Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 187 (2021) 128–133

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Identification, Information and Knowledge in the internet of Things, 2020

attackGAN: Adversarial Attack against Black-box IDS using Generative Adversarial Networks

Shuang Zhao[a], Jing Li[b], Jianmin Wang[c], Zhao Zhang[a], Lin Zhu[d], Yong Zhang[a,*]

[a]*School of Electronic Engineering, Beijing University of Posts and Telecommunication, Beijing,*
Email:{zhaoshuang, guoda, zhaozhang, yongzhang}@bupt.edu.cn, China
[b]*Zhejiang Huayun Electric Power Industrial Group co,ltd, Zhejiang,*
Email: lijing_huayun@sina.com, China
[c]*Zhejiang Huayun Information Technology co,ltd, Zhejiang, Email: wangjianmin@hyit.com.cn, China*
[d]*China Mobile Research Institute, Beijing, Email: zhulinyj@chinamobile.com, China*

Abstract

With the rapid development of Internet of Things technology, a large number of devices are connected to the Internet of Things, and at the same time, a large number of network attacks and security threats are introduced. Intrusion detection system (IDS) is one of the effective methods for protecting network. With the rise of artificial intelligence technology, intrusion detection system based on ML/DL is widely applied. However, neural network is vulnerable to adversarial perturbation. Most of existing adversarial attacks cannot guarantee the basic function of traffic data. In this paper, we propose an improved adversarial attack model based on Generated Adversarial Network called attackGAN, and design a new loss function to achieve effective attack against the black-box intrusion detection system on the premise of ensuring network traffic functionality. Experiments show that the proposed attackGAN can improve the success rate of adversarial attack against the black-box IDS compared with Fast Gradient Sign Method (FGSM), Project Gradient Descent (PGD), CW attack (CW) and the GAN-based algorithms.

Keywords: Generated Adversarial Network; adversarial attack; intrusion detection system

1. Introduction

The rapid development of Internet of Things technology has brought about the explosive growth of user resources and the number of connected devices, but it has raised security concerns. With the rise of artificial intelligence technology, intelligent algorithms represented by machine learning and deep learning have been applied to network intrusion

[*] Corresponding author: Yong Zhang
E-mail address: yongzhang@bupt.edu.cn

10.1016/j.procs.2021.04.118

detection systems. Deep learning technology brings a leap-forward development opportunity for protecting network, but it also provides a new attack surface for the attackers. The emergence of malicious adversarial attacks poses a huge challenge to the security and reliability of network intrusion detection systems. Recent studies have shown that the neural network is vulnerable to adversarial perturbation. By adding a small amount of disturbance to the original samples, the generated adversarial samples make the intrusion detection system misjudge [13].

With the rapid development of deep learning algorithm, it also faces more and more challenges from adversarial attacks. Since Szegedy [11] proposed that neural network is vulnerable to adversarial attacks, the researchers constantly proposed new adversarial attack methods. In the field of network security, some attack algorithms are also proposed. Grosse et al. [4] proposed an adversarial perturbation of malware classifiers based on Deep Neural Networks (DNNs). They used Fast Gradient Signal Method (FGSM) and significant mapping attacks based on jacobian matrices to create adversarial malware instances. In some cases, the attacker cannot access the architecture and weights of the neural network to be attacked, the target model was the black-box for the attacker.

At present, a small number of existing achievements have been made in attacking intrusion detection systems with Generative Adversarial Network (GAN) [2]. In GAN, the discriminant model is used to distinguish the generated sample from the real sample, and the generated model is trained to make the discriminant model misclassify the generated sample into the real sample. MalGAN [5] was proposed to circumvent the detection system of malware using generation modeling technology. The MalGAN model used the forward neural network as the generator, the substitute detector as the discriminator, and the random noise as the input to generate malicious samples. However, how to make the substitute detector learn the inner details of the black-box model to the maximum extent remains to be further studied. IDSGAN [7] was proposed to generate adversarial attacks against intrusion detection systems, which was based on Wasserstein GAN [9] and included the generator, the discriminator and the black-box IDS. The discriminator was used to simulate the black-box intrusion detection system, and the generator can generate a sample of malicious traffic. Although they claimed to retain the functional characteristics of the traffic data, they actually changed the two functional characteristics to invalidate the generated traffic data.

