Detection DDoS Attacks Based on Neural-Network Using Apache Spark

Chang-Jung Hsieh[1, a], Ting-Yuan Chan[2, b]

[1]Dept. of Information Management, National Chin-Yi University of Technology, Taichung, Taiwan
[2]Dept. of Information Management, National Chin-Yi University of Technology, Taichung, Taiwan
[a]www6809@gmail.com, [b]c07408@ncut.edu.tw

Abstract

Network security issues are becoming serious with the growth of Internet, in many types of network attacks, The Distributed Denial of Service (DDoS) has become the vital threat, The Akamai report reveals first quarter of 2015 there are more than doubled attacks in the same quarter last year, In addition, the types of DDoS attacks changing frequency, in the past, attackers used huge volumes of traffic in short time to make victim host unavailable, but now some attackers used low volumes of traffic for a long time making attacks difficult to detect. Nowadays, there already have many methods for the DDoS detection, but no one can fully detect, the difficulty lies in DDoS attacks often have huge volumes of traffic, in first quarter of 2015, there were 8 attacks exceeding 100 Gbps, and their characteristic highly variable, make hard to detect, in order to overcome it, this paper propose DDoS detection method based on Neural Networks, implemented in the Apache Spark cluster, we used 2000 DARPA LLDOS 1.0 dataset to train and perform experiments to our detection system in a real network environment, the results show our detection system able to detect attacks in real time and average detection rates were over 94%.

Key words: DDoS Detection、Traffic Analysis、Big Data、Apache Spark

Introduction

The purpose of Distributed Denial of Service (DDoS)[1] is expend the resources of a victim host to make services break off, utilize numerous of zombie computer have Trojans, send huge volume of abnormal traffic to victim host, in general this type of attacks are not broken system or steal data from victim host, but it still often causes enterprise loss of property, since this DDoS attack is very easy to implement, it became prevalent and serious network security issue in recent years, There are many types of DDoS attacks, they have different attack rate, traffic volume and even forged source address. For these reasons, the detection methods mostly use the algorithm which has ability to predict, more over the detection system must be able to separate the abnormal traffic from genuine traffic, guarantee the normal traffic can be able to access, the main issue is algorithm will spend a lot of time to compute, especially in case of high volume, sometimes attacks were affected victim host, but the detection system can't detect abnormal traffic immediately.

The key objective of this study is to construct a detection system have high accuracy and immediate, The Artificial neural networks(ANNs)[2] is used to identify and detect abnormal traffic, and used famous DDoS tools like ARPA 2000 LLDOS 1.0[3] and self-generated to training ANNs, make our detection system be able to identify the characteristic of abnormal traffic, the feature of ANN is their network will not expand by data volumes, so it is suitable for the huge volume analysis like DDoS detection, and we used Apache Spark an open source cluster computing framework, The feature of this framework is In-Memory computing, break through the bottleneck of hard disk, make run programs up to 100x faster than Hadoop MapReduce in memory.

Related Work

Artificial neural networks(ANNs) are abstract mathematical models of brain structures and functions, ANNs could be a prediction ability AI machine, through high performance computing, the feature of ANNs is their structure will not expand with the number of input and output does not change, so the memory used will be under control, and have high reliability that ability to predict unknown input data, ANNs is used in many fields, there are many types of ANNs, it contains Back-propagation Network, Hopfield Network, Radial Basis Function Network, they will be used in different situation by their suitable conditions, Alan Saied、Richard E. Overill and Tomasz Radzik[4] used Back-propagation Network to detect DDoS attack, and have high accuracy about 98%.

DDoS attack detection has some patterns different with genuine traffic, our detection system detects DDoS attacks based on specific characteristic features, we used some fields in the IP header to calculate DDoS attacks features which proposed by previous researcher Reyhaneh Karimazad and Ahmad Faraahi[5].

Apache Spark[6] is a big data processing framework work on a cluster, Apache Spark use In-Memory computing, run programs up to 100x faster than Hadoop MapReduce in memory, and able to use with another Hadoop component, like the Hadoop Yarn to manage the resource of cluster and Hadoop HDFS to store data.

This study combines the advantages of other research, construct computing framework to process huge volume of network traffic by Apache Spark, after that we use ANNs to analyze network traffic, the key objective of this study is construct a detection system fast and with accuracy.

