Predicting Denial of Service Attacks Using Network Traffic Dataset

Kağan Özgün

May 16, 2023

Abstract

-

1 Introduction

Shortly after the emergence of personal computers, people began to encounter the definitions of cyber attack and cyber attacker. The methods used by these malicious people who want to damage computers or access data on these devices, have changed over time with the advancement of technology. Due to the fact that the internet was not widespread in the beginning, malicious software spread over input and output devices was used a lot. With the increase in the use of the Internet, the attackers started to make their attacks mainly over the network. With the spread of systems such as antivirus or intrusion detection system (IDS) used to prevent this situation, attackers began to look for methods that could be launched more insidiously. Denial of service (DoS) attacks, which became widespread in the early 2000s [Ran22], have also been one of the attack types that emerged for this purpose. DoS attacks generally aim to temporarily or long-term takedown of a targeted resource. In order to increase the effect of these attacks, similar attacks made from many points are called a distributed denial of service attack(DDoS). DDoS attacks are also divided into various categories according to the types of packets and requests they use, the frequency change of the number of requests used, and their purposes. In the prevention of these attack types, devices called IDS are used, which mainly include rule-based analysis and send network packets to the devices that provide service after passing this analysis. Since rule-based devices are insufficient in emerging attack types, it is preferred to use machine learning-based methods that can learn from incoming data.

Detecting file attacks instantly or detecting incoming attack types is very important for detecting future attacks and learning attack types. For this purpose, many different classification studies have been carried out in the literature and very successful results have been obtained in both binary and multi-class classification processes, especially in deep learning-based studies. What is more important than the instant detection of the incoming attack type by classification is to determine that the attack will start a certain time before the attack comes. Predicting an attack can give network security experts a huge advantage in preventing an attack.

Within the scope of this study, it is aimed to predict DoS attacks that will occur for a certain period of time by using an LSTM-based model. Model training and evaluation phases include two different commonly used public network dataset. Used techniques and approaches detail in the following section and result of the proposed solution and reference model compared in the evaluation section.

1.1 Problem definition

DoS attacks, which attack methods aimed at restricting the accessibility of applications or directly preventing the application from being accessible, are one of the greatest dangers, especially for applications with many users and sessions. In particular, during campaign times or if the application infrastructure is already under load, long-term problems may occur in the application exposed to such attacks. It may take time for the systems to be accessible again. Accessibility is a significant factor in areas such as banking, insurance and e-commerce, and the reliability of the application is severely damaged if it is under DoS attack.

1.2 Motivation

Although various studies have been carried out on detecting and preventing DoS attacks, it is still one of the most common attack types today. DoS attacks, first encountered and defined in 1999, became more widespread after the 2010s, and the severity of the attacks increased over the years[Ran22]. For handling the increase in DoS attacks more successful and flexible attack detection systems developed and different approaches proposed. In recent years the usage of artificial intelligence-based methods increased for this reason and they generate better predictions for attack detection. It gives very successful results in estimation, especially due to the memory structure used by LSTM-based models. For this reason, predicting DoS attacks with LSTM-based models can produce good results. In this study, it is aimed to predict DoS attacks with an LSTM-based model before the attack occurs, and to ensure that cyber security experts or systems take the necessary precautions.

1.3 Solution hypothesis

TODO

1.4 Contribution

TODO

2 Methodology

TODO

2.1 Data Preparation

TODO

Before the training phase of the work, datasets are prepared for training. Firstly some of the columns were removed from datasets then, categorical columns exploded as binary value columns. Sampling methods were used to solve the data imbalance issue after this step, new sampled datasets were created with 30% attack and 70% normal traffic rate. Also, data were normalized using a min-max scaler. Finally, the data split for the train and test phase with 70% train and 30% test ratio. Detail of the data preparation given in the sub sections.

2.1.1 CIC-DDoS2019

The CIC-DDoS2019 dataset includes 13 different types of DoS attacks and 70,619,331 rows of data. Dataset has 87 features about network traffic information. Most of the data is generated as attack data 99.83% rows include attack data and the rest of the data includes normal traffic.

Source IP, destination IP, source port, destination port, "SimillarHTTP" and timestamp features were removed from the dataset. Then, the data was balanced using data sampling from different labels. Also, the min-max scaler is applied to data for normalizing data. Finally, the dataset was split as 30% test 70% train data.

2.1.2 UNSW-NB15

UNSW-NB15 dataset includes 9 different types of attacks and one label is DoS, and other attack data was removed from the dataset. Dataset has 49 features and 2,540,047 rows of network traffic information. Most of the data is generated as attack data 99% rows include normal data and the rest of the data includes attack traffic.

