Master thesis proposal

Identification of adversarial examples in network traffic using detectors

Arkadiusz Warzyński

November 29, 2024

# Background

Over the past few years, several studies have been conducted on the phenomenon of adversarial examples, which can be a potential threat to artificial intelligence models. The most significant threat in the case of classification algorithms is the possibility of misclassification, which in many applications could mean incorrectly recognizing a disease, a road sign or ignoring a phishing attempt. Current solutions in use in this area guarantee sufficient classification quality to be considered acceptable and safe under normal conditions. However, it is possible to make intended, precisely prepared changes to the analyzed data so that it leads to misclassification through the chosen model. One possible solution to the problems is detectors, which are an extension of the original model architecture and introduce additional layers responsible for performing certain modifications to the classified observations. The goal is to highlight differences in observations from the original sets and adversarial examples.

# Objectives

The objective of the work is to apply the concept of the adversarial examples detector, which has so far been mostly discussed for datasets related to the computer vision problem, to the detection of adversarial examples generated for network traffic. The scope of the work includes the implementation of known methods for generating adversarial examples, such as the Fast Gradient Sign Method, and the preparation of suitable architecture for the classification model, which will consider the presence of the adversarial examples detector, and its implementation. It is also necessary to evaluate the quality of classification before and after the detector is introduced into the selected classifier architecture.

# Data description

The CICIDS collection contains network traffic generated in a laboratory environment by profiling 25 abstract network users. User behavior was modeled using the B-profile system by simulating the use of HTTP, HTTPS, FTP, SSH and e-mail services. The data collection period covered 5 days. At the same time, it is one of the newest and most widely used datasets for the problem of intrusion detection in network traffic, which shows great imbalance due to the small representation of data on attacks carried out.