Karlsruher Institut für Technologie

Institut für Angewandte Informatik und Formale Beschreibungsverfahren

Graduiertenkolloquium Angewandte Informatik

Human-Understandable Explanations for Neural Networks

M.Sc. Anna Nguyen

The 21st century is characterized by a flood of data. To deal with this amount of data, research on neural networks has gained significant momentum over the past few years. Due to the huge success in pattern recognition, they have become a powerful tool for classification and forecasting in statistics, computer science, and economics. Despite their application in many disciplines, neural networks are black box models. They do not give any insights into the structure of the approximated function. Recent research has attempted to explain these black-boxes. However, the focus so far has been to explain decisions of a neural network in a technical way for computer science experts. As neural networks get more common, it is crucial to develop approaches that allow for explanations of neural networks understandable to non-experts. This means that a human can understand the cause of even a simple decision made by the neural network and can consistently interpret the model's result.
For this reason, this work provides a framework to gain *human-understandable explanations for neural networks* by considering specific goals of explanations. These are transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction.

In this work, we address those explanation goals with four conceptual approaches. First, we present *FAIRnets,* which aims to process metadata such as intended use and architecture information in order to make neural networks more transparent and efficient. Second, to open the neural network box, we define a new explanation quality metric *ObAlEx* for image classification. Using object detection approaches, explanation approaches, and ObAlEx, we quantify the focus of CNNs on the actual evidence which covers scrutability, trust, and effectiveness. Third, we propose *FilTag*, an approach to explain Convolutional Neural Networks even to non-experts by tagging the filters with keywords. These tags provide an explanation of what the filter does. Individual image classifications can then be intuitively explained in terms of the tags of the filters that the input image activates. These explanations enhance scrutability and trust. Last but not least, we present *TransPer*, an explanation framework for recommender systems based on neural networks. We define explanation measures based on Layer-Wise Relevance Propagation to understand the recommendation quality and find new ideas on how to improve the recommender system. This captures transparency, trust, persuasiveness, and satisfaction.

The given talk will focus on approaches FAIRnets, ObAlEx, and TransPer to showcase our framework and give insights into developed explanation techniques for neural networks.

Termin: Freitag, 11.06.2021,14:00Uhr
Ort: Onlineveranstaltung
Zoom-Meeting beitreten
https://kit-lecture.zoom.us/j/67513018205
Meeting-ID: 675 1301 8205
Kenncode: 046870
Veranstalter: Institut AIFB, Forschungsgruppe Web Science

Zu diesem Vortrag lädt das Institut für Angewandte Informatik und Formale Beschreibungsverfahren alle Interessierten herzlich ein.

A. Oberweis, H. Sack, A. Sunyaev, Y. Sure-Vetter (Org.), M. Volkamer, J. M. Zöllner

Besucheranschrift:
KIT-Campus Süd
Institut AIFB – Geb. 05.20
Kaiserstr. 89
76133 Karlsruhe

Postanschrift:
KIT-Campus Süd
Institut AIFB – Geb. 05.20
76128 Karlsruhe