Methodology

The system architecture is shown in Fig 1. Our detection system divided into five phases. (1) Packet Collector, (2) Hadoop HDFS, (3) Format Converter, (4) Data Processor, (5) Neural Network detection module.

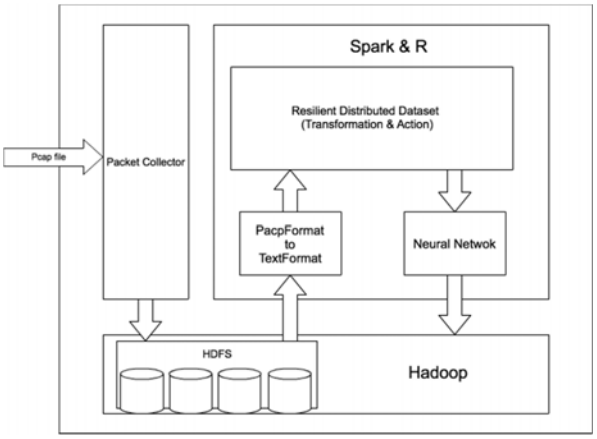

Fig. 1 Detection system architecture.

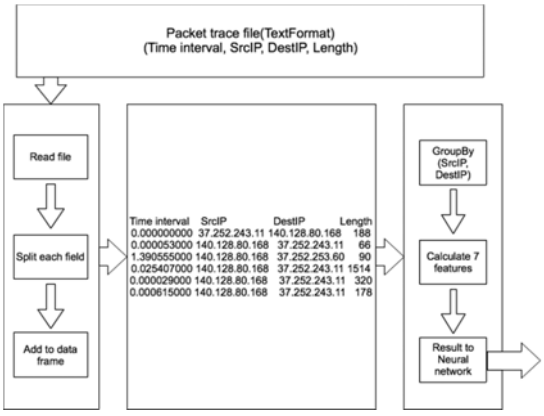

Fig. 2 Detection system data flow and output format.

A. Packet Collector

Packet Collector capture incoming packet with a fixed time window about every 1000 packets and generate packet trace file such as Tcpdump, we catch six fields from IP header that accommodate source IP address, destination IP address, time interval and length, the packet trace file would be stored to Hadoop HDFS.

B. Hadoop HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system, HDFS has the fault-tolerant ability, it stores the file duplicate to another node, and have a high throughput for applications that suitable to store the large volume of traffic data.

C. Format Converter

The common format of Spark is a text file. But the packet trace file (.pcap) is a binary format, so we have to convert the packet trace file to text file, and split by each line to Data Processor.

D. Data Processor

This phase is used resilient distributed dataset (RDD) to compute some necessary features proposed by Reyhaneh Karimazad and Ahmad Faraahi [5]. We grouped the packet, and calculated the features which have the same source IP address and destination IP address, and the Data Processor is compiled by R language which is the famous language used in statistics and data mining, Fig 2. Shows our detection system data flow and output format.

These characteristic features help us detect the DDoS attack from the genuine traffic, after calculated, these features will be a neural network input parameter.

Seven features are used to DDoS attack.

- Number of Packets
- Average of Packet Size
- Time Interval Variance
- Packet Size Variance
- Number of Bytes
- Packet Rate
- Bite Rate

E. Neural Network

Our detection system detects DDoS attacks by seven parameters, so the Neural Network input layer has seven nodes, these parameters are obtained from Data Processor that consists of Number of Packets, Average of Packet Size, Time Interval Variance, Packet Size Variance, Number of Bytes, Packet Rate and Bite Rate.

In order to train, analyzing ability, our neural network the number of nodes in hidden layer should be 2n+1, n is the number of input layer based on other research.

The output layer in this neural network has one node, the output will be 0 or 1, 0 represent normal, and 1 represent an attack, the neural network architecture is shown in Fig. 3

Experiments and analysis

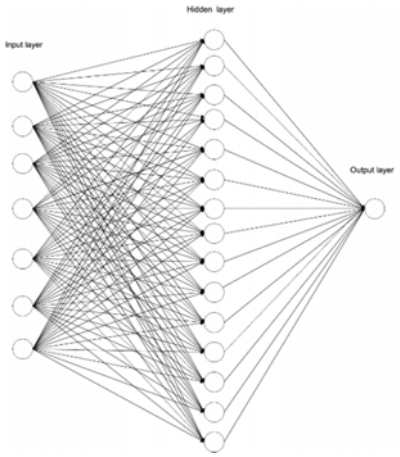

Fig. 3 Neural network architecture

We used 2000 DARPA LLDOS 1.0[3] datasets in our experiments. We generated normal traffic data by application, and 2000 DARPA LLDOS 1.0 is used to represent attack traffic, we merge normal and attack traffic datasets to train and test our detection system.

