2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL) | 979-8-3503-7382-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVIDL62147.2024.10603582

IDS-GAN: Adversarial Attack against Intrusion Detection Based on Generative Adversarial Networks

[1st] Di Wang
School of Cyberspace Security
PLA Information Engineering University
Zhengzhou, China
18240793465@163.com

[2nd] Xuemeng Wang
School of Cyberspace Security
PLA Information Engineering University
Zhengzhou, China
wuximo_email@163.com

[3rd *] Jinlong Fei
School of Cyberspace Security
PLA Information Engineering University
Zhengzhou, China
feijinlong_2021@163.com
[*] Corresponding author: Jinlong Fei; feijinlong_2021@163.com

***Abstract*—Intrusion detection protects network security and data integrity by monitoring and identifying malicious activities and abnormal behaviors in the network. With the increasing volume of network data, machine learning-based intrusion detection has gained popularity as the predominant approach. However, research has demonstrated that machine learning is susceptible to adversarial samples, where even a small, intentionally crafted perturbation can result in incorrect model outputs. In this paper, the idea of adversarial samples is applied to traffic obfuscation in IDS, and an adversarial obfuscation method based on generative adversarial network is proposed. This method clearly divides functional and nonfunctional features, and only adds perturbation to nonfunctional features, so as to avoid model detection while maintaining functionality of traffic. In addition, this method does not require structure and internal parameters of the model, which is a black box obfuscation method. In the experiment, the NSL-KDD dataset was employed to assess the obfuscation performance of the method on two target models. The results indicate that the method demonstrates effective obfuscation capabilities for both DoS and Probe malicious traffic.**

Keywords-traffic obfuscation, adversarial sample, machine learning, intrusion detection

I. INTRODUCTION

Intrusion Detection System (IDS) detects possible security threats and attacks by monitoring and analyzing the network communication flow in real time to help prevent unauthorized access, denial of service and other threats, so as to improve the overall security of the system [1].

With the advancement of artificial intelligence technology, machine learning-based intrusion detection has experienced significant progress and widespread adoption has gradually become the mainstream. However, research has demonstrated that machine learning models are susceptible to adversarial samples [2], [3], and obfuscators can add perturbation to traffic to cause incorrect output of intrusion detection models, thus escaping IDS. However, for traffic obfuscation, some require a comprehensive understanding of the model's structure and parameters [4],[5]; some researchers simply transfer adversarial methods in image without distinguishing between functional and nonfunctional features, thus destroying the traffic functionality [6],[7]; some gradient-based methods leverage the gradient information to craft adversarial samples [8], [9], but only apply to those models with derivability and gradient calculation, such as deep neural networks, and are not applicable to models such as decision trees that do not rely on gradients for training and prediction. In summary, these problems remain to be solved.

In this paper, to evade IDS based on machine learning, we introduces a novel approach called IDS-GAN, which utilizes a generative adversarial network (GAN) to develop an adversarial obfuscation method. The perturbation is calculated by training GAN. IDS-GAN is mainly composed of generator, discriminator and pre-trained intrusion detection model. Generator is is responsible for generating perturbations, discriminator is employed to differentiate between adversarial traffic and original malicious traffic. Additionally, intrusion detection model is utilized to discriminate between malicious traffic and normal traffic. The adversarial traffic that is closer to the real malicious traffic is generated through continuous training and can evade model detection. When adding perturbation, only nonfunctional features are selected for perturbation to preserve the functionality. In addition, IDS-GAN does not require a comprehensive understanding of the model's structure and parameters. The experiment conducted in this paper employs the NSL-KDD dataset to evaluate the obfuscation performance of the IDS-GAN method on two target models. The results demonstrate that this approach exhibits effective obfuscation performance for both Denial-of-Service (DoS) and Probe traffic.