Muhammad et al. [12] proposed and verified an adversarial machine learning attack based on GAN against the black-box intrusion detection system. The GAN-based algorithm was the first adversarial attack that can successfully evade the intrusion detection system and ensure the functionality of traffic data simultaneously. Although the accuracy of their experimental results is reduced, the degree of reduction is limited, and the attack success rate needs to be further improved.

In view of adversarial attack against network intrusion detection system, the traffic data belongs to the discrete data. How to produce imperceptible and effective adversarial samples should be further studied under the premise of ensuring its functionality. In general, the principle of GAN is the adversarial samples and the real samples are input into the discriminator. Feedback based on the results of discriminator, the generator keeps training and adjusting until the discriminator cannot distinguish between the real samples and the adversarial samples. The main contributions of this paper are shown as follows:

- We proposed an improved adversarial attack model based on Wasserstein GAN called attackGAN. By adding the feedback of the intrusion detection system, it can effectively realize evade attack, and at the same time guarantee the functionality of network traffic.
- Combined with the model structure, we designed the corresponding loss function using the result of the intrusion detection system as a constraint, which improved the quality of the adversarial samples.
- When the black-box intrusion detection system is assumed to be ML/DL algorithms, compared with the existing GAN-based adversarial attack algorithms [12], Fast Gradient Sign Method (FGSM) [3], Project Gradient Descent (PGD) [8], CW attack (CW) [1], the proposed attackGAN can achieve a higher attack success rate.

2. ATTACKGAN MODEL

The proposed attackGAN is based on Wasserstein GAN. Generative adversarial networks based on game theory, by adding a small number of subtle disturbance to the original traffic sample, the attacker tries to trick the discriminator into believing that the sample is real. The discriminator tries to distinguish between the sample extracted from the original data and the adversarial sample generated by the generator. In the case of gradual convergence, the adversarial

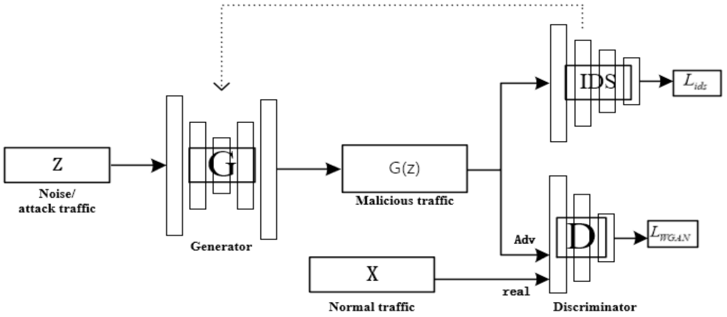

Fig. 1. The architecture of attackGAN

sample is as similar as possible to the original sample. In order to obtain the appropriate adversarial samples, two constraints should be satisfied: one is to maintain the function of the traffic samples, and the other is to be aggressive, so that the adversarial samples will not be detected by the intrusion detection system. Fig. 1 illustrates the overall architecture of attackGAN, which mainly consists of three parts: the generator G, the discriminator D, and the intrusion detection system IDS.

Firstly, the input of the generator G is the noise sample or the attack sample Z, and the output is the adversarial sample $G(z)$. The adversarial sample $G(z)$ is fed into the discriminator D, which is used to distinguish the generated adversarial sample $G(z)$ from the normal traffic sample X. The goal of D is to encourage that the generated sample is indistinguishable with the sample from its normal class. The loss function is L_{WGAN}, which represents the difference between the predicted label and the actual label.

$$L_{WGAN} = E_{x\sim p_r}[D(x)] + E_{z\sim p_g}[1 - D(z)] \tag{1}$$

where p_r is the distribution of the normal sample, p_g is the distribution of the generated sample.

Secondly, in order to achieve the goal of evade attack, the black-box attack is considered. The input of the black-box intrusion detection system IDS is the adversarial sample $G(z)$. The output result of IDS is fed back to the generator G to help generate more effective adversarial attack samples. The goal of G is that the discrimination results of the generative adversarial samples are the normal samples. The loss function is L_{ids}, which represents the difference between the output detection result and target label t_{adv}.

$$L_{ids} = E_{z\sim p_g} l_f[IDS(z), t_{adv}] \tag{2}$$

where t_{adv} represents the target label and l_f represents the cross entropy function.

L_{WGAN} is used to encourage the adversarial samples to be similar to the original samples X, and L_{ids} is used to generate more effective adversarial samples. Finally, by jointly optimizing G and D, the generator and the discriminator are obtained by solving the maximum-minimum game, and then the black-box attack is achieved. The overall objective function is as follows:

$$\underset{G}{min}\underset{D}{max} L = L_{WGAN} + \lambda L_{ids} \tag{3}$$

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

3. Experiment and evaluation

We evaluated the effectiveness of attackGAN using the NSL-KDD dataset [10]. First, the dataset and experimental configuration used in the experiment are introduced. Then, an unified evaluation index is proposed to adapt to different attack algorithms. Finally, the performance of various algorithms using the same dataset is comprehensively compared.

3.1. Dataset

NSL-KDD dataset is a new version of KDD Cup 99 dataset. Each network connection is marked as normal or abnormal. Anomaly types are subdivided into 4 categories with a total of 37 attack types, of which attack types are divided into five categories: Denial of Service (DoS), the User to Root (U2R), Root to Local (R2L), Probing (Probe) and normal.

For each record, the NSL-KDD training dataset contains 41 fixed feature attributes and class identifiers. The extraction of features from the original traffic datas has been widely explained in [10], where they designed a four-layer feature extraction scheme. Among them, 41 features can be divided into intrinsic, content, time-based traffic, and host-based traffic.

In the study of adversarial attack theory, a very important constraint is to maintain the functional behavior of adversarial samples. Different attack types correspond to different functional features, which are the basic features of attack implementation. In the field of network intrusion detection, the change of some functional characteristics of the traffic sample results in the failure of the attack. Therefore, the premise of attackGAN to resist the attack is that the functional features of the attack category do not changed, we only change the non-functional features. In [6], the functional features of each type of attack in NSL-KDD dataset are given. DOS attack: Intrinsic, Time-based; U2R attack: Intrinsic, Content; R2L attack: Intrinsic, Content; PROBE attack: Intrinsic, Time-based, Host-based.

In data preprocessing, the 41-dimensional features of NSL-KDD dataset need to be processed, mainly including two aspects: The first is to convert the input samples of attakGAN into numeric vectors, and convert the three non-numeric data into numeric data, including protocol type, service and flag. The second is normalization.

Algorithm 1 attackGAN

```
Require: The normal example X; The noise for generate adversarial example Z; The trained neurak network IDS;
Ensure: The trained generator $G$;
Initialize the generator $G$ and the discriminator $D$
for number of training iterations do
    for k-steps do
        sample $x^{(i)}$ minibatch of $m$ normal examples X ; sample $z^{(i)}$ minibatch of $m$ noise examples Z
        update the discriminator $D$ descent its gradient
            $\nabla_{\theta_d} - \frac{1}{m} \sum_{i=1}^{m} \left[ D(x^{(i)}) - D(G(z^{(i)})) \right]$   (4)
        clipping threshold of the discriminator is set to [-c,c]
    end for
    sample $z^{(i)}$ minibatch of $m$ noise examples Z
    update the generator $G$ descent its gradient
        $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \left[ -D(G(z^{(i)})) + \lambda \text{ł}_f [IDS\,(z^{(i)}), t_{adv} \right]$   (5)
end for
```

