2023-06069 - PhD Position F/M Adversarially Robust Machine Learning-based Network Intrusion Detection System

Type de contrat : CDD
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

Contexte et atouts du poste

Within the framework of the ANR PEPR project "Superviz".

Mission confiée

Background

Adversarial attacks against machine learning (ML) systems aim to manipulate training and/or test examples to cause misclassification of ML models. Widely observed in computer vision and natural language process research, slightly changing the pixel values of an input image or words of an input corpus may drastically mislead the classification output from a machine learning model. It is also well known as evasion attack [1][2]. Similarly, injecting noise to the training data set of a target machine learning model can also cause misclasification over the attacker-desired classes like in poisoning [3] and backdoor attacks [4]. Demonstrated in previous works, such adversarial risks [1][2][3][4] can bring signficant harm to the trustworthiness of machine learning systems. With the wide deployment of machine learning techniques in intrusion detection and classification, the vulnerability of machine learning systems becomes a critical factor that impacts the utility and reliability of the machine learning-based security practices. It is hence a must to assess the adversarial risk of machine learning in security data analysis and discuss how to harden machine learning-based detection systems.

The objective of this thesis

In this research, we focus on assessing the risk/impact of ML-based detection algorithms in the presence of adversarial threats and designing and implementing potential mitigation solutions to defend against these attacks and strengthen ML-based detection applications.

Our study explores the following research problems:
- We first define realistic adversarial threat models against ML-based network intrusion detection systems. In particular, we focus on input perturbation at test time (evasion attack) and data perturbation at learning time (data poisoning attack). Unlike adversarial attacks on images, network and system security log disruption aims to evade detection methods while ensuring successful intrusions. We need to consider domain constraints on data disruption patterns, such as which behavioral attributes of an intrusion incident can (or cannot) be modified/deleted. In addition, the input data of intrusion detection systems typically contains unstructured categorical attributes, e.g., short textual descriptions of security events. How to perturb these discrete attributes is inherently a combinatorial problem and remains open in adverse learning research.
- We study the identification of key factors that determine the risk of weakness in the face of an adversarial attack of an ML-driven intrusion detection model. We start by examining classical ML models, such as support vector machines and decision trees, and then extend our scope to more advanced models based on Deep Neural Networks (DNNs), especially in approaches that take into account the temporal dimension of attacks (LSTMs, Transformers). We seek to answer the following questions 1) What properties (smoothness, model complexity, ability to generalize out of distribution, etc.) of the detection model are responsible for the adversarial vulnerability of the ML-based detector? 2) Is the adversarial vulnerability also associated with the security data used to train the detection model (feature sensitivity, feature redundancy, data sparsity, etc.)? How can quantified measures be defined to reflect the level of adverse vulnerability of the ML-based detector?
- Our goal is to propose defense mechanisms based on the identified risk factors to enhance the robustness by construction of ML-based detection models. For example, differential privacy has been shown to be an effective tool to improve the robustness of DNN-based classifiers. In addition, we investigate how to establish health checking methods to identify potentially poisoned training and test inputs in ML-based intrusion detection services.
We propose to evaluate our approaches using publicly available network intrusion datasets collected from real devices. Two examples of data sources are CIC-IDS-2018 [5] and DAPT2020 [6]. The former provides a large-scale labeled dataset containing network traffic of normal and intrusion behaviors. This dataset provides rich descriptions of network traffic profiles, e.g., pcap files, to facilitate the description of intrusion incidents. The second gives a collection of network traffic simulating Advanced Persistent Attacks (APT). This is a good test bed for evaluating how ML-based intrusion detection models perform against commonly deployed APT attack techniques.

Expectations

The candidate for this thesis is expected to have accomplished courses on Machine Learning and/or have experience of implementing Machine Learning algorithms using Python for practical data mining problems. Especially, expertise in using Pytorch will be required in the project. Theoretical developments are also expected based on statistics and theory of machine learning and approximation. Knowledge about intrusion detection systems will be preferred.

References

[1] M. A. Ayub, W. A. Johnson, D. A. Talbert and A. Siraj, "Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning," *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2020, pp. 1-6, doi: 10.1109/CISS48834.2020.1570617116.

[2] B. Biggio, I.Corona, D.Maiorca, B.Nelson, N.Srndic, P.Laskov, G.Giacinto and F.Roli, "Evasion Attacks against Machine Learning at Testing Time", Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013.

[3] Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume Iii, and Tudor Dumitras. When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks. In 27th USENIX Security Symposium (USENIX Security 18), pp. 1299–1316, 2018. ISBN 978-1-939133-04-5.

[4] Manoj, Naren and Avrim Blum. "Excess Capacity and Backdoor Poisoning." *Neural Information Processing Systems* (2021).

Informations générales

- **Thème/Domaine :** Représentation et traitement des données et des connaissances
  Statistiques (Big data) (BAP E)
- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2023-10-02
- **Durée de contrat :** 3 ans

Contacts

- **Equipe Inria :** CIDRE
- **Directeur de thèse :**
  Han Yufei / yufei.han@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 200 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3500 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 180 start-up. L'institut s'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

The candidate for this thesis is expected to have accomplished courses on Machine Learning and/or have experience of implementing Machine Learning algorithms using Python for practical data mining problems. Especially, expertise in using Pytorch will be required in the project. Theoretical developments are also expected based on statistics and theory of machine learning and approximation. Knowledge about intrusion detection systems and anomaly detection will be preferred.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

[5] CIC-IDS 2018 dataset:https://www.unb.ca/cic/datasets/ids-2018.html

[6] DAPT-2020 dataset: https://sailik1991.github.io/files/DAPT_at_MLHat2020.pdf

Principales activités

This thesis will be conducted at INRIA Rennes and co-supervised with the AI researchers of CEA List team in Paris. The Inria Rennes - Bretagne Atlantique Centre is one of Inria's eight centres and has more than thirty research teams. The Inria Center is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc. The monthly gross salary for the PhD candidate amounts around 2000 euros. For every applicant, please submit online your resume, cover letter and letters of recommendation.

Compétences

Technical skills and level required : Machine Learning, Statistics, Information theory, Pytorch, Intrusion Detection

Languages : English

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage