Robustness and Adversarial Machine Learning

Mohammed Abuhamad
Loyola University Chicago


Soteria

Detecting Adversarial Examples in Control Flow Graph-based Malware Classifiers. [3]

AdvEdge

Optimizing Adversarial Perturbations Against Interpretable Deep Learning [1]. Other exploration can be found in [4].

Figure 1. The architecture of Soteria. IoT samples are fed to the feature extraction process, where each sample is represented by multiple feature vectors. The feature vectors are forwarded to adversarial example detector. All non-AEs are then forwarded to the classifier to be classified into its corresponding family.

Figure 2. Soteria feature extraction process. IoT samples binaries are disassembled to extract their corresponding CFGs. Then, two node labeling techniques are used (Dense-based and level-based), then, several random walks are done over each labeled graph. The trace of the random walk is then used for feature extraction.

Figure 3. DL-FHMC system flow. First, corresponding CFGs of the IoT software are extracted, then, 23 algorithmic features are extracted from the CFGs. Afterward, an IoT malware detection system classifies samples into benign and malware, all malware samples are directed to IoT malware classification system, while benign samples are directed into suspicious behavior detection system (SBD) for further investigation.

Figure 4. Example images for (a) benign, (b) regular adversarial and (c) dual adversarial and interpretations on ResNet (classifier) and CAM (interpreter).

We present AdvEdge and AdvEdge+, two attacks to mislead the target DNNs and deceive their combined interpretation models. We evaluate the proposed attacks against two DNN model architectures coupled with four representatives of different categories of interpretation models. The experimental results demonstrate our attacks’ effectiveness in deceiving the DNN models and their interpreters.

References