National Natural Science Foundation of China No. 62302520

II. Related work

Rigaki et al. [4] applied the adversarial methods in image such as FGSM and JSMA to IDS, and generated adversarial traffic using NSL-KDD dataset. The results show that random forest is the most robust and its accuracy decreases by only about 5%. Wang et al. [6] added the performance evaluation of DeepFool and C&W adversarial methods in traffic obfuscation, and the results showed that the adversarial traffic generated based on C&W had poor obfuscation ability.

Sharon et al. [10] proposed a novel time-based adversarial attack TANTRA, which can bypass multiple IDS. TANTRA uses LSTM for training, which adjusts the timestamp of malicious traffic packets by learning the timing features of benign packets without changing the content of the packets. In this way, the timing features of malicious traffic packets are changed to avoid detection.

Different from most studies that directly obfuscated in the feature space of traffic, Han et al. [11] proposed a more practical packet-based obfuscation method, which realized evasion detection of intrusion detection models by adding virtual packets, changing the time stamp and randomizing the size of virtual packets.

Similar to this paper, many researchers use GAN-based adversarial traffic generation algorithm. Alhajjar et al. [12] used genetic algorithm, particle swarm optimization algorithm and GAN to generate adversarial traffic on NSL-KDD dataset. Despite the consideration of domain constraints related to traffic, there were still challenges in clearly defining the division of functional features. Additionally, perturbations in the traffic might potentially disrupt the original traffic functionality. Usama et al. [13] also proposed a GAN-based obfuscation method. They made a clear distinction between functional and non-functional features in NSL-KDD dataset and perturbed only the non-functional features, so as to retain the traffic functionality. However, this method only reduced the detection rate from 85% to 56%. Although it can evade the detection to a certain extent, further improvement is needed to improve the obfuscation performance.

III. Methodology

IDS-GAN is based on AdvGAN [14], and its architecture is shown in Figure 1. The objective is that adversarial traffic $X' = X + \delta$ can successfully evade IDS by adding perturbation δ to the original malicious traffic X while preserving the traffic functionality. IDS-GAN mainly consists of generators G, discriminators D and pre-trained models F. Generator G is responsible for generating perturbations, discriminator D is utilized to differentiate between adversarial traffic and original malicious traffic, and model F serves as an intrusion detection model to distinguish between malicious and normal traffic. Regarding the training process where the generator and discriminator compete with each other, the adversarial obfuscated traffic is generated which is closer to the real malicious traffic and can evade detection.

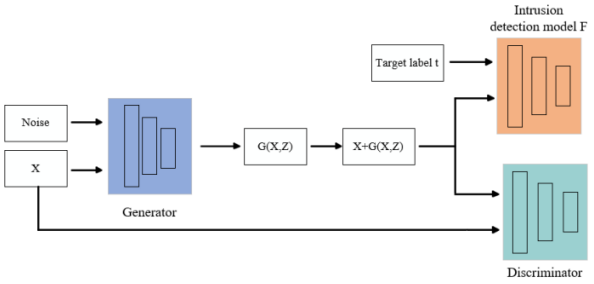

Figure 1. The architecture of IDS-GAN.

Generator G: We assume that X represents original traffic, δ represents perturbation and Z represents noise. The generator is to generate perturbation δ and add δ to X to create adversarial traffic. And add noise Z to X as input to generator. The generator has 5 fully connected layers, each of which contains 128 neurons. And the output and input dimensions remain consistent, that is, the dimension of perturbation δ is the same as that of the original traffic. But for perturbation vector δ, the dimension corresponding to functional features needs to be set to 0 to ensure that only nonfunctional features are perturbed to preserve the traffic functionality.

Discriminator D: The discriminator is responsible for differentiate between real malicious traffic and adversarial traffic. Through training, the discriminator promotes the generator to generate adversarial traffic that is closer to real malicious traffic. The discriminator also has 5 linear layers, each has 256 neurons. The discriminator plays a vital role in training the generator, providing feedback that enables the generator to compute the loss function, update parameters accordingly, and make the generated adversarial traffic close to the real malicious traffic.

