Explainable AI – Term paper

Adversarial Images to trick MNIST model

Introduction

Twenty years from now nearly everything will be powered by some a Machine Learning based model (cars, planes, companies, etc.). Over time it will get more and more interesting to trick those models by executing Adversarial Attacks which means to change input to fool the model. Likewise, it will be important to make models against this kind of malicious attacks.

Idea

The idea of this project is to train a model based of the handwritten digits dataset (MNIST) and analyze some of the images using an explainer method out of the course. Afterwards I will choose an algorithm to generate Adversarial images and try to trick the image classifier based on the newly generated dataset.

Procedure

1. Train a Model using Tensorflow and the MNIST dataset
2. Analyze some images in the validation set using an explainer methodology covered in the course (LIME, SHAP, …)
3. Search method to generate Adversarial Images that let the performance (recall) of the model go down.
4. Analyze the resulting *Adversarial Images* using previously selected explainer methodology

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