The Packet Collector captured 51040 packets for the normal traffic and 74480 packets for attack traffic, after calculated, we obtained seven items of features, Table 1 shows a few of the results from Data Processor, the first three are features of normal network traffic, and others are attack network traffic.

Table 1
Result of Data Processor

No.	Source IP	Destination IP	Number of Packets	Average of Packet Size	Time Interval Variance	Packet Size Variance	Number of Bytes	Packet Rate	Bite Rate	Type
1	140.128.80.168	104.115.210.132	103	121.57	0.160	22941.48	12522	6.121	744.1	Normal
2	117.56.6.3	140.128.80.168	6	220.16	0.121	157852.16	1321	6.793	1495.7	Normal
3	140.128.80.168	104.115.215.45	4	68	5.251×10^{-4}	16	272	55.555	3777.7	Normal
4	76.199.77.195	131.84.1.31	2	60	4.805×10^{-10}	0	120	1.226×10^{4}	7.361×10^{5}	Attack
5	91.40.94.62	131.84.1.31	2	60	1.8×10^{-10}	0	120	1.388×10^{4}	8.333×10^{5}	Attack
6	3.232.241.222	131.84.1.31	2	60	3.285×10^{-9}	0	120	9.132×10^{3}	5.479×10^{5}	Attack

In this case, we found the Number of Packets in normal traffic data is larger than attack traffic, it means the source IP address and destination IP address of genuine traffic is more centralized than attack traffic, because of the attack traffic have 32 bits IP spoofed generated by 2000 DARPA LLDDOS 1.0 dataset, the source IP address may have 2^{32} situations cause the attack traffic distributed than genuine traffic.

Time Interval Variance is clearly different between genuine and attack traffic, the attack traffic is generated by the zombie computer and send at the close time, the value in the attack would be close to 0, and time interval variance in genuine traffic would higher than attack.

Packet Size Variance would very close to zero, because of the packet was batch generated by attack program, so all the packet size would be same.

Packet Rate may raise in attack, because of the attacker use the large volume of traffic to attack, and our system will get lots of packets in this time.

Finally, three features Average of Packet Size, Number of Bytes and Bite Rate may not as easy as the other features to identify is attacking or not. Because of these features is depends on the packet size, so if the packet size which is set by the attacker is close to the current genuine traffic, this situation is represented on TABLE I no.3, it may lower the accuracy caused by the false positive or false negative.

Within this context, network traffic is a very complex environment and not easy to predict, sometimes thus these seven features don't have the threshold value to clearly define what traffic is attacking or not, so we analysis these features of the neural network by learning history data helping us identify the network traffic, the Fig. 4 shows our well-trained neural network architecture.

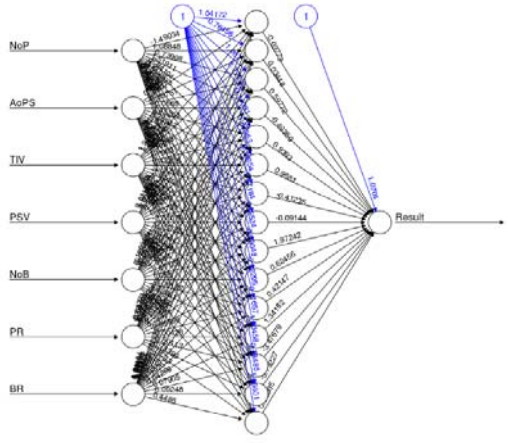

Fig.4 Well-trained neural network architecture

In the experiments, we training and test our detection system in the simulated network traffic, in order to make our test environment closer to the real world, we are sampling 30% data from normal and attack traffic to training our neural network, and 70% used to simulate the 950s network traffic to test, we mixed the 375s normal traffic and 575s attack traffic and divided into 5 phases, the phase 1, 3, 5 is attack, 2 and 4 is normal, the Fig.5 shows the expected out in this experiment.