Table 1. The architecture of intrusion detection model.

layer	type	neurons	activation	dropout
Input		122		
1	Linear	1024	ReLU	0.01
2	Linear	768	ReLU	0.01
3	Linear	512	ReLU	0.01
4	Linear	256	ReLU	0.01
5	Linear	128	ReLU	0.01
Output		1		

Intrusion detection model F: It is a pre-trained deep neural network model, and its purpose is to identify adversarial traffic and calculate the loss between the classification label and the target label of adversarial traffic to feed back to the generator. In this paper, we used DNN proposed by Vinayakumar [15] as the intrusion detection model. The architecture is shown in Table 1.

To generate adversarial traffic capable of evading detection, the corresponding parameter update is carried out by optimizing the loss function. In the training process, the loss function of IDS-GAN primarily comprises following three components:

L_F ensures that the generated adversarial obfuscation traffic can be classified by the intrusion detection model as the target label t, that is, normal traffic:

$$L_F = \mathbb{E}_{X \sim P_{malious}} \mathcal{L}[F(X + G(X,Z), t)] \quad (1)$$

L_P restricts the magnitude of the perturbation, ensuring that making slight modifications to original traffic can evade detection.:

$$L_P = \mathbb{E}_{X \sim P_{malious}} \|G(X,Z)\|_1 \quad (2)$$

L_D ensures that generated adversarial traffic is as close to the real malicious traffic as possible. The loss function is described in detail as follows:

$$L_D = \mathbb{E}_{X \sim P_{malious}}[D(X)] + \mathbb{E}_{X \sim P_{malious}}[1 - D(X + G(X,Z))] \quad (3)$$

In summary, the loss function of IDS-GAN is as follows:

$$\min_G \max_D \; L = L_F + \alpha L_D + \beta L_P \quad (4)$$

Through grid search, α and β values are 0.3 and 0.1.

IV. Evaluation

A. Experimental Setup

1) Dataset

This paper uses NSL-KDD dataset, which mainly includes Dos, Probe, U2R and R2L malicious traffic as well as normal traffic. Each traffic contains 41-dimensional feature vector. The features can be divided into intrinsc, content, time-based traffic and host-based traffic. Table 2 shows the functional features of traffic.

Table 2. Functional features of traffic.

Attack	Intrinsic	Content	Time-based traffic	Host-based traffic
DoS	√		√	
Probe	√		√	√
U2R	√	√		
R2L	√	√		

NSL-KDD contains nonnumerical features. Since the model only accept numerical features as input, we should convert nonnumerical features to numerical features normally by using one hot encoding. For instance, consider the "protocol_type" attribute, which primarily consists of three protocol types: TCP, UDP, and ICMP. After one hot encoding, it can be represented as 100,010,001. NSK-KDD originally has 41 features, and after encoding, the feature vector expands to 122 dimensions. In order to eliminate the scale difference among features and improve performance of classifier, the Min-Max normalization is employed to map the feature values to the range of [0,1].

2) Evaluation Metrics

To assess the effectiveness of IDS-GAN, two evaluation metrics, detection rate and evasion increase rate, are adopted, and each metric is defined as follows.

Detection Rate (DR) : The detection rate refers to the ratio of malicious traffic correctly identified by the model among all instances of malicious traffic. The Original Detection Rate (ODR) is the DR of the model against the original traffic, and the Adversarial Detection Rate (ADR) is the DR of the model against adversarial traffic. It demonstrates the ability of adversarial obfuscation traffic evasion detection and the robustness of detection model.

$$DR = \frac{T_{correct}}{T_{total}} \quad (5)$$

$T_{correct}$ indicates the number of malicious traffic correctly detected by the model, and T_{total} indicates the total number of all malicious traffic.

Evasion Increase Rate (EIR): The evasion increase rate is defined as the increasing rate of undetected malicious traffic in adversarial traffic relative to undetected malicious traffic in the original malicious traffic.

$$EIR = 1 - \frac{ADR}{ODR} \quad (6)$$

The goal of IDS-GAN is to achieve lower ADR and higher EIR.