3.2. Experimental configuration

In the experiment, Pytorch is used as the deep learning framework to construct attackGAN to prove the effectiveness of the model we proposed. All the experiments are completed on the NSL-KDD dataset. Our experimental environment is: CPU: Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, OS: Ubuntu 16.04

During the training of attackGAN model, probe attack is taken as an example. In order to retain the functionality of the traffic sample, only its content feature is changed, and the dimension of the noise vector is 13. The number of neural network layers of generator and discriminator is two, the number of neurons in each layer is 128. The learning rate is 0.0001, and the optimizer used is RMSprop. The weight clipping threshold of the discriminator is set to 0.01. The deep neural network trained by the training set is used as the intrusion detection system, with 4 layers and 8 neurons in each layer, which is saved as the *IDS* model. Moreover, we select five ML/DL algorithms, including

Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Deep Neural Network (DNN), which are trained as the black-box intrusion detection systems to synthesize comparative analysis.

3.3. Evaluation indicators

The trained intrusion detection system predicts a label for each attack sample. If the predicted label is normal, it will be an effective attack. In order to compare the attackGAN with other algorithms, such as Fast Gradient Sign Method (FGSM), Project Gradient Descent (PGD), CW attack (CW), the performance of attackGAN is intuitively reflected, and three evaluation indexes of detection accuracy, attack success rate and evade increase rate are defined.

The detection accuracy (D_R) reflects the proportion between the attack samples correctly detected and all the actual attack samples. The formula is Eq. (6).

$$D_R = \frac{A_{de}}{A_{all}} \tag{6}$$

where A_{all} is the number of all the attack samples, A_{de} is the number of correctly detected attack samples.

Besides, in order to reflect the performance of adversarial attack algorithms under different intrusion detection systems, the success rate of attack (A_{SR}) is defined. The formula is Eq. (7).

In addition, the evade increase rate (E_{IR}) is also defined to compare the performance of the proposed attackGAN with different algorithms. The formula is Eq. (8).

$$A_{SR} = D_{RO} - D_{RA} \tag{7}$$

$$E_{IR} = 1 - \frac{D_{RA}}{D_{RO}} \tag{8}$$

where D_{RO} is the detection rate of the original malicious traffic samples, D_{RA} is the detection rate of the generated adversarial samples.

3.4. Experimental results

In the experiment, the attack performance of attackGAN and four attack algorithms GAN-based, FGSM, PGD, CW when the black-box IDS are five different ML/DL algorithms are also compared. The experiment results are shown in Fig. 2 and Fig. 3.

It can be seen from Fig. 2: because the black-box IDS assumes that different machine learning or deep learning algorithms have different training effects, the original detection accuracy is different, so the detection accuracy cannot be directly compared. When the black-box IDS is assumed as SVM, the attack success rate of attackGAN is 81.37% higher than that of the GAN-based attack algorithm is 40.87%. The higher the values of A_{SR}, the higher the aggressiveness of the generated adversarial samples, and the better the performance of the corresponding adversarial attack algorithm. On the one hand, the proposed attackGAN generates adversarial samples that interfere with the black-box IDS detection and increase the attack success rate. On the other hand, compared with the three adversarial attack algorithms, attackGAN has the highest attack success rate, which proves that the attackGAN algorithm has superior attack performance.

It can be seen from Fig. 3: to ensure the functionality of the traffic sample is the premise to resist attacks against the IDS. When the black-box IDS is assumed to be SVM, DT, RF, NB, and DNN, the changes in detection accuracy are not fully measured only by the success rate of attack, so we have to compare the evade increase rate. When the black-box IDS is assumed as NB, the evade increase rate of attackGAN is 87.18% higher than that of the GAN-based attack algorithm is 18.59%, FGSM (17.76%), PGD (29.78%) and CW (21.25%). Moreover, the values of E_{IR} are higher than those of the comparison algorithm, which proves that the generated adversarial samples of attackGAN are more effective.

