Proceedings of the 7th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'21)
~~Prague, Czech Republic~~ Virtual Conference - July, 2021
Paper No. CIST 121
DOI: 10.11159/cist21.121

Detection of Adversarial DDoS Attacks Using Generative Adversarial Networks with Dual Discriminators

Chin-Shiuh Shieh [1], Wan-Wei Lin [1*], Thanh-Tuan Nguyen [1,3], Yong-Lin Huang [1], Mong-Fong Horng [1], Chun-Chih Lo [1], and Kun-Mu Tu [2]

[1] Department of Electronic Engineering
National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan, ROC
[2] EverGuard Technology Co. Ltd., Taiwan, ROC
[3] Department of Electronic and Automation Engineering, Nha Trang University, Vietnam.
[*] i109152103@nkust.edu.tw

Abstract - DDoS (Distributed Denial of Service) has become a pressing and challenging threat to the security and integrity of computer networks and information systems. The detection of DDoS attacks is essential before any mitigation approaches can be taken. AI (Artificial Intelligence) and ML (Machine Learning) have been applied to the detection of DDoS attacks with satisfactory achievement. However, new types of attacks emerge as the technology for DDoS attacks keep evolving. This study investigates the impact of a new sort of DDoS attack – adversarial DDoS attack. We synthesize attacking traffic using Wasserstein Generative Adversarial Networks with Gradient Penalty (GP-WGAN). Experiment results reveal that the synthesized traffic can penetrate the systems, including Random Forest, k-Nearest Neighbor, and Multi-Layer Perceptron, without being detected. This observation is an alarming and pessimistic wake-up call implying the urgent need for countermeasures to adversarial DDoS attacks. To this problem, we propose an adversarial GAN Intrusion Detection System (AG-IDS) featuring dual discriminators. The additional discriminator is designed to capture adversarial DDoS traffic. As indicated in the experimental results, the proposed AG-IDS can be an effective solution to adversarial DDoS attacks.

***Keywords*:** DDoS, Machine Learning, Generative Adversarial Network, Intrusion Detection System

1. Introduction

By flooding malicious traffic, DoS (Denial of Service) attacks deplete the network bandwidth and computing resources of a targeted system, preventing the target system from offering regular services to legitimate users. DDoS goes even further on a much larger scale. DDoS attacks take over the control of a large number of comprised systems, called a botnet, and launch coordinated attacks on the victim system, as illustrated in Figure 1. Along with the emergence and advancement of disruptive Internet technologies, DDoS attacks are evolving and proliferating in scale, frequency, and sophistication. Organizations face potential threats to their network environment that may cause severe impacts to their operations, such as business downtime, data breach, or even ransom demands from hackers [1].

Upon the occurrence of DDoS attacks, actions for DDoS mitigation should be taken, as suggested in [2]. The detection of DDoS attacks is essential before any mitigation approaches can be taken. In the early era, the alarm of DDoS attacks is triggered by rules programmed by traffic engineers. This approach apparently failed to catch up with the dynamic and evolving natures of DDoS attacks. As AI (Artificial Intelligence) and ML (Machine Learning) unleash their great potential in different fields, the academy and industry also explore the feasibility of applying ML to DDoS detection. Certain successes have been achieved, as reported in [3]. With ML, features for classification must be selected by human experts or by certain feature selection schemes. On the other hand, feature selection is an integral part of DL (Deep Learning). Some successful stories on DL for DDoS detection will be reviewed in Section II.

Both ML and DL demand labeled traffic as training data as applied to DDoS detection. The quality of the training set decides the performance of a DDoS detection system. A new type of DDoS attack, named adversarial DDoS attack, could bring about danger to the approaches mentioned above. GAN (Generative Adversarial Network) [11] is well recognized in the generation of fake but real-looking data, such as image synthesis. We believe that GAN is also capable of the generation of malicious but legitimate-looking traffic and therefore confuses the DDoS detection systems. We synthesize attacking

traffic using Wasserstein Generative Adversarial Networks [13] with Gradient Penalty (GP-WGAN) [14]. As we shall see in Section IV, the synthesized traffic can penetrate DDoS detection systems, including Random Forest, *k*-Nearest Neighbor, and Multi-Layer Perceptron, without being detected. To deal with this deficiency, we propose an adversarial GAN Intrusion Detection System (AG-IDS) equipped with dual discriminators. The additional discriminator is designed to capture adversarial DDoS traffic. As revealed in the experimental results, the proposed AG-IDS can be an effective solution to adversarial DDoS attacks.

