Unraveling the DNN Black-Box: Explaining Decisions of Multi-Task Deep Learning Architectures

Explaining the decision process of a deep neural network (DNN) black-box has gained significant interest with the introduction of *relevance propagation* based methods. In a nutshell, state-of-the-art algorithms interpret DNN decisions by generating so-called *heat-maps* that overlay with the input image, highlighting the relevant and relative contributions of the input image to the calculated output score. In this project we will focus on this in the context of DNNs that are trained to solve classification problems within a multi-task formulation. Such multi-output networks can be naturally trained to solve a problem by decomposing them into several smaller scale prototypical problems. We will explore these methods on large-scale image datasets for classification between different species of animals, which are even challenging for humans (e.g., Caltech-UCSD 200 Birds dataset). Finally we will analyze these interpretations in adversarial training settings.

Goals & Tasks

- Review of the state-of-the-art on DNN interpretability methods.
- Implementing and experimenting with DNNs trained to solve large-scale image classification tasks with multi-task output representations.
- Exploring and developing novel ideas to adversarially train and interpret the decisions of multi-output DNNs

Qualifications

- Interest in deep learning.
- Experience with Python based deep learning frameworks such as TensorFlow or PyTorch are beneficial.
- Registered to one of the following:
  - ✓ **Bachelor Thesis**
  - ✓ **Seminar Project**
  - ☐ **Master Thesis**

  Extension for a Master Thesis is possible on interesting results.

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at