Source IP, destination IP, source port, destination port and "ct_ftp_cmd" features were removed from the dataset. Categorical features which are state, proto and service; are encoded with one-hot-encoding. Then, the data was balanced using data sampling from different labels. Also, the min-max scaler was applied to data for normalizing data. Finally, the dataset was split as 30% test 70% train data.

2.2 Model Development

Details about model that developed in this work given in the sub section. Data preparation and model development phase of the project published in public git repository [Ozg].

TODO

3 Related work

TODO

A study on designing a system that detects especially HTTP flood and TCP flood type DoS attacks by examining tags and flags in network data with statistical methods has been done by Shaaban et al[Ram19]. As a result of this study, a model that is said to detect DoS attack types for 3 seconds after the start of the attack is proposed, but no metrics for the performance of this model are given. Method used in this project includes more complex deep learning algorithms comparing to Shaaban et al's work. Additionally, evaluation result are detailed and performance evaluation metrics like accuracy, false negative rates and recall used in proposed method evaluation section. Also, a reference model results added to evaluation process.

In another study, Sitiawan and other researchers aimed to detect Ping Flood attacks, a kind of DoS attack, using IoT network data with the K-Means method [Sti21]. In this study, both accuracy value and confusion matrix results are shared as performance metrics, and a successful result was revealed according to these results. However, the dataset used in the study contains approximately 95% of attack data and 5% of normal data, which is different from real-life scenarios. For this reason, the realistic performance of this approach will be different from the experiment. Our methodology includes three different dataset and each of the dataset processed before model development phase for simulating real world scenarios.

In their study, which was published in 2022, Gopi et al. developed a model on DoS attack detection with an ANN-based deep learning model[Gop22]. In this study, it is desired to present a successful model that can predict faster than a standard Artificial Neural Network (ANN) model using dimension reduction approaches. However, similar to the previous study dataset used includes more attack data than normal network traffic. Dataset used in this publication has 30% of normal traffic and 70% of attack traffic. In the result section of the paper proposed model performance is compared with an ANN model. Comparing to this work, our methodology has different kind of deep learning models and a reference model for generate realistic evaluation. Besides, balanced datasets that includes 70% of normal traffic and 30% of attack traffic used in proposed methodology.

In another study by Yan Li and Yifei Lu, a model that detects DoS attacks was developed using a model that includes LSTM and Bayes methods. A comparison was made with the model containing only LSTM in the conclusion part[Li,19]. In the conclusion part of this study, a sample study, the result of a base model and the result of the proposed model are compared, and the results are given with different metrics. However, as in other studies, the dataset used was chosen in a balanced way, containing 50% normal 50% attack traffic, unlike the network traffic encountered under normal conditions. In Yan Li and

Yifei Lu's study deep learning models hat include LSTM layer used similar to our proposed model and in evaluation section different evaluation criteria and detailed results provided in paper. However, different kind of dataset are not used in their work. Working with single dataset decrease the reliability of the results and proposed method in realistic environment. For increasing reliability of the proposed method and results three different dataset used in our proposed work.

Developing a DL-based model for detecting DoS attacks includes complex neural network structures and complex calculations. Some of the researchers proposed a DL model with complex feature extraction and feature manipulation methods to increase the efficiency of the models. Luhan Zou and other contributors proposed a model with a feature grouping technique[Zou22]. This approach is named feature-attended multi-flow long short-term memory (FAMF-LSTM) by the authors. The technique includes a future engineering part with grouping input features depending on similarities then they feed different groups to a LSTM-based model for predicting DoS attacks in the dataset. They used BoT-IoT and UNSW-NB15 datasets for the training and test phase of the work. Results show their proposed hypothesis is correct. The proposed model got better accuracy compared to models without feature manipulations. However, fine-optimized DL-based models can get better results without feature manipulation. In some specific cases, feature manipulation can generate better results but for DoS detection using BoT-IoT and UNSW-NB15 datasets, additional feature operation is not needed. Optimized RNN or LSTM models can achieve better results compared to the method proposed by Luhan Zout et. al.