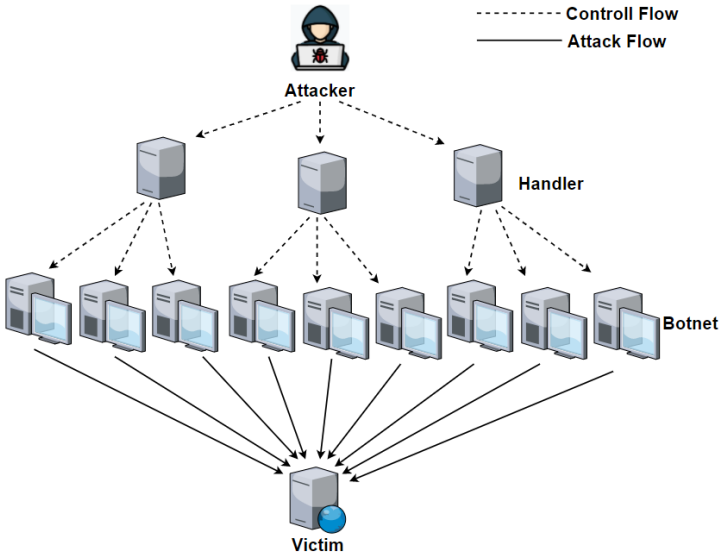

Figure 1. DDoS attack with a botnet.

The rest of this paper is organized as follows. Previous works and GANs are briefly reviewed in Section II. The framework of the proposed approach is presented in Section III. Then, experiment results are reported in Section IV. Finally, some conclusions are drawn in Section V.

2. Related Works

2.1. ML and DL for DDoS Detection

Various ML technologies have been employed, mainly as classifiers, in the detection of DDoS attacks. There are Support Vector Machine (SVM) [4], *k*-Nearest Neighbors (KNN) [5], Naïve Bayes Classifier [6], Random Forest (RF) [7], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [8], and Artificial Neural Network (ANN) [17], to name a few. With SVM, based on labeled training data, a hyperplane is constructed in the transform domain to classify unseen data. In KNN, *k* nearest neighbors of incoming data are located. A majority of these k neighbors decide the classification of the incoming data. Naïve Bayes classifier is a classification technique based on Bayes' theorem assuming independence among predictors. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. A RF is a collection of decision trees. The majority of the outcomes of individual decision trees determine the classification. DBSCN finds core samples of high density and expands clusters from them. It well suits data that contains clusters of similar density. ANNs emulate biological neural networks. Given label data, ANNs learn the mapping function using the back-propagation algorithm.

There are also successful stories for DL in DDoS detection. Yan Li [9] combines Long Short-Term Memory (LSTM) and Bayesian methods to detect DDoS attacks. LSTM is suitable for events with long intervals and delays in the time domain. In other words, LSTM can control the value of the indefinite length of time and decide whether the information should be retained or removed. The author uses LSTM to identify the confidence index of DDoS attacks and further uses the Bayesian

method to make a second judgment to improve detection accuracy. K. Yang et al. [10] adopt the autoencoder for the detection of DDoS attacks. Autoencoder is a multi-layer neural network with an unsupervised training algorithm. It removes less relevant information and noise during the training process and retains essential information. It is, in effect, some sort of feature selection. The training and operation speeds can be significantly improved.

2.2. Generative Adversarial Networks

GAN, proposed by Ian J. Goodfellow et al. in 2014 [11], demonstrates a wide variety of applicability in recent years. A GAN consists of two functional blocks, namely generator and discriminator, as shown in Figure 2. These two blocks act as parties in the game theory and compete with each other. The generator aims at the generation of artificial data to fool the discriminator. Conversely, the discriminator is responsible for the judgment of authenticity. As a GAN converges after a long time of training, the generator is supposed to able to generate artificial data indistinguishable from the real one.

