Adversarial Machine Learning for Detecting Malicious Behaviour in Network Security

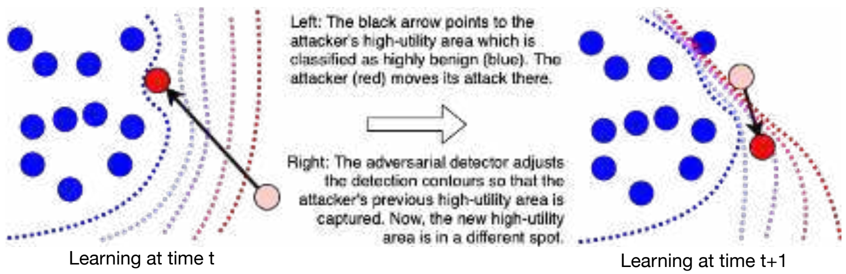
Author: Ing. Michal Najman
Supervisor: Mgr. Viliam Lisý, MSc, Ph.D
Czech Technical University in Prague, Faculty of Electrical Engineering

Attack detectors are seen to be circumvented by malicious actors who purposely adjust their activity to mimic benign behaviour. We use the following industrial problem as a running example:

- A URL reputation service provides a rating of a website to protect a user
- A malicious actor monitors the rating of its malicious websites to check for disclosure and does so in such a way its activity is not detected
- **Our goal**: to design a detector that identifies malicious users and benign users based on their activity history, i.e. the sequence of URLs which a user requested for evaluation

By uniquely fusing **risk minimisation** and **game theory**, we arrive at a detector's learning algorithm that outputs a **detector robust to adversarial attacks**. The detector is tested on real-world data by Trend Micro Ltd.

Intuition Behind Detector's Learning Algorithm



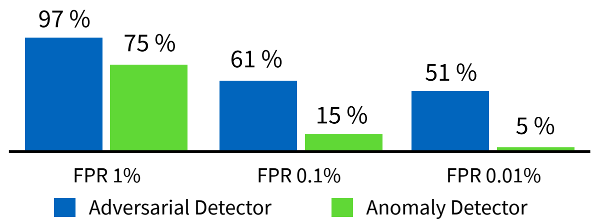
PROBLEM & PROPOSED METHOD

- **The adversarial detection task** is limited by the following **constraints**:
  - Only benign activity recorded and malicious actors adjust their behaviour to circumvent the detector ▸ we propose a **model of an attacker** to generate the benign activity (attacks)
  - Industrial applications require to **constrain the false positive rate** (FPR)
- We derive **the detector's learning algorithm** combining uniquely the expected risk minimisation framework and the game theory:
  - **The Neyman-Pearson task** [1]: minimise the risk on the attacks while keeping the risk on benign users below a threshold
  - **The Stackelberg equilibrium**: the detector is fixed after deployment and the attackers perform attacks that aim to circumvent it ▸ bilevel optimisation problem

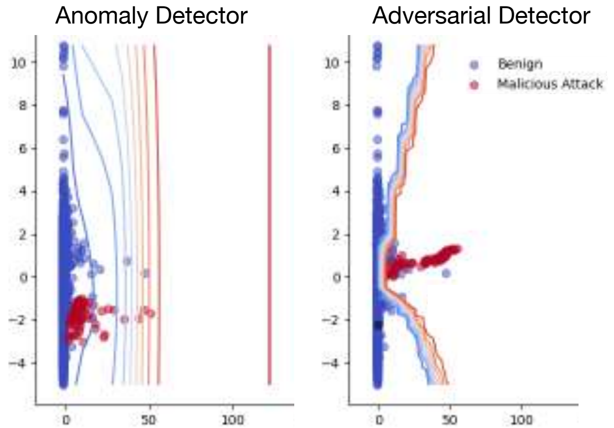
MODEL OF ATTACKER

- The attacker is a **utility-maximising agent**
- The **utility is differentiable** and consists of:
  - A penalty for being detected
  - A penalty for not attacking
  - A penalty for attack complexity
- **Action space**: set of all URL sequences
- The best-response **attack approximation**:
  - Iterative improvement using gradient descent
  - Based on PGD [5] and FGSM [6]

Detector Robustness (Attack Detection Rate)



The bar chart shows that **the adversarial detector is more robust to attacks than the anomaly detector**. Attack Detection Rate is the true negative rate and amounts to the portion of attacks that are detected during test time. The higher the value, the more robust the detector is.



The PCA view of the feature space: The attacks are in red, the blue dots are benign data, FPR = 0.01%. Notice that the **adversarial detector** (right) adjusted its maliciousness contours to the attacker's high-utility areas. Whereas the **anomaly detector** (left) only exploited benign data distribution.

ADVERSARIAL DETECTOR

- Given a sequence of requests to a URL reputation service, the detector outputs the maliciousness rate
- The optimal detector is approximated with a custom five-layer neural network with SeLU activations
- **The detector's learning algorithm:**
  - The detector iteratively learns to detect best-response attacks with gradient descent *(image left)*
  - Inspired by Exploitability Descent [2], StackGrad [3] and gradient descent with constraints [4]
- Key principle of the detector's weights θ update:

$$\Delta\theta = \lambda_t \cdot \mathbb{E}\,\nabla_\theta D_{\theta_t}(\text{data}) - \mathbb{E}\,\nabla_\theta D_{\theta_t}(\text{attacks})$$

The update $\Delta\theta$ is a mixture of the expected gradients on the benign data and the attacks. The multiplier λ_t controls the mixing ratio and D_{θ_t} stands for the detector's output, i.e. maliciousness rate.

EXPERIMENTS & CONCLUSIONS

- **The adversarial detector outperforms an anomaly detector** on real-world data *(top center)*
  - The anomaly detector is based on k-nearest neighbours
- The **adversarial detector successfully detects attackers** querying the URL reputation service and **meets the desired FPR constraint**
- The maliciousness contours of the adversarial detector better reflect the attacker's high-utility areas whereas the anomaly detector only wraps the benign data *(bottom center)*
- The thesis is a **proof of concept for adversarial machine learning applications**

Implemented using: PyTorch
Source codes available at: bit.ly/thesis-adversarial-ml

References: [1] Neyman et al., „On the problem of…", 1933. [2] Lockhart et al., „Computing approximate equilibria…", 2019. [3] Amin et al., „Gradient methods for stackelberg security…", 2016. [4] Janisch et al., „Classification with costly features…", 2019. [5] Kurakin et al. „Adversarial machine learning at scale", 2016. [6] Goodfellow et al., „Explaining and harnessing examples…", 2015.

najmanm@gmail.com, viliam.lisy@agents.fel.cvut.cz