Detecting the attacks depends on network data structure and attack-normal traffic rate in the dataset. In the real-world scenario, most of the network traffic is generated by normal users so it includes normal traffic heavily. However, generated datasets generally include more attack traffic compared to normal traffic of the network. This situation cause difference between the success of the model on the test dataset and the real-world scenario. Due to solve that problem, using dataset balancing techniques can be reasonable. Jirasin Boonchai and other researchers used the data balancing method for handling this problem. They used the Synthetic Minority Oversampling Technique (SMOTE) technique for solving the imbalance issue of the CIC-DDoS2019 dataset[Boo22]. Dataset samples were created using SMOTE for each class label and the new sampled dataset includes 50% of attack and 50% of normal traffic. After sampling they trained DNN-based and convolutional autoencoder-based models and compare the result of the models with Naive Bayes and Logistic Regression models. Proposed models got better predictions compared to reference models but accuracy and other metrics are not good enough compared to solutions in the literature. For solving the imbalance problem, a similar approach is used in our work but SMOTE is not used. A sampling method depends on python numpy library used for generating balanced sampled datasets. 30% attack traffic and 70% normal traffic ratio are preferred in the proposed method in this work because this ratio is more realistic compared to 50-50 ratio.

Brief comparison of the related work and proposed methodology provided

Table 1: Comparison of the related work methodology

	Include DL Model	Multiple Dataset	Multiple Evaluation Criteria
Proposed Solution	**YES**	**YES**	**YES**
[Ram19]	NO	NO	NO
[Sti21]	NO	YES	YES
[Gop22]	YES	NO	YES
[Li,19]	YES	NO	YES
[Zou22]	**YES**	**YES**	**YES**
[Boo22]	YES	NO	YES

in Table 1. Comparison table includes three specific information about related works and proposed method.

4 Results

TODO

4.1 Results of CIC-DDoS2019

TODO

4.2 Results of UNSW-NB15

TODO

5 Discussion

TODO

6 Conclusion

TODO

References

[Boo22] Boonchai, Jirasin and Kitchat, Kotcharat and Nonsiri, Sarayut. The Classification of DDoS Attacks Using Deep Learning Techniques. In *2022 7th International Conference* *on Business and Industrial Research (ICBIR)*, pages 544–550, 2022. https://doi.org/10.1109/ICBIR54589.2022.9786394 doi:10.1109/ICBIR54589.2022.9786394.

[Gop22] Gopi, R. and Velayutham, Sathiyamoorthi and Selvakumar, S. and Manikandan, Ramasamy and Chatterjee, Pushpita and Zaman, Noor and Luhach, Ashish. Enhanced method of ANN based model for detection of DDoS attacks on multimedia internet of things. *Multimedia Tools and Applications*, 81:1–19, 08 2022. https://doi.org/10.1007/s11042-021-10640-6 doi:10.1007/s11042-021-10640-6.

[Li,19] Li, Yan and Lu, Yifei. LSTM-BA: DDoS Detection Approach Combining LSTM and Bayes. In *2019 Seventh International Conference on Advanced Cloud and Big Data (CBD)*, pages 180–185, 2019. https://doi.org/10.1109/CBD.2019.00041 doi:10.1109/CBD.2019.00041.

[Ozg] Ozgun, Kagan. Detecting Denial of Service Attacks Using Network Traffic Data. URL: https://github.com/kaganozgun/dos-detection-on-network-data.

[Ram19] Ramzy Shaaban, Ahmed and Abdelwaness, Essam and Hussein, Mohamed. TCP and HTTP Flood DDOS Attack Analysis and Detection for space ground Network. In *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pages 1–6, 2019. https://doi.org/10.1109/ICVES.2019.8906302 doi:10.1109/ICVES.2019.8906302.

[Ran22] Rangapur, Aman and Kanakam, Tarun and Jubilson, Ajith. DDoSDet: An approach to Detect DDoS attacks using Neural Networks, 2022. URL: https://arxiv.org/abs/2201.09514, https://doi.org/10.48550/ARXIV.2201.09514 doi:10.48550/ARXIV.2201.09514.

[Sti21] Stiawan, Deris and Suryani, Meilinda Eka and Susanto and Idris, Mohd Yazid and Aldalaien, Muawya N. and Alsharif, Nizar and Budiarto, Rahmat. Ping Flood Attack Pattern Recognition Using a K-Means Algorithm in an Internet of Things (IoT) Network. *IEEE Access*, 9:116475–116484, 2021. https://doi.org/10.1109/ACCESS.2021.3105517 doi:10.1109/ACCESS.2021.3105517.

[Zou22] Zou, Luhan and Wei, Yunkai and Ma, Lixiang and Leng, Supeng. Feature-Attended Multi-Flow LSTM for Anomaly Detection in Internet of Things. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2022.

https://doi.org/10.1109/INFOCOMWKSHPS54753.2022.9798066
doi:10.1109/INFOCOMWKSHPS54753.2022.9798066.