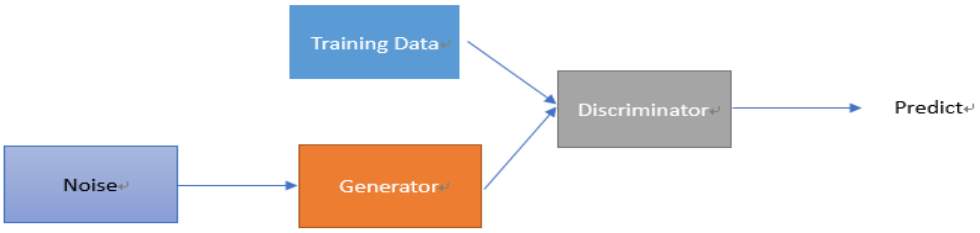

Figure 2. Functional block diagram of a generative adversarial network.

Despite its success in different fields, the training and convergence of GANs still pose significant challenges to practitioners. The original GAN is prone to convergence issues, such as loss functions failing to provide training direction, lacking diversity in generated data, and so on. Wasserstein GAN (WGAN) [12] introduces the Wasserstein distance to solve the original GAN's gradient vanishing problem. The WGAN is proved to more stable in the training process, with fewer crashing cases.

Although WGAN does improve the original GAN, it still suffers from the problem of inferior generated data and occasional divergence. These difficulties are the results of the Weight Clipping strategy adopted by WGAN. The Weight Clipping strategy is used to forcibly satisfy the Lipschitz constraint, leading the weights of WGAN to converge to extreme values and therefore degrades the generation efficiency or even model collapsion. To this problem, GP-WGAN proposes the Gradient Penalty strategy [13]. After using the Gradient Penalty strategy, the gradient learned by GP-WGAN is uniformly distributed. The Gradient Penalty strategy makes WGAN training more stable, and a higher-quality data generation can be achieved.

2.3. Adversarial DDoS Attacks

Hankers and researchers also envision GAN's potential in bringing up a new form for DDoS attack – adversarial DDoS attacks. In other words, GANs can be used to generate malicious but legitimate-looking traffic. In CYBERSEC-2019, Trend Micro Inc. reported cases of adversarial DDoS attacks. Hankers exploit Adversarial Machine Learning (AML) technology to find ML-based detection systems' weaknesses and deceive them with adversarial traffic. Studies are devoting to the investigation of adversarial DDoS attacks. Based on the observation that DDoS attacks highly resemble regular flash crowds, Least Square GAN (LSGAN) was proposed in [14] for the generation of artificial traffic. With LSGAN, up to 99% of generated DDoS traffic was incorrectly classified as legitimate flash crowds. An interesting architecture, named MalGAN, is presented in [15] for adversarial DDoS attacks. There are two neural networks and a black box detector. The black box detector plays the role of a victim system. Instances generated by the generator are fed to the black box detector for classification. The results of the classification are then directed to the discriminator. The system converges when the discriminator can no longer distinguish an adversarial attack from benign.

2.4.Detection of Adversarial DDoS Attacks

For the time being, the detection of adversarial DDoS attacks is still in infancy. Adversarial DDoS attacks deceive ML-based detection systems with GAN-generated traffic. An Adversarial Detection Module (ADM) in the loop can be an effective solution to this problem. ADMs capture and log adversarial DDoS attacks. Logged labeled traffic is then used to retrain the detection system incrementally. The problem in previous approaches is that the role of ADM is played by human data engineers. It renders discriminating and tagging a time-consuming and manpower-demanding process. The main contribution of this study is that we use an additional discriminator to automatize the process for the capturing and tagging of adversarial DDoS traffic.

3. Adversarial GAN Intrusion Detection System (AG-IDS)

The retraining of ML-based detection systems with hand-labelled adversarial DDoS traffic is extremely time-consuming. IBM Security suggests the incorporation of the ADM to prevent the crash of ML-based models. Studies indicate that the employment of the ADM can save up to 95% of the training time. However, traditional ADM was not equipped with the capability of self-learning. Thus, it is still vulnerable to adversarial DDoS attacks.

