Internet of Things, Ad Hoc and Sensor Networks Technical Committee Newsletter

(IoT-AHSN TCN)

Volume 1, No. 19

December, 2023

CONTENTS

PREFACE

The IEEE ComSoc Ad Hoc and Sensor Networks Technical Committee (IoT-AHSN TC) sponsors papers, discussions, and standards on all aspects of IoT, ad hoc and sensor networks. It provides a forum for members to exchange ideas, techniques, and applications, and share experience among researchers. Its areas of interest include systems and algorithmic aspects of sensor and ad hoc networks, networking protocols and architecture, embedded systems, middle-ware and information management, novel applications, flow control and admission control algorithms, network security, reliability, and management. In an attempt to make all the TC members as well as the IoT-AHSN worldwide community aware of what is going on within our main areas of concerns, this newsletter had been set up. The newsletter aims at inviting the authors of successful research projects and experts from all around the world with large vision about IoT-AHSN-related research activities to share their experience and knowledge by contributing in short news.

The nineteenth issue of the IoT-AHSN TC Newsletter focuses on the theme “Generative AI for Internet of Things”. Specifically, this issue includes 1 news article: A Concise Survey on Adversarial Attacks against Network Intrusion Detection Systems. We thank the contributor for their efforts to help make the IoT-AHSN TC Newsletter a success. We hope that the methods/approaches presented in this issue could significantly benefit researchers and application developers who are interested in IoT and ad hoc/sensor networks.

Finally, it is worth noting that Dr. Mohammed Atiquzzaman receives the 2023 IoT-AHSN Technical Achievement and Recognition Award for his significant contributions and impacts to the technological advancement of the Internet of things, ad hoc and sensor networks.

Newsletter Co-Editors
Qiang Ye (Dalhousie University, Canada)
Moez Esseghir (University of Technology of Troyes, France)
Lu Lv (Xidian University, China)

TC OFFICERS AND NEWSLETTER EDITORS

TC Officers

Name	Affiliation	Email
Sharief Oteafy (Chair)	DePaul University, USA	soteafy@depaul.edu
Shuai Han (Vice Chair)	Harbin Institute of Technology, China	hanshuai@hit.edu.cn
Rodolfo Coutinho (Secretary)	Concordia University, Canada	coutinho@ece.concordia.ca

Newsletter Editors

Name	Affiliation	Email
Qiang Ye (Editor in Chief)	Dalhousie University, Canada	qye@cs.dal.ca
Moez Esseghir (Technical Editor)	University of Technology of Troyes, France	moez.esseghir@utt.fr
Lu Lv (Technical Editor)	Xidian University, China	lulv@xidian.edu.cn

A Concise Survey on Adversarial Attacks against Network Intrusion Detection Systems

Shiyun Wang
Faculty of Computer Science
Dalhousie University
Halifax, Canada
sh776410@dal.ca

***Abstract*—Computer networks' evolution brings new topologies, protocols, and security threats. Network Intrusion Detection Systems (NIDS), particularly machine learning-based ones, excel for adaptability and swift detection. However, they're vulnerable to adversarial examples, leading to misclassification. Adversarial attacks exploit this weakness, crafting traffic to evade detection, posing significant threats. This paper surveys such attacks, reviews prior research, classifies attacks, and assesses their performance across datasets. It stresses the need for updated NIDS benchmark datasets and thorough assessments.**

***Index Terms*—machine learning, intrusion detection, cyber attacks, perturbation methods, adversarial attacks**

I. Introduction

In recent years, the evolution of computer networks has brought about new topologies and protocols, alongside an increase in security threats, compromising data confidentiality, integrity, and availability [1]. To mitigate these risks, intrusion detection systems (IDS) play a pivotal role. IDS are categorized into signature-based, anomaly-based, and hybrid-based, and can be deployed as network-based (NIDS) or host-based (HIDS) [2]. Given the scalability and comprehensive network coverage, this paper focuses on NIDS.

