Marcin Pietron

Explainable AI in process of complexity reduction of DL models and in boosting theirs performance

Abstract:

The performance of AI systems based on deep learning models is exceeding the human level on an increasing number of tasks like image classification and segmentation, object detection, sentiment analysis, speech recognition, language translation or game playing. Deep learning models don?t need feature extractors, they are applied in a black box manner, no information is provided about what exactly makes them arrive at their predictions.

Our experience in reducing DL and ML models (because of the size and adaptation them to real time systems) help to understand (sensitivity analysis - based on a gradient, heat map computation, conditionality reduction) how specific layer is sensitive, which filters have low contribution in activation or are correlated together, which region or subset of weights can be removed from a layer, or how input data variations have influence on prediction accuracy and categories distinction for further model shrinking. It enables to acquire knowledge about model ability to generalize each category. Recently we can also observe changes in context DL models. LSTM and GRU cells in recurrent networks are exchange with less complex different types of attention mechanism. There is no answer if we can exchange them in most of the task and can achieve the same accuracies like in RNNs. Additionally, there is no explicit rule what combination of attention mechanism to use for specific task.

Most of the state-of-the-art rules extraction techniques are often designed for specific network architectures and specific domains, and therefore not easily adaptable to new applications. With the emergence of Deep Learning, rules extraction and interpret-ability methods that can work for networks with thousands of neurons are in high demand. We have already implemented some memetic approaches and RL methods for reducing complexity of the DL networks by acquiring the knowledge before this process. Our further goal is to use GA and reinforcement leaning combined with other ML methods to acquire knowledge of the AI model mechanism to do not only reduction of a model, but also improvement in accuracy verification of the system in a specific task. The specific goal to explain will be defined as a combination of fitness or rewards form.

Presented approaches will help to better understand how the prediction is made and which parts and regions take main part in this process. It will give more control on a prediction process. Further it will help to shrink the model by removing some less important parts of huge learning models in the specific tasks and can improve the model by changing its architecture and adapt it to specific task and input data. Another goal is to find using genetic and reinforcement learning techniques new methods or rules how DL models deal and make decision in specific tasks. In our presentation existing already implemented approaches will be

described and new methods which we want to include to extract more hidden information from DL models in wide range of specific tasks.
Explainable Machine Learning-based Artificial Intelligence
June 11
Short talk
University of Science and Technology in Cracow
Marcin Pietron, Maciej Wielgosz
.

- Administration