The proposed AG-IDS is a GAN with dual discriminators, as shown in Figure 3. The additional discriminator implements the concept of ADM, in charge of the discrimination of adversarial DDoS traffic. Traffic generated by the GP-WGAN is part of the training set in the training of the AG-IDS. In operation, incoming traffic is subject to the inspection of the ADM discriminator and then passed to the regular discriminator for further examination.

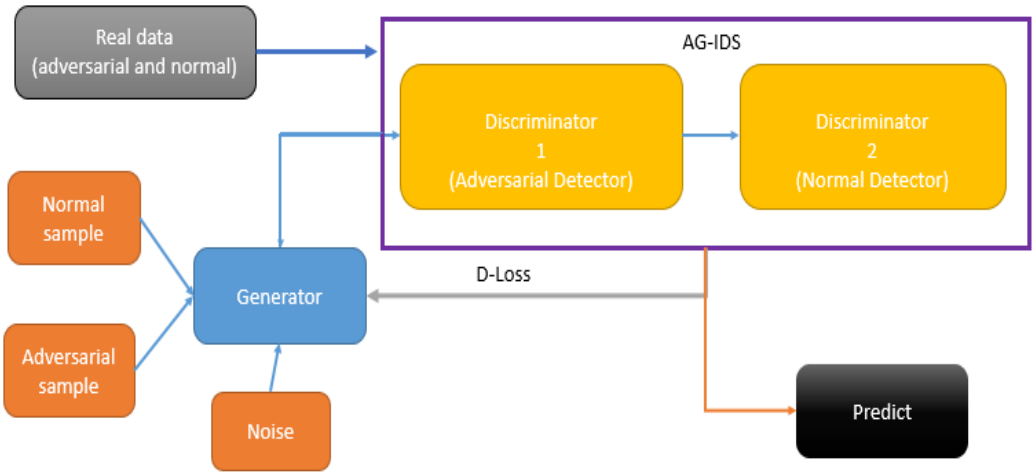

Figure 3. Architecture of AG-IDS.

3.1. Generator

The generator in a standard GAN takes samples from the problem domain to generate new instances to defeat the discriminator. For AG-IDS, to ensure the generation of adversarial DDoS traffic, the generator is fed with both normal and adversarial samples generated by the GP-WGAN. The loss function is defined as follows:

$$L_G = \mathbb{E}_{IG}\left[\log\left(1 - D_2 D_1\left(G(IG)\right)\right)\right] \tag{1}$$

where G is the generator, D_1 is the ADM discriminator, and D_2 is the discriminator for the differentiation of normal and malicious traffics. IG is the training set containing adversarial attack, conventional attack, and normal traffics.

3.2. Discriminator

We have adversarial DDoS traffic generated by GP-WGAN included in the pre-training of discriminator D_1. By doing so, we can have the correct gradient in the contending with the generator. Normal traffic is also considered in the training of discriminator D_2 to endow it with the capability in differentiating normal and malicious traffics. The loss function is defined as follows:

$$L_D = \mathbb{E}_{x \epsilon Adv}\left[\log\left(D_1(X)\right)\right] + \mathbb{E}_{y \epsilon Nor}\left[\log\left(D_2(Y)\right)\right] \quad (2)$$

L_D is fed back to the generator for it parameter adjustment.

Using two discriminators makes generator training and DDoS detection more efficient. In the experimental analysis below, we will see how inadequately when using only one discriminator against a synthesis attack. By separating the task, *D*1 detects adversarial attacks, *D*2 detects normal attacks, making the model transparent and effective against synthetic attacks.

4. Experiments

NSL-KDD [16] is employed in this study as the data set. NSL-KDD is a popular data set containing various types of attacks include two network attack layers, DoS and Probe, and two host-based attack classes, Remote-2-Local (R2L) and User-2-Root (U2R) which listed in Table 1. In our experiments, we focus on the DoS attack which including 41 network flow features.

Table 1. Types of attacks in NSL-KDD

Attack Type	# of Instances
DoS	45,927
Probe	11,656
U2R	52
R2K	995
Normal	67,343

Performance indices include the confusion matrix, as shown in Table 2, and the TPR and FPR, as defined in (3) and (4), respectively.

