Universität Stuttgart
Institut für Aerodynamik und Gasdynamik
Wankelstraße 3, 70563 Stuttgart
NRG

Feature Importance in Neural Networks

Bachelor Thesis/ HiWi Position (m/w/d)

Background

Artificial intelligence (AI) is becoming increasingly important in many areas of our everyday lives, e.g., in administration, jurisdiction, medicine, or science. Neural networks (NNs), which build the foundation of modern AI, are often criticized for being black-box models lacking of an explanation of their predictions which is comprehensible for humans. Such explanations are required for many reasons, for instance, in order to ensure the model has learned causal and not just statistical relations or to verify that a NN's decisions are not discriminating minorities. A very popular means of explanation is to determine the importance of the NN's inputs, i.e., their relative contributions to a particular prediction. DeepSHAP, which is one method to determine such importance scores, has found widespread usage, but relies on two fundamental assumptions. First, the features must be statistically independent, which is barely the case in scientific problems. Secondly, it assumes the NN was a linear model, which is plain wrong.

Project Description

This thesis is supposed to answer the question, whether DeepSHAP, whose output is often used to increase the trust in NNs, is itself trustworthy. To this end, the explanations obtained by DeepSHAP shall be compared with those obtained by the more copmutationally expensive direct evaluation of the SHAP values. In particular, the work is structured as follows:

- Literature review
- Gather a few popular ML datasets with correlated and uncorrelated features
- Setup an automated NN training and testing pipeline in Python
- Implement the direct computation of SHAP values and their DeepSHAP approximation
- Benchmark DeepSHAP against SHAP with respect to its accountability
- Documentation of the results

Your Profile

- Solid Python programming skills
- Basic understanding of feed-forward neural networks

Application

If you are interested, even if you don't fulfill all the listed requirements, please send your application (incl. CV and latest transcript) via e-mail.

Start: as soon as possible

Contact Person:
Hannes Mandler
Institut für Aerodynamik und Gasdynamik
E-Mail: hannes.mandler@iag.uni-stuttgart.de
Tel.: +49 711 685 63461