Introducing My New PhD Research: Adversarial Robustness in Network Intrusion Detection System

Mouzaoui Matthieu

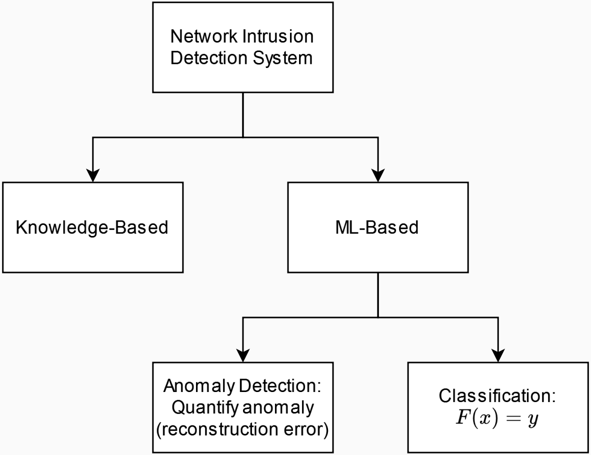
06-03-2024

PhD student, PIRAT); , Inria

Background

- New PhD student in PIRAT.
- ML, probability background.
- Supervised by:
    - Yufei Han, Inria.
    - Michel Hurfin, Inria.
    - Gabriel Rilling, CEA-List.
    - Gregory Blanc, Télécom-Sud Paris.
- Title: Adversarially Robust Machine Learning based Network Intrusion Detection System.

NIDS Model



Focus on ML-NIDS. ML-based $\implies$ vulnerable to adversarial attacks.
First spotted against Neural Networks in [Szegedy et al., 2014].

Adversarial Sample Example

Example of Adversarial Sample

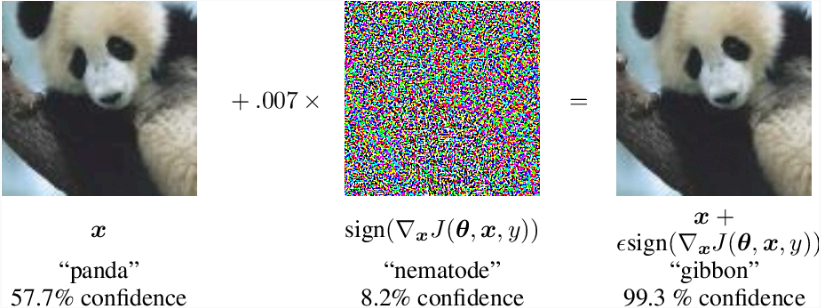


Figure 1: Adversarial sample generation, from [Goodfellow et al., 2015]

Adversarial Attacks against ML model

Targeted Phase
Training or inference time.

Adversarial Sample
Model $\mathbf{x} :\mapsto F(\mathbf{x})$. Given $\mathbf{x}$, find perturbation $\boldsymbol{\delta}$ such that $t = F(\mathbf{x}+\boldsymbol{\delta}) \neq F(\mathbf{x})$ or, if $\mathbf{r} = \mathbf{x}+\boldsymbol{\delta}$, $\tilde{r} = \mathit{Decode}(\mathit{Encode}(\mathbf{r}))$, $\|\tilde{r} - \mathbf{r}\|_p \leq \alpha$.

Evasion
$\mathbf{x}$ a malicious sample, the attacker wants $F(\mathbf{x}+\boldsymbol{\delta}) =$ *'benign'*. $\rightarrow$ evasion.

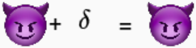
Optimization problem
Maximize loss of classifier / cross the threshold, minimizing norm of perturbation.

Evasion in network domain
Developed in Computer Vision:

- Features: pixel, range known.
- Dependencies

Constraints specific to ML-NIDS

$\mathbf{x} + \boldsymbol{\delta}$ should satisfy some properties:

- Validity (can be transmitted). 
- Plausibility (similar to real traffic). 
- Preserved Semantic (coherent with its purpose). 

- Robustness to preprocessing ($\boldsymbol{\delta}$ not removed). 

Most papers focus on **feature-level** attacks, features = Netflows.

Constraints from [Pierazzi et al.,] and [Vitorino et al., 2023]

Orientations

Still in review process, however, identified 2 gaps:

- Validity. Now: ensured by expert knowledge.
- Preserved Semantic. Now ”justified” though bound of $\|\boldsymbol{\delta}\|_{l_p}$.

Inverse feature mapping. Uses graph representation.

References i

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015).
Explaining and harnessing adversarial examples.

Pierazzi, F., Pendlebury, F., Cortellazzi, J., and Cavallaro, L.
Intriguing properties of adversarial ML attacks in the problem space.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014).
Intriguing properties of neural networks.

Vitorino, J., Praça, I., and Maia, E. (2023).
Towards adversarial realism and robust learning for iot intrusion detection and classification.
Annals of Telecommunications, 78(7–8):401–412.