SCIENCE · PASSION · TECHNOLOGY

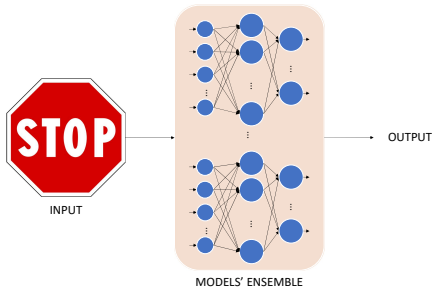
Open Thesis / Project

# Moving Target Defense Against Adversarial Examples on Embedded Devices

Embedded Information Processing Team

## Motivation

Adversarial attacks represent attempts to design malicious inputs for ML models to make them yield wrong predictions. Input perturbations are usually of a small scale, e.g., a sticker pasted on a road sign, and are perceived as valid inputs by humans. In contrast, deep neural networks (DNNs) are known to be vulnerable to adversarial attacks. The knowledge of a trained DNN (or only its outputs) enables attackers to construct efficient adversarial examples to mislead the model. Moving target defense (MTD) is a model-agnostic defense mechanism against adversarial attacks, based on frequently updating the model and leaving an attacker no time to learn model output peculiarities to design effective adversarial examples. State-of-the-art MTD approaches are based on frequent model re-training and/or aggregating the output from an ensemble of models while frequently replacing its members. Both approaches generate high computational and memory overhead, which is problematic for embedded devices and limits their applicability in practical applications. We have an elegant solution to the problem and need you to implement it and measure its efficiency. **Interested? Contact us for more details!**




## Goals and Tasks

In this project you will be exploring a way to achieve an adequate inference speed for a frequently changing DNN ensemble designed against adversarial attacks. The project includes the following tasks:

- Literature review on adversarial learning and defense mechanisms;
- Implement methods for fast inference in adversarial deep learning models' ensemble using linear mode connectivity; compare your method to existing methods;
- Summarize the results in a written report and prepare an oral presentation.

## Requirements / Skills:

- Programming skills in Python and C++; interest in efficient code writing;
- Prior experience with deep learning frameworks (preferably PyTorch).

## Target Group

Students in ICE and Computer Science.

## Thesis Type

Master Project / Master Thesis.

## Contact:


- Dr. Olga Saukh (saukh@tugraz.at)
- MSc Katarina Milenković (katarina.milenkovic@pro2future.at)