Regularising Adversarial Training and Identifying Attacks

Introduction: Deep learning models have become a common feature in industry as well as everyday life. They are used to make decisions in autonomous driving, healthcare, security, and more. However, it has been shown that these models are vulnerable to small adversarial perturbations to their input, which can completely change their prediction. This limits the deployment of deep learning models in industrial applications. This research topic aims at training more robust neural networks.

Objectives: Combine adversarial training with (a subset of) the following methods:

- Proven regularizers (L1-, L2-Norm, Dropout)
- Optimization techniques (SGDM, Adam, AdamW, etc.)
- Curriculum Learning (create stronger attacks at later training stages)
- Hierarchical classification (learn to detect attacks + predict a class)

Those approaches are expected to improve the robustness of the model and enable the simultaneous classification and identification of adversarial attacks.

Requirements: Strong knowledge of deep learning; experience in programming with Python

Contact: rene.raab@fau.de, leo.schwinn@fau.de