Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Master Thesis

Simplifying a Neural Network with Concept Activation Vectors

Neural networks (NNs) are unmatched in prediction accuracy for some tasks, such as image classification. But besides making accurate predictions, an NN must also satisfy legal, economical and ethical constraints. Veryfying such constraints often requires human judgment and thus some sort of *human understanding of the NN*. One key challenge is that such understanding can only be approximate, because NNs have far more parameters than one can comprehend.

A recent approach to approximate NNs is *Testing Concept Activation Vectors (TCAV) [1]*. TCAV computes how much a NN relies on selected high-level concepts, say, “stripes” for image data. TCAV’s novelty is to map such concepts to “concept activation vectors” (CAVs). These are direction vectors in the real-valued space of a NN layer. Given an input to the NN, a CAV is considered “relevant” to the output of the NN if it points in the same direction as the gradient of the layer. For a given sample of inputs, TCAV counts how often each CAV is relevant. This yields a *TCAV score* for each concept. With these scores, users can verify whether the NN relies on their expected or required concepts.

This thesis takes CAVs a step further: instead of describing a NN with CAVs, this thesis integrates CAVs into the NN. Each CAV presents a binary linear classifier that detects a concept using the values of a hidden layer. Equivalently, each CAV presents a “concept neuron” that fires whenever the concept is detected. Integrating such concept neurons into the NN could make it easier for users to comprehend the NN itself. Based on this idea, the following questions are of interest:

- How do concept neurons help to “understand” NNs? For example, can one alter *rule extraction* algorithms to rewrite the NN as a list of rules about concepts [2]?
- How to select concept neurons so that the NN achieves good prediction accuracy? For instance, is their TCAV score relevant? Can one automate the selection of concept neurons for some NNs?

This results in the following tasks:

- Reviewing existing approaches to “understand” NNs, especially those that also identify neurons with "concepts".
- Defining a new approach to find and integrate concept neurons into a NN, based on CAVs. Implementing the approach in a NN for image classification.
- Investigating to what extent different kinds and numbers of concept neurons influence prediction accuracy and explaining this.
- Leveraging the concept neurons to “understand” the NN, for example by extending rule extraction algorithms. Discussing the quality and limitations of the gained “understanding”.

To help you with this work, we offer thorough mentoring and support from our research group, together with the required computing infrastructure. We expect basic knowledge in Python programming and machine learning, and a high motivation to explore, plan and work on the topic independently.

[1] Kim, Been, et al. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." *International Conference on Machine Learning*. PMLR, 2018.

[2] Zilke, Jan Ruben, Eneldo Loza Mencía, and Frederik Janssen. "DeepRED – Rule Extraction from Deep Neural Networks." *International Conference on Discovery Science*. Springer, Cham, 2016.

Ansprechpartner

Moritz Renftle, M.Sc. moritz.renftle@kit.edu +49 721 608-44066 Raum: 338

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34

Master Thesis