3) Target Models

The experiment included two target models. One was the DeepNet model proposed by Gao et al. [16] , which had 4 linear layers with 256 neurons each. The other is the model built in this paper, IDS-MLP, which has five linear layers with 244 neurons each. This paper only considers binary classification situation, malicious traffic is labeled as '1' and normal traffic is labeled as '0', without further distinguishing which attack type belongs to. The ODR of the model is shown in Table 3.

Table 3. The detection rate of the target model.

Model	Attack	ODR
IDS-MLP	Dos	84.32%
	Probe	81.31%
	R2L、U2R	15.09%
DeepNet	Dos	83.01%
	Probe	81.92%
	R2L、U2R	14.15%

B. Experimental Results

1) Overall obfuscation performance evaluation

Table 4. The adversarial detection rate and evasion increase rate of the target model.

Model	Attack	ADR	EIR
IDS-MLP	Dos	26.58%	68.48%
	Probe	21.04%	74.12%
	R2L、U2R	14.16%	6.16%
DeepNet	Dos	21.81%	73.73%
	Probe	18.92%	76.91%
	R2L、U2R	12.47%	11.87%

The performance of IDS-GAN against IDS-MLP and DeepNet was evaluated on NSL-KDD dataset, as shown in Table 4. It indicates that IDS-GAN has better performance on Dos and Probe traffic, while the poor performance of U2R and R2L may be due to the small amount, so IDS-GAN cannot be fully trained. The adversarial detection rate of both IDS-MLP and DeepNet for Dos and Probe is lower than 26.58%, which is greatly reduced compared with the original detection rate, indicating that adversarial traffic created by IDS-GAN

demonstrates strong evasion capabilities against model detection. Furthermore, the evasion increase rate of the two target models for Dos and Probe is higher than 68.48%, indicating that the growth rate of the amount of traffic evading detection is high. At the same time, through the comparison of adversarial detection rate and evasion increase rate of IDS-MLP and DeepNet, it can be further found that the adversarial traffic detection rate of IDS-MLP is higher, and the evasion increase rate is lower, indicating that IDS-MLP has stronger robustness facing adversarial traffic.

2) Obfuscation performance with the reduction of perturbed features

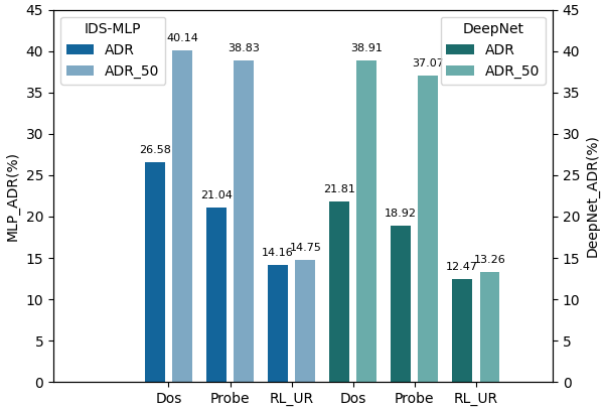

Figure 2. Comparison of adversarial detection rates for modifying 50% of non-functional features vs modifying all non-functional features.

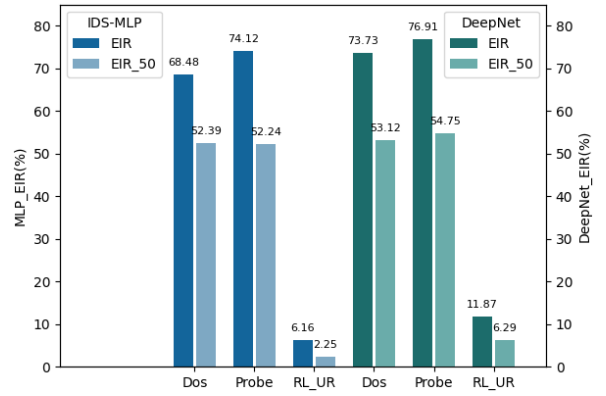

Figure 3. The evasion increase rate of modifying 50% of non-functional features vs modifying all non-functional features.

