DDOS attack survey

Masoud Erfani

November 2019

1 Introduction

DDoS is short for Distributed Denial of Service. DDoS is a type of DOS attack where multiple compromised systems, which are often infected with a Trojan, are used to target a single system causing a Denial of Service (DoS) attack. DDoS attacks are costing enterprises anywhere between $50,000$ to 2.3 million dollars per year, which makes DDoS detection an important topic. Detection of DDoS attacks help network administrators to take immediate actions and mitigate the impact of such attacks.

Machine Learning has been applied for DDoS detection. These algorithms are dependant on the features and based on these features they will predict the label. There have been different research papers that focused on just calculating features from the network but, recently flow-based features have been incorporated with Machine Learning algorithms that acquired better results.

In this article, we focus on machine learning approaches based on flow-based features for DDoS detection.

2 Related work

In [3] authors have proposed a machine learning approach to detect DDoS for consumer IoT devices. Botnets such as Mirai launch attacks by suing insecure consumer IoT devices on critical Internet infrastructure. 250 vulnerabilities have been found in the 10 most popular IoT devices. So, it is critical to present an approach to detect these attacks. Authors have indicated that using IoT specific network behaviors for feature selection and combining with ML(machine learning) algorithms will lead to high accuracy for DDoS detection in IoT network traffic.
Their ML pipeline is as follows 1-Data collection 2-Feature extraction and binary classification. Features are designed based on IoT-specific network behaviors and using network flow characteristics such as packet length, inter-packet intervals, and protocol. This pipeline is designed to work on network middleboxes namely firewall, network switches and routers to detect anomalous traffic and corresponding devices.
The anomaly detection is done in four steps: 1- Traffic capture: the traffic capture process records the source IP address, destination IP address, source port, destination port, packet size and timestamp of all IP packets sent from smart home devices.
2 - Grouping packets: grouping is done based on device and time. 3- feature selection: This process can be done stateless or stateful. For each packet based on domain knowledge of IoT device behavior, the stateful consists predominantly of packet header fields and stateless aggregate flow information over a very short time window. Stateless can be derived from flow-independent features of individual packets and are most lightweight. On the other hand, stateful features capture how network traffic evolves. Stateless features are namely packet size, Inter packet interval, and protocol. Attack packet size is small since the attacker tries to keep the size of packets as small as

possible to maximize the number of connections. Also, Dos attack traffic has close to zero inter-packet intervals. In attack packets, TCP protocol outnumbers UDP ones. Furthermore, stateful features can be named bandwidth and IP address cardinality and novelty. IoT devices are characterized by a limited number of endpoints with which they communicate. Their set of destination IP addresses rarely change over time. So, two features are extracted to reflect mentioned notes: 1- Count of distinct destination IP addresses within the 10-second window. 2- The change in the number of distinct destination IP address between time windows.
4- Binary classification

Finally, ML algorithms namely Random Forest(RF), K-nearest neighbor(KNN), Support vector machine(SVM), decision tree(DT), and neural networks(NN) were applied. Results indicated that KNN and DT acquire high accuracy. Also, stateless features outperform stateful, which means that real-time anomaly detection of IoT attack traffic may be practical because stateless features are lightweight and derived from network flow attributes.

In [7], authors focused on ML(machine learning) algorithms to detect DDoS attacks in network communication flows. Taking advantage of a continuous learning algorithm that learns the normal pattern of network traffic, behavior of network protocols and identifies a compromised network flow. Their approach focused on using flow-based traffic characteristics to analyze the difference in pattern between normal and anomaly packet. Then ML algorithms were implemented.
Three groups of DDoS attacks were analyzed namely volume-based attacks, protocol-based attacks, and application-layer attacks. Volume-based attacks generate a huge volume of network traffic toward a target device such as UDP flood, ICMP flood, and SYN flood. Protocol-based attacks make use of the vulnerability present in specific protocols like TCP. Finally, application-layer attacks target specific application protocols such as HTTP and generating protocol packets in such a way that the server will have multiple long open sessions causing exhaustion of resources.
Their data-set for experiments was CICDS 2017 that contains a mix of both benign network traffic and most up-to-date Dos attacks. Realistic network traffic used a benign profile system that abstracts the behavior of human interactions and generates naturalistic benign background traffic. This data-set includes 85 features and in their paper, proper features were selected based on using recursive feature elimination techniques. Proper features will help the classification model to fit the data better. Correlation between features was analyzed to eliminate features that contribute to the same information about data. The more the Gini index decreases for a feature, the more important the feature is. Fourteen features were used in the final data-set. The classifier takes the data-set as input and categorizes them using a learning algorithm that will identify the instance to the best fit category. Classification models should be general to predict the current category of unknown samples. Logistic Regression, KNN, RF, and NN are executed for the experiment part. Evaluation of classification models is based on the number of correct predictions on the test data-sets which is given by the confusion matrix.