Table 2. Confusion matrix.

	Attack	**Normal**
Attack	TP (True Positive)	FP (False Positive)
Normal	FN (False Negative)	TN (True Negative)

$$TPR = {TP}/{(TP + FN)} \quad (3)$$

$$FPR = {FP}/{(FP + TN)} \quad (4)$$

The proposed GP-WGAN and AG-IDS are implemented using PyTorch on Intel i7-8700. PyTorch is a popular framework for deep learning. It is distinguished for its simplicity and ease in debugging. PyTorch enables the customization of existing models. There is no need to construct a model from scratch.

Table 3 reports the results of applying RF, KNN, and MLP to NSL-KDD. All three ML-based DDoS detection perform well.

Table 3. ML-model performance on NSL-KDD

ML-Model	TPR	FPR
RF	97.56%	4.3%
KNN	94.31%	5.7%
MLP	95.32%	4.8%

We now turn our attention to the impact of adversarial DDoS traffic. 80% of data from NSL-KDD is used for training the GP-WGAN, 10% is for the test, and 10% is for verification. The GP-WGAN has 5 layers with 148, 256, 128, 128, and 64 neurons using ReLU activation function for each layer. Adam gradient descent was used for the neural network training with the parameters lr = 0.001. By experiment, we choose gradient penalty λ = 0.01 for the lowest error reconstruction. In this analysis, the adversarial attacks are generated from input DDoS data and noise is equivalent to the capacity of the NSL-KDD data set. It is apparent, from Table 4 data that all three ML models fail to detect adversary DDoS attacks with very low Adv.TPR indices compared to the conforming performance metric on NSL-KDD. We can see this quandary overall in Figure 4.

Table 4. ML-model compared performance on NSL-KDD and Adversarial DDoS attacks

ML-Model	TPR	Adv.TPR
RF	97.56%	7.3%
KNN	94.31%	5.2%
MLP	95.32%	4.8%

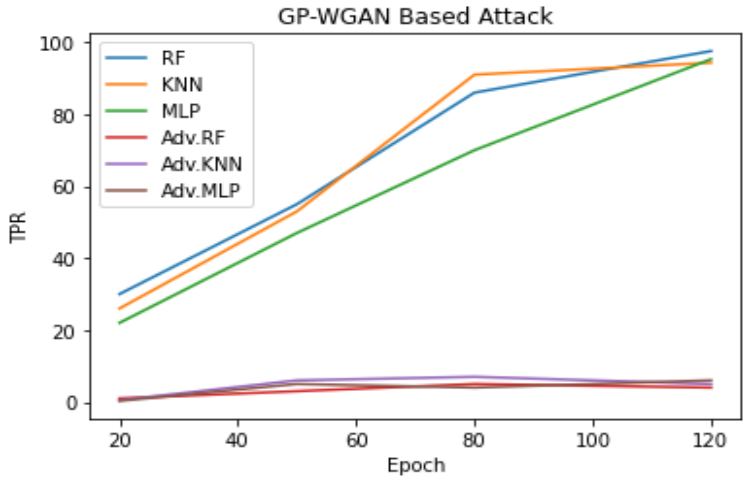

Figure 4. Detection rate of NSL-KDD and GP-WGNA generated traffic.

We also examine the performance of ANN, original GAN, and AG-IDS on conventional attacks, as shown in Table 5. Conventional DDoS attacks can be easily interrupted by all three approaches.

A new metric is defined in (5) to evaluate the capability in detection adversarial DDoS attacks.

$$ADR = \frac{N_D}{N} \tag{5}$$

where N is the number of all adversarial DDoS instances and N_D is number of detected instances.

The performance of ANN, original GAN, and AG-IDS on adversarial DDoS attacks is reported Table 6. It's evident that ANN and the original GAN are unable to detect adversarial DDoS attacks. Impressively, the proposed AG-IDS has an interrupt rate up to 87.39%.