4. Conclusion

In this paper, we propose attackGAN for adversarial attack against the black-box intrusion detection system. On the premise of ensuring the functionality of the traffic samples, the adversarial samples generated by attackGAN can

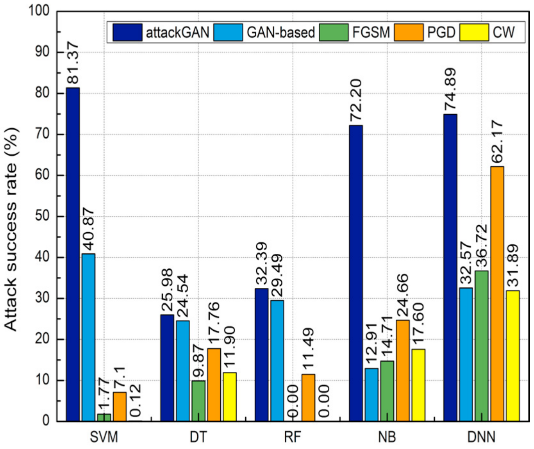

Fig. 2. The comparison of Attack success rate

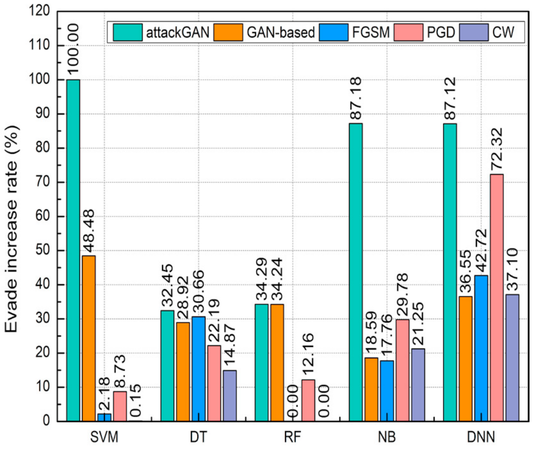

Fig. 3. The comparison of Evade increase rate

successfully evade the detection of the IDS. Experiment results show that the proposed attackGAN has a higher the success rate of attack than existing adversarial attack algorithms. In the future, we need to use adversarial attack and defense game to drive the future development of network security technology.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No.61971057), MoE-CMCC ”Artifical Intelligence” Project (No. MCM20190701) and the State Major Science and Technology Special Projects (No. 2014ZX03004002).

References

[1] Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), IEEE, San Jose, USA. pp. 39–57.

[2] Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 .

[3] Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples, in: 3nd International Conference on Learning Representations, ICLR 2015, San Diego, USA.

[4] Grosse, K., Papernot, N., Manoharan, P., Backes, M., Mcdaniel, P., 2016. Adversarial perturbations against deep neural networks for malware classification. arXiv preprint arXiv:1606.14435v2 .

[5] Hu, W., Tan, Y., 2017. Generating adversarial malware examples for black-box attacks based on gan. arXiv preprint arXiv:1702.05983 .

[6] Lee, W., Stolfo, S.J., 2000. A framework for constructing features and models for intrusion detection systems. ACM transactions on Information and system security (TiSSEC) 3, 227–261.

[7] Lin, Z., Shi, Y., Xue, Z., 2018. Idsgan: Generative adversarial networks for attack generation against intrusion detection. arXiv preprint arXiv:1809.02077 .

[8] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, Canada.

[9] Martin Arjovsky, S., Bottou, L., . Wasserstein generative adversarial networks, in: Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia.

[10] Revathi, S., Malathi, A., 2013. A detailed analysis on nsl-kdd dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research & Technology (IJERT) 2, 1848–1853.

[11] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2014. Intriguing properties of neural networks, in: 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada.

[12] Usama, M., Asim, M., Latif, S., Qadir, J., et al., 2019. Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems, in: 15th International Wireless Communications & Mobile Computing Conference (IWCMC), IEEE, Tangier, Morocco. pp. 78–83.

[13] Xiao, C., Zhu, J., Li, B., He, W., Liu, M., Song, D., 2018. Spatially transformed adversarial examples, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, Canada.