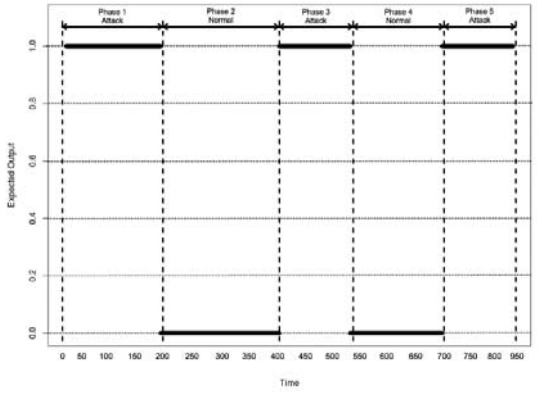

Fig. 5 Expected Output

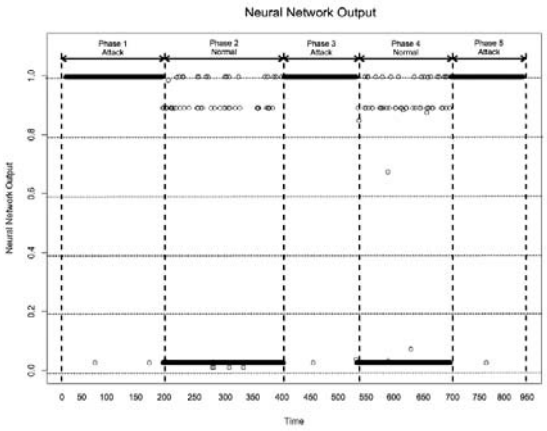

Fig. 6 Neural Network Output

In Fig.6 shows the results of neural network, and we can see the most of result is correct, but a little part is not correct, when the expected output is 0 but in fact is 1 we called false positives, and the excepted is 1 but in fact is 0 called false negatives, in this test, we got 995 true positives, 782 true negatives, 106 false positives and 5 false negatives, the accuracy is 94%.

The reason we got false positives is some features of normal traffic is similar to attack traffic, this type of packet is shown in Table 1 No.3, this situation in the normal traffic may occur on that normal connection which has large traffic, so their features will similar to attack traffic, and false negatives caused by the attacker used low traffic or generate more concentrated source IP address, destination IP address and not to send packets in the close time.

Conclusion

In this paper, we present a DDoS detection system used big data platform integrated the neural network, the detection system is composed of the open source big data computing framework Apache Spark and compiled by R language which is low cost and easy to implement. The experiment result proves our system is able to analyze the high velocity, and high volume network flow in real time and separated the genuine and attack traffic successfully, the result represents the Apache Spark is suitable to process large volume network traffic, and neural network identifies the features of packets effectively, the accuracy is about 94%.

The limitation of this study is sometime the packets will have unexpected features, it means in our system will get the incorrect results, the accuracy will be decreased. To increase the accuracy, in future work, the researcher attempt is to find out the other features of network traffic to detect the DDoS accurately, and use the newer DDoS attack tool or datasets to training the detection system, maybe this can help detection module to handle more situation in the complex network environment.

References

[1] Fengxiang, Zhang, and Shunji Abe. "A Heuristic DDoS Flooding Attack Detection Mechanism Analyses based on the Relationship between Input and Output Traffic Volumes." Computer Communications and Networks, 2007. ICCCN 2007. Proceedings of 16th International Conference on. IEEE, 2007.

[2] T.M. Mitchell, Machine Learning 81–117, 128–145, 157–198, 1st ed., McGraw- Hill Science/Engineering/Math, New York (1997) 52–78, Chapters 3,4,6,7.

[3] Zissman, M. "DARPA intrusion detection scenario specific data sets." (2000).

[4] Saied, A., Overill, R. E., & Radzik, T. (2016). Detection of known and unknown DDoS attacks using Artificial Neural Networks. Neurocomputing, 172, 385-393.

[5] Reyhaneh Karimazad and Ahmad Faraahi. An Anomaly-Based Method for DDoS Attacks Detection using RBF Neural Networks. IPCSIT vol.11 (2011) © (2011) IACSIT Press, Singapore.

[6] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D. & Xin, D. (2015). Mllib: Machine learning in apache spark. arXiv preprint arXiv:1505.06807.