Machine learning-based techniques have demonstrated superiority over traditional signature-based NIDS. This advantage arises from their ability to adapt to subtle changes in attack patterns, making evasion more challenging for attackers [2]. Additionally, machine learning-based NIDS can continuously learn and adapt, enabling efficient identification of new attack variants. However, these ML-based techniques are notably vulnerable to subtle, imperceptible input perturbations, termed adversarial examples, which can result in sample misclassification and pose a risk to NIDS.

This paper aims to provide an overview of the implementation of adversarial attacks in NIDS by surveying the literature on adversarial attacks against NIDS. We first propose a classification of adversarial attacks in Section II, dividing adversarial attacks into two major categories: white-box attacks and black-box attacks. We then describe the performance of adversarial attacks against NIDS on different datasets in Section III. Finally, we discuss and conclude the paper in Section IV.

II. Adversarial Attacks on NIDS

This section presents white-box and black-box adversarial attacks on NIDS.

A. White-box Attacks

White-box attacks are predicated on the assumption that the attacker possesses comprehensive knowledge identical to that of the targeted IDS.

1) Projected Gradient Descent: In the Projected Gradient Descent (PGD) attack [3], adversarial examples are created by optimizing a negative loss function, as depicted in Eq. 1.

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot \text{sgn}(\nabla_x L(\theta, x, y))) \tag{1}$$

S represents the spherical limit, Π represents the process of first computing the loss gradient to the original example to obtain the adversarial sample, then subtracting the original sample from the adversarial sample to obtain the perturbation, constraining it to the ℓ_∞-ball range, and finally adding the perturbation to the original example to obtain the final result.

2) Jacobian-Based Saliency Map Attack: Jacobian-based Saliency Map Attack (JSMA) primarily relies on the forward gradient to assess the influence of each data on the model's classification result and identify the input features of x that change most significantly. It is important to note that while JSMA introduces minor perturbations to the feature subset, this comes at the expense of significantly increased computational complexity and time.

3) CW Attack: The Carlini and Wagner's (CW) attack adds imperceptible perturbations to attack samples, causing the model to provide an incorrect label with high confidence. The CW attack defines an objective function, denoted as g, such that $f(x + r) = l$ if and only if $g(x + r) \leq 0$. Thus, the CW attack can be expressed as follows:

$$\begin{aligned} \text{minimize} \quad & \|r\|_p + c \cdot g(x + r) \\ \text{subject to} \quad & x + r \in [0, 1]^n \end{aligned} \tag{2}$$

The CW attack offers three distinct strategies for generating adversarial examples: L_2 (utilizing multi-start gradient descent), L_0 (employing an iterative process), and L_∞ (requiring an iterative optimization method) attacks.

B. Black-box Attacks

Black-box attacks assume that the attacker only knows the output of the model, such as labels or confidence scores.

1) Transfer-based Attacks: In black-box scenarios, white-box attacks rely on transferability, termed transfer-based attacks [1]. These attacks involve developing surrogate models that mimic the decision boundaries of the target model, allowing the application of white-box attacks to generate effective adversarial examples. It is important to emphasize that the success of transfer-based attacks depends on the surrogate model's ability to accurately replicate the decision boundary of the target model [1].

2) Opt. Attack: Opt. attack [4] strategy aims to find an optimal sample x^* to minimize $p^{(x_{\text{mal}})}$, which denotes the probability of being classified as malicious by the substitute detector [4]. The larger $p^{(x_{\text{mal}})}$, the less likely the sample is to be considered benign. The algorithm employs the linear approximation method to iteratively approach the linear programming problem to obtain the final adversarial example x^* [4]. The number of iterations in the gradient descent method increases as the amount of data and the Opt algorithm's complexity increase, making it highly resource-intensive. Additionally, the memory required grows quadratically with the number of variables [4]. Consequently, the efficiency of the Opt attack is reduced.