Table 5. ML and DL performance on NSL-KDD

ML/DL-Model	TPR
ANN	88%
Original GAN	93%
AG-IDS	95%

Table 6. ML and DL performance on adversarial DDoS attacks

ML/DL-Model	ADR
ANN	0%
Original GAN	0%
AG-IDS	87.39%

5. Conclusions

This article investigates the potential threat of a new form of DDoS attacks, adversarial DDoS attacks. In the experiments for validation, NSL-KDD is used to train the proposed GP-WGAN architecture. The resulted generator is capable of the generation of legitimate-looking malicious traffic. Experimental results reveal that the generated adversarial DDoS attack can easily penetrate ML-based detection systems, including RF, KNN, and MLP. This phenomenon is an alarming and pessimistic wake-up call implying the urgent need for countermeasures to adversarial DDoS attacks. In response to this new threat, the AG-IDS architecture is proposed in this study. With an additional discriminator for the capturing and tagging of adversarial DDoS traffic, adversarial DDoS attacks can be effectively intercepted, as demonstrated in the experimental results.

Acknowledgement

This research was partly funded by the Minister of Science and Technology, Taiwan, under the grant number 109-2637-E-992-006 and MOST 109-2622-E-992-033.

References

[1] Genie-Networks, *DDoS Attack Statistics and Trends Report for 2020*, 2021. https://www.genie-networks.com/gnnews/ddos-attack-statistics-and-trends-report-for-h1-2020/

[2] A. Bakr, A. E. Ahmed, and H. A. Hefny, "A Survey on mitigation techniques against DDoS attacks on cloud computing architecture," *Journal of Advanced Science*, vol. 28, no. 12, pp. 187-200, 2019.

[3] S. S. Priya, M. Sivaram, D. Yuvaraj, and A. Jayanthiladevi, "Machine learning based DDoS detection," *2020 International Conference on Emerging Smart Computing and Informatics*, pp. 234-237, 2020.

[4] J. Cheng, J. Yin, Y. Liu, Z. Cai, and C. Wu, "DDoS attack detection using IP address feature interaction," *IEEE International Conference on Intelligent Networking and Collaborative Systems*, pp.113-118, 2009.

[5] N.H. Vu, "DDoS attack detection using K-Nearest Neighbor classifier method," *International Conference on Telehealth/Assistive Technologies*, pp. 248–253, 2008.

[6] A. Fadlil, I. Riadi, and S. Aji, "Review of detection DDoS attack detection using Naïve Bayes classifier for network forensics," *Bulletin of Electrical Engineering and Informatics*, vol. 6, no.2, pp. 140-148, 2017.

[7] C. Wang, J. Zheng and X. Li, "Research on DDoS attacks detection based on RDF-SVM," *The 10th International Conference on Intelligent Computation Technology and Automation*, 2017.

[8] U. Dincalp, "Anomaly based distributed denial of service attack detection and prevention with machine learning," *2nd International Symposium on Multidisciplinary Studies and Innovative Technologies*, 2018.
[9] Y. Li and Y. Lu, "LSTM-BA: DDoS detection approach combining LSTM and Bayes," *7th International Conference on Advanced Cloud and Big Data*, pp.180-185 2019.
[10] K. Yang, J. Zhang, Y. Xu, and J. Chao, "DDoS attack detection with AutoEncoder," IEEE/IFIP Network Operations and Management Symposium, pp. 1-9, 2020.
[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 27th International Conference on Neural Information Processing Systems, vol. 2, pp. 2672–2680, 2014.
[12] M. Arjovsky, S. Chintala, L. Bottou, "Wasserstein generative adversarial networks," 34th International Conference on Machine Learning, pp. 214-223, 2017.
[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," *c*, pp. 5769-5779, 2017.
[14] D. Sun, "A new mimicking attack by LSGAN," *2017 IEEE 29th International Conference on Tools with Artificial Intelligence*, 2017.
[15] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on GAN," arXiv:1702.05983, 2017.
[16] Canadian Institute for Cybersecurity, NSL-KDD https://www.unb.ca/cic/datasets/nsl.html
[17] T. A. Ahanger, "An effective approach of detecting DDoS using artificial neural networks," *2017 International Conference on Wireless Communications, Signal Processing and Networking*, pp. 707-711, 2017.