Explainable AI – Term paper

Adversarial Images to trick MNIST model

Introduction

Twenty years from now nearly every car on the road has been replaced by an autonomous vehicle. Every turn, lane switch, acceleration and break is powered by a deep learning network running inside the car. Assuming we are cruising through a city and the car stops abruptly. Probably it was just a big print of a girl with a balloon. What happened? The car might just have recognized an *Adversarial Image*.

Adversarial images are images that have been intentionally adjusted to deceive machine learning models but look innocent to humans and can be used for *Adversarial Attacks*.

Idea

The idea of this seminar is to generate and analyze *Adversarial Images* that were generated to trick an image classifier for handwritten digits. The classifier was trained based on the MNIST dataset.

Procedure

1. Train a Model using Tensorflow and the famous MNIST dataset
2. Analyze some of the images in the validation set using an explainer methodology covered in the course (LIME, SHAP, …)
3. Generate *Adversarial Images* that let the model predict unexpected classes
4. Analyze the resulting *Adversarial Images* using explainer methodologies out of the course (LIME, SHAP, …)

Raffael Schmid, 11.11.2021

Links

Adversarial images and attacks with Keras and TensorFlow	https://www.pyimagesearch.com/2020/10/19/adversarial-images-and-attacks-with-keras-and-tensorflow/
Understanding Adversarial Attacks on Deep Learning	https://arxiv.org/abs/1907.10456
Explaining and Harnessing Adversarial Examples	https://arxiv.org/abs/1412.6572
Hacking the Brain with Adversarial Images	https://spectrum.ieee.org/hacking-the-brain-with-adversarial-images