Lehrstuhl für Sicherheit in der Informatik
Prof. Dr. Claudia Eckert

Testing and Evaluation in Adversarial Machine Learning (HiWi)

Motivation

Adversarial machine learning is a technique employed in the field of machine learning which attempts to fool models through malicious input. This technique can be applied for a variety of reasons, the most common being to attack or cause a malfunction in standard machine learning models. An attacker can break machine learning systems, such as by poisoning the data used by the learning algorithm or crafting adversarial examples to directly force models to make erroneous predictions.

Suppose a researcher proposes a new defense procedure and evaluates that defense against a particular adversarial example attack procedure. If the resulting model obtains high accuracy, does it mean that the defense was effective? Possibly, but it could also mean that the researcher's implementation of the attack was weak. A similar problem occurs when a researcher tests a proposed attack technique against their own implementation of a common defense procedure. By testing, we mean evaluating the system in several conditions and observing its behavior, watching for defects.

To resolve these difficulties, we intend to create a testing tool specialized for adversarial machine learning attacks and defences that researchers and product developers can use the tool to test their models against standardized, state-of-the-art attacks and defenses.

Requirments

- Python programing
- Having experience in adversarial machine learning

Your Task

Currently, we involve in a European project called SPARTA [1] that one of the main goal of the project is to investigate approaches to make systems using AI more reliable and resilient through enhanced explainability and better threat understanding by providing methods and tools for analysis of security threats for AI systems, and solutions for protection.

For this Hiwi position, we are looking for skillful and motivated student to support us in the implemention of the specialized testing and evaluation tool for adversarial machine learning.

Contact

Mohammad Reza Norouzian
Technische Universität München
Chair for IT Security (I20)
Boltzmannstraße 3, 85748
Garching Tel. + 49 89 289 18584
norouzian@sec.in.tum.de

1 https://www.sparta.eu/