonomous cars), in healthcare, for legal systems, finances, and military.

To address those challenges the project aims to develop methods of **Explainable Artificial Intelligence (XAI)** which might help to uncover and reduce the problem of bias in data. The project involves investigation and integration of explainability into new and existing Artificial Intelligence systems and mostly focuses on **Deep Neural Networks** in the field of Computer Vision. One of the ways of categorizing XAI methods is to divide them into **local** and **global explanations**. Local analysis aims to explain a single prediction of a model, whereas a global one tries to explain how the whole model works in general. The project aims to develop novel methods of both local and global explainability to help explain deep neural network-based systems in order to **justify** them, to **control** their reasoning process, and to **discover** new knowledge.

At first, it is planned to develop **globally aware local explanations** for prediction justification, then develop **global explanations for detecting undesirable bias** in data. The final step will be to **develop trainable attention for eliminating influences of undesirable bias** in data on the model.