A recently published paper [8] mentioned that detection and designing a real-time detector with low computational overhead is still of the main concerns. So, they proposed a new taxonomy for DDoS attacks. Also, a new data-set namely CICDDoS 2019 was generated by them. Furthermore, they proposed a new detection and family classification approach based on a set of network flow features and provided the most important feature sets to detect different types of DDoS attacks with their corresponding weight.
The new data-set is labeled with 80 network features that have been extracted and calculated from all benign and denial of service flows. Building a model to capture patterns by training data and using four common ML algorithms namely ID3, RF, NB, and LR(logistic regression) was carried out by authors.
They designed and implemented two networks namely attack-network and victim network. An

active network is a separated network that executes different types of DDoS attacks. The victim network is a high-security infrastructure with firewalls, router, switches, and several common operating systems along with an agent that provides benign behaviors on each PC. In the data-set, 11 different DDoS attacks were created to diversify the DDoS attacks. 12 attacks such as NTP and DNS for a training day and 7 attacks such as portscan and NetBios in a testing day.

To find the best detection features for each DDoS attack, they tested the first 80 extracted features using Random Forest Regressor. Then, they examined the performance and accuracy of selected features with four common ML algorithms on training and testing data. For example, packet length std is one of the most important features for benign traffic.
DDoS attacks show clear behavior in sending packets to the victim, unlike benign traffic that usually does not. This behavior affects related features, and this can be the reason why these features are useful for detecting the attacks.

In [1] the research team studied the effectiveness of using ML algorithms to detect DDoS attacks by SDN(software denied networking). A recent paradigm that aims to improve network management by centralizing network information and control. Machine Learning algorithms were implemented on nmeta2(an SDN-based traffic classification architecture) and evaluated on a physical network testbed to demonstrate their efficacy during a DDoS attack especially in accurately classifying non-malicious traffic. Seven well known and established classifiers namely LDA(linear discriminant), QDA(quadratic discriminant analysis), SVM, KNN, NB, DT, and RF were used based on two features. The amount of information sent in one direction and duration of connection were the features that they took advantage of detection. A DDoS flooding attack results in a large volume of traffic being sent over a shorter period than a normal session. Features for the mentioned attack are flow duration and number of packets sent by the source. During the training phase, writers used K-fold cross-validation to counter overfitting. Four evaluation metrics were used by them namely initialization time, packet processing time, prediction time, and the number of predictions. Initialization time is the time taken for a classifier to be in a state where it can make a prediction. Also, the packet processing time is the time taken for a classifier to gather the required information to make a prediction. Based on the results classifiers with shorter initialization periods may not be desirable.

A comprehensive empirical evaluation of ML-based DDoS detection techniques was presented by [6]. Their framework is developed on different attack scenarios to investigate the performance of a class of ML-based methods. Class imbalance problem on machine learning-based DDoS detection was their evaluation metric. Class imbalance problem happens when the number of observations of one class is far less than the other classes. Different ML methods including classification, clustering, nearest-neighbor based, and statistical were compared with assuming that the entire network traffic is available for each DDoS detection technique. Results indicate that feature selection should be method-specific, and the capability of detecting attack traffic by ML methods is evident.

Imbalance problem if happens during the test, then accuracy is no longer enough to assess the performance of the detective scheme. Also, if it happens during the training phase, it may prevent the learning process of the classification algorithm. To assess the impact of the class imbalance in training data-sets, the authors generated a set of training data with different degrees of imbalance. A simple random under-sampling method to create five subsets from the training data-set was their method. The results showed that the impact of the class imbalance problem in datasets should not be neglected.

Another research group suggested implementing a semi-supervised K-means method for DDoS detection[10]. According to the authors, supervised approaches need a lot of labeled data, and un-

supervised techniques have relatively low detection rates and high positive rates. Their approach is as follows: 1- presenting a Hadoop-based hybrid feature selection algorithm to solve the problem of outliers and local optimal 2- providing the semi-supervised k-means algorithm using hybrid feature selection to detect attacks. The algorithm follows these steps: 1- data preparation 2- data preprocessing 3- model training. Also, another semi-supervised machine learning approach for DDoS detection was proposed, which is an online-sequential semi-supervised ML approach that was presented by another group [5]. This method is based on network entropy estimation, co-clustering, information gain-ratio, and extra-trees algorithms. The unsupervised part of their proposal reduces the irrelevant normal traffic data for DDoS detection, which will lead to false-positive reduction and increase inaccuracy. The supervised part reduces the false-positive rates of the unsupervised part and classifies the DDoS traffic accurately.

