Chair for Integrated Systems
TUM Department of Electrical and Computer Engineering
Technical University of Munich

Seminar

Black-box and White-box Methods for Explainable Artificial Intelligence

Deep Neural Networks (DNNs) are dominating various tasks such as classification, object detection, speech recognition and natural language processing. Despite their undisputable success, understanding the decision-making process of DNNs and identifying key factors involved in the process remains an on-going challenge. Gaining further insights in the decision-making process of DNNs is especially crucial for safety-critical applications such as autonomous driving, where explainable artificial intelligence (XAI) aims to produce interpretations for machine learning based decision. Explainability methods in the field of computer vision (CV) can be divided into two categories, namely proxy (LIME, ...) and direct strategies (Guided Grad-CAM). As the name implies, proxy strategies rely on a proxy model that approximates the DNN of interest, where the decision-making of the DNN at hand is interpreted by querying the proxy model. These proxy approaches are considered as "black-box" approaches of explainability methods. Contrary, direct strategies represent "white-box" approaches of explainability methods requiring the networks parameters of the target DNN such as gradients and activations of respective layers. Utilizing these internal parameters, direct strategies are able to identify the key factors within the input that are crucial for the decision-making process of the target DNN.

In the scope of this seminar topic, state-of-the-art explainability methods are identified and surveyed. Additional to highlighting the principles of state-of-the-art techniques, advantages and disadvantages of the presented technique should be portrayed.

Contact

Lukas Frickenstein

Email: Lukas.Frickenstein@bmw.de

Advisors

Nael Yousef Abdullah Al-Fasfous