Exposing Adversarial Vulnerabilities of Deep Neural Networks in Practice

Deep neural networks (DNNs) are shown to be vulnerable to a variety of malicious attack paradigms. One category of so-called *adversarial attacks* exploit intentionally crafted additive input perturbations that are hardly perceptible to humans, leading to incorrect decision making with high confidences. On a different category, *adversarial weight attacks* can negatively influence the inference process of a neural network for a particular test input during the deployment stage, by flipping specific bits of the weight representations stored in the memory and altering the weights of the DNN for malicious purposes. In this project we want to explore these type of phenomena from a practical perspective, subject to classification tasks that are beyond conventional image recognition benchmarks. An illustrative practical example would be attacking a biometric person identification system on a longitudinal scale (causing week-to-week failure), which relies on identifying a user/person from retinal scans (processing of retina images) or brain waves (time-series electroencephalography (EEG) signals). Finally we will investigate state-of-the-art defenses to these problems in order to assess their efficiency in practical use, and develop other novel ideas towards addressing this problem.

Goals & Tasks

- Review of the state-of-the-art on adversarial (weight) attacks on DNNs.
- Implementing real-life practical tasks with DNNs and simulating attack scenarios.
- Exploring and extending novel defense methods to adversarial (weight) attacks.

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**

Contact

Ozan Özdenizci
ozan.ozdenizci@igi.tugraz.at