3) AttackGAN: AttackGAN [5] belongs to the category of Generative Adversarial Networks (GANs), which comprises three main components: a generator G, a discriminator D, and an IDS. G takes either a noise or attack sample Z as input and produces an adversarial sample $G(z)$ as output. $G(z)$ is then fed to D, which aims to distinguish it from normal traffic samples X. The loss function, L_{WGAN}, quantifies the disparity between predicted labels and actual labels. The IDS takes $G(z)$ as input, and its output is then fed back to G to assist in generating more effective adversarial attack samples. The loss function is denoted as L_{IDS}, which quantifies the disparity between the output detection result and the target label t_{adv}. The overall objective function is as follows:

$$\min_{G} \max_{D} L = L_{\text{WGAN}} + \lambda L_{\text{IDS}} \tag{3}$$

Where $\lambda \in (0, 1)$ represents the relative importance of the two loss functions mentioned above.

III. Performance Analysis

The NSL-KDD dataset, an enhanced version of the KDD'99 dataset utilized in the DARPA'98 IDS evaluation program, is categorized into basic, traffic, and content features [6]. The UNSW-NB15 dataset encompasses nine types of attacks, including DoS, worms, and backdoors, and boasts over 2 million records [7].

In their respective studies [5]–[7], researchers evaluated the effectiveness of different adversarial attacks on various datasets. In the NSL-KDD Dataset, [6] discovered that the CW attack demonstrated less destructive behavior in contrast to JSMA. JSMA, which utilizes a higher proportion of unbalanced features, appeals to malicious actors seeking to manipulate fewer features. [7] conducted a comparative analysis of CW and PGD attacks on ANN, CNN, and RNN models using both the NSL-KDD and UNSW-NB15 datasets. They observed that the CW attack was most effective on the NSL-KDD dataset but had a comparatively lower impact on the UNSW-NB15 dataset. Interestingly, the performance of PGD was similar across both datasets. Lastly, AttackGAN was evaluated against PGD and CW attacks using five different ML/DL algorithms (SVM, RF, DT, RNN, and NB) for black-box IDS scenarios. The experimental results demonstrated that AttackGAN's attack success rate on these five IDS surpassed that of other algorithms.

These varied experimental outcomes highlight the fluctuating efficacy of attack algorithms across various datasets and IDS systems employing different algorithms. Comparative evaluations should encompass diverse experimental setups and evaluation criteria. Moreover, it's essential to recognize that most performance assessments primarily focus on evading NIDS detection, often neglecting to consider the preservation of the attacks' malicious intent as a critical evaluation metric.

IV. Disscussion and Conclusion

This survey explores the application of adversarial attacks in NIDS, examining both white-box and black-box scenarios. Our analysis of various adversarial techniques across datasets reveals a significant finding: the performance of an attack can vary greatly depending on the dataset used for testing, highlighting the pivotal role of datasets in NIDS evaluations. However, we face a notable challenge: the scarcity of comprehensive datasets tailored for NIDS assessments.

Furthermore, we emphasize a critical aspect often overlooked in existing evaluations: the malicious intent of adversarial attacks. While current assessments prioritize evasion detection, they neglect the attacks' malicious nature and potential harm. A comprehensive evaluation should consider not only the attacks' evasiveness but also their effectiveness in causing harm.

References

[1] Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar. Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1):538–566, January 2023.

[2] Kamran Shaukat, Shifu Luo, Vijay Varadharajan, Irfan A. Hameed, and Mingrui Xu. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*, 8:222310–222354, 2020.

[3] Aleksander Mańdry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2018.

[4] Jiming Chen, Xiangshan Gao, Ruilong Deng, Yang He, Chongrong Fang, and Peng Cheng. Generating adversarial examples against machine learning-based intrusion detector in industrial control systems. *IEEE Transactions on Dependable and Secure Computing*, 19(3):1810–1825, May-June 2022.

[5] Shuang Zhao, Jing Li, Jianmin Wang, Zhao Zhang, Lin Zhu, and Yong Zhang. Attackgan: Adversarial attack against black-box ids using generative adversarial networks. *Procedia Computer Science*, 187:128–133, 2021.

[6] Zheng Wang. Deep learning-based intrusion detection with adversaries. *IEEE Access*, 6:38367–38384, July 2018.

[7] Rana Abou Khamis and Ashraf Matrawy. Evaluation of adversarial training on different types of neural networks in deep learning-based idss. In *2020 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, October 20-22 2020.