This experiment primarily investigates how varying the quantity of perturbed features influences the obfuscation performance. In order to maintain the traffic functionality, 50% of non-functional features are randomly selected and modified. The degree of change of ADR and EIR with the number of modified features is shown in Figure2 and Figure3. The results show that as the number of modified features decreases, EIR decreases and the adversarial detection rate increases, such as: When the number of modified features was 50% of the original non-functional features, the adversarial detection rate of DeepNet on Probe increased from 18.92% to 37.07%, and the evasion increase rate decreased from 76.91% to 54.75%. Because with the increase of the number of unmodified features, The adversarial traffic retains more feature information of the original traffic, which is easier to be correctly identified by intrusion detection models, resulting in the increase of ADRe and the decrease of EIR.

3) Obfuscation performance of white box scenarios

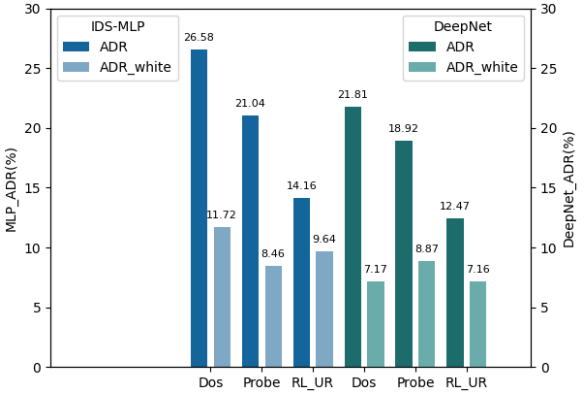

Figure 4. Comparison of adversarial detection rates in black box and white box scenario.

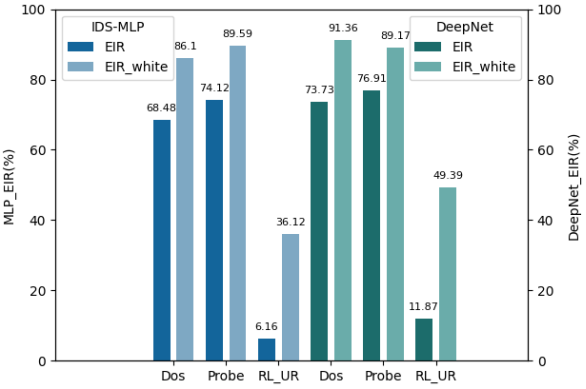

Figure 5. Comparison of evasion increase rate in black box and white box scenario.

The experiment further evaluated obfuscation performance of IDS-GAN in the white-box scenario. In this scenario, it is assumed that the obfuscator has access to the target model's structure and the internal parameters such as the loss function. During the training of IDS-GAN, IDS-MLP and DeepNet are respectively used as model F. The obfuscation performance of adversarial traffic generated by IDS-GAN against IDS-MLP and DeepNet was evaluated, as shown in Figure 4 and Figure 5. It demonstrates that in comparison to the black box scenario, the adversarial detection rate is lower in the white box scenario, and the evasion increase rate is higher. For example, compared with the black box scenario, ADR of IDS-MLP against Dos decreases from 26.58% to 11.72%, and EIR increases from 68.48% to 86.10%. This is because in the white box scenario, the obfuscator possesses a comprehensive understanding of the model's structure and parameters, so the adversarial traffic can be designed specifically to interfere with the model's prediction results to the greatest extent.