Another group focused on deep learning for DDoS detection in open flow-based SDN[2]. Their approach can learn the pattern from sequences of network traffic and trace network attack activities in a historical manner. It acquired better performance in comparison with previous machine learning ways. Their model reduces the degree of dependence on the environment, simplifies the real-time update of the detection system and decreases the difficulty of upgrading or changing detection strategy. The deep learning model consists of 1- input layer 2- forward recursive layer 3- reverse recursive layer 4- fully connected hidden layer 5- output layer. According to their experiments, DDoS attack detection scheme based on deep learning has an advantage of high detection accuracy and little dependence on hardware and software devices.

Another machine learning approach for DDoS detection focused on applying enhanced support vector machines with real-time generated datasets [9]. The authors generated a data-set that has 14 attributes and 10 types of latest DDoS attack classes. Then, enhanced multi-class support vector machines are used for the detection of attacks into various classes. Normal and attack traffic was classified by their proposed algorithm, which learns the normal and attack patterns from the training file. During the test phase, EMCSVM (enhanced support vector machines) differentiate the attack and normal traffic using learned patterns.

The last paper in this survey presented a hybrid technique for DDoS detection with supervised learning algorithms. They proposed a novel hybrid framework based on the data-streams approach for DDoS detection with incremental learning. A technique that divides the computational load between client and proxy sides based on their resource to organize the task with high speed. On the client-side the process is as follows: 1- data collecting of the client system 2- feature extraction based on forwarding feature selection 3- divergence test. If the divergence gets bigger than a threshold, the attack is detected otherwise data will be processed to the proxy side.
They applied different machine learning algorithms namely NB, RF, DT, MLP, and KNN on the proxy side to make better results. Experiments indicated that proper performance for detecting attacks and the ability to distinguish new attack types is achieved. Based on their results, RF(random forest) acquired the best results. [4].

3 Conclusion

In this survey, we presented different machine learning approaches that were incorporated with flow-based features from network traffic for DDoS detection. Since designing a real-time detector is still a problem, different ML algorithms can be applied for detection in future researches.

References

[1] Bakker, J. N., Ng, B., and Seah, W. K. Can machine learning techniques be effectively used in real networks against ddos attacks? *2018 27th International Conference on Computer Communication and Networks (ICCCN)* (2018), 1–6.

[2] Chuanhuang LiYan Wu, Xiaoyong Yuan, Z. S. W. W. X. L. L. G. Detection and defense of ddos attack–based on deep learning in openflow-based sdn.

[3] Doshi, R., Apthorpe, N., and Feamster, N. Machine learning ddos detection for consumer internet of things devices. In *2018 IEEE Security and Privacy Workshops (SPW)* (2018), IEEE, pp. 29–35.

[4] Hosseini, S., and Azizi, M. The hybrid technique for ddos detection with supervised learning algorithms. *Computer Networks 158* (2019), 35–45.

[5] Idhammad, M., Afdel, K., and Belouch, M. Semi-supervised machine learning approach for ddos detection. *Applied Intelligence 48*, 10 (2018), 3193–3208.

[6] Liang, X., and Znati, T. An empirical study of intelligent approaches to ddos detection in large scale networks. *2019 International Conference on Computing, Networking and Communications (ICNC)* (2019), 821–827.

[7] Lopez, Alma D.; Mohan, A. P., and Nair, S. Network traffic behavioral analytics for detection of ddos attacks. In *SMU Data Science Review: Vol. 2 : No. 1 , Article 14.* (2019), pp. 29–35.

[8] Sharafaldin, I., Lashkari, A. H., Hakak, S., and Ghorbani, A. A. Developing realistic distributed denial of service (ddos) attack dataset and taxonomy. In *2019 International Carnahan Conference on Security Technology (ICCST)* (Oct 2019), pp. 1–8.

[9] Subbulakshmi, T., Balakrishnan, K., Shalinie, S. M., Anandkumar, D., Ganapathi-Subramanian, V., and Kannathal, K. Detection of ddos attacks using enhanced support vector machines with real time generated dataset. *2011 Third International Conference on Advanced Computing* (2011), 17–22.

[10] YONGHAO GU, KAIYUE LI, Z. G., and WANG, Y. Semi-supervised k-means ddos detection method using hybrid feature selection algorithm. *IEEE* (2019).