V. Conclusions

In this paper, we propose an adversarial obfuscation method for intrusion detection based on generative adversarial network named IDS-GAN, which only perturbs non-functional features to preserve traffic functionality while evading detection. The method is mainly composed of generator, discriminator and pre-trained model. Generator is used to generate perturbation, discriminator is used to distinguish generated adversarial traffic from original malicious traffic, intrusion detection model is employed to differentiate between malicious traffic and normal traffic. The adversarial traffic that is closer to the real malicious traffic is generated through continuous training and can evade model detection. The experiment uses NSL-KDD dataset and evaluates the obfuscation performance of IDS-GAN on two target models. The results indicate that the approach exhibits the most effective obfuscation performance against DoS and Probe attacks, and can reduce ADR of the model to less than 27%, and increase EIR to more than 68%.

Acknowledgment

This work was supported by the National Natural Science Foundation of China No. 62302520.

References

[1] H.J. Liao, C.H. R. Lin, Y.C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2013. https://doi.org/10.1016/j.jnca.2012.09.004

[2] C. Szegedy et al., "Intriguing properties of neural networks," in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. http://arxiv.org/abs/1312.6199

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 ieee symposium on security and privacy (sp), Ieee, 2017, pp. 39–57. https://doi.org/10.1109/SP.2017.49

[4] A. Chernikova and A. Oprea, "Fence: Feasible evasion attacks on neural networks in constrained environments," ACM Trans. Priv. Secur., vol. 25, no. 4, pp. 1–34, 2022. https://doi.org/10.1109/SP.2017.49

[5] M. J. Hashemi, G. Cusack, and E. Keller, "Towards evaluation of nidss in adversarial setting," in Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks, 2019, pp. 14–21. https://doi.org/10.1145/3359992.3366642

[6] Z. Wang, "Deep learning-based intrusion detection with adversaries," IEEE Access, vol. 6, pp. 38367–38384, 2018. https://doi.org/10.1109/ACCESS.2018.2854599

[7] M. Rigaki and A. Elragal, "Adversarial deep learning against intrusion detection classifiers," in 2017 NATO IST-152 Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience, IST-152 2017; Czech Technical UniversityPrague; Czech Republic; 18-20 October 2017, CEUR-WS,2017, pp.35–48. https://doi.org/10.1109/GLOBECOM38437.2019.9014337

[8] P. Papadopoulos, O. Thornewill von Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, and W. J. Buchanan, "Launching adversarial attacks against network intrusion detection systems for iot," J. Cybersecurity Priv., vol. 1, no. 2, pp. 252–273, 2021. https://doi.org/10.3390/jcp1020014

[9] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks," in 2019 IEEE global communications conference (GLOBECOM), IEEE, 2019, pp. 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9014337

[10] Y. Sharon, D. Berend, Y. Liu, A. Shabtai, and Y. Elovici, "Tantra: Timing-based adversarial network traffic reshaping attack," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 3225–3237, 2022. https://doi.org/10.1109/TIFS.2022.3201377

[11] D. Han et al., "Practical traffic-space adversarial attacks on learning-based nidss," ArXiv Prepr. ArXiv200507519, 2020.

[12] E. Alhajjar, P. Maxwell, and N. Bastian, "Adversarial machine learning in network intrusion detection systems," Expert Syst. Appl., vol. 186, pp. 115782, 2021. https://doi.org/10.1016/j.eswa.2021.115782

[13] M. Usama, M. Asim, S. Latif, and J. Qadir, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in 2019 15th international wireless communications & mobile computing conference (IWCMC), IEEE, 2019, pp. 78–83. https://doi.org/10.1109/IWCMC.2019.8766353

[14] C. Xiao, B. Li, J. Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," in 27th International Joint Conference on Artificial Intelligence, IJCAI 2018, International Joint Conferences on Artificial Intelligence, 2018, pp. 3905–3911. https://doi.org/10.24963/ijcai.2018/543

[15] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," Ieee Access, vol. 7, pp. 41525–41550, 2019. https://doi.org/10.1109/ACCESS.2019.2895334

[16] M. Gao, L. Ma, H. Liu, Z. Zhang, Z. Ning, and J. Xu, "Malicious network traffic detection based on deep neural networks and association analysis," Sensors, vol. 20, no. 5, p. 1452, 2020. https://doi.org/10.3390/s20051452