Tor Traffic Classification using Deep Learning

Raval Radha, Deepak Upadhyay
Student, GTU- School of Engineering and Technology
Assistant Professor, Gujarat Technological University

Abstract: Tor is an anonymous Internet communication system based on the second generation of onion routing network protocol. Tor protects user's freedom privacy against surveillance and censorship by making it more difficult for an observer to correlate visited websites in the Internet with the real physical-world identity. Sometimes this browser can used by cybercriminal to do their illegal activities. Tor accomplished that by ensuring adequate protection of Tor traffic against traffic analysis and feature extraction techniques. Further, Tor ensures anti-website fingerprinting by implementing different defences like TLS encryption, padding, and packet relaying. In this paper proposed technique able to identify whether host is generating Tor-related traffic. For that well-known deep learning algorithms in order to evaluate the effectiveness of the proposed feature set in a real world environment. In addition this technique demonstrates that the proposed method is able to recognize the kind of activity the user under analysis is doing on the Tor network.
.

Index Terms—FFN, DL, HTTP, SVM, VOIP,P2P, FTP, HMM

I. INTRODUCTION

Traffic classification is basic to numerous activities including network management and security monitoring, traffic modelling, planning and Quality of Service (QoS)provising. The industry as well as research community make many efforts and produced many techniques for traffic classification. However, continues growth of internet and its offer services along with the encrypt and disguise the services make this traffic classification task more challenging to the internet research community. one of the key thing that make traffic classification more challenging is encryption, a key technology that hide the internet user, protect user freedom and privacy and providing them anonymity and protect themselves against the surveillance system.

Tor (Dingledine et al., 2004) is currently the most popular privacy enhancing tool. It can Anonymized the identity of users as well as their Internet activity by encrypting and tunnelling the traffic through a distributed network of servers, known as Tor nodes. Tor provides strong implementation, which protects against both sniffing and analysis making a secure communication to protect data confidentiality and users privacy.TLS protocol is used in tor communication to provide the required encryption.

For example, Bob and Alice communicating on a public internet connection, by using the mean of Tor, they can ensure they communication can't be intercepted or monitored by eavesdropper and that information passed back and forth is encrypted and secured. Tor is free open source software that works on almost every platform, once tor installed user can use this web browser to Anonymized their traffic. Traffic passes between tor nodes and user are via strong encryption. Tor works perfectly with other browser like Firefox and Chrome with their tor bundle.

In this paper we focus on the classification of Tor traffic that is downgrading privacy to some ex- tent by exposing the activity within the Tor traffic. Given a traffic flow, we aim to detect whether it is Tor traffic or not by using UNB-CIC tor-nontor dataset. For classification purpose, we are using deep learning's feed forward neural network architecture and comparing the accuracy of model with other machine learning algorithm.

II. RELATED WORK

Tor has been subjective of many research papers, focusing many of them on compromising Tor's anonymity or improving its performance. Another interest related with tor network and closer to this problem is almost all cases of analysis is done in the tor node. In only two to three paper tor traffic classification is done between client node and entry node using time based features.

In this Author detect that connection is to or from the tor network. This detection is based on the tor updated server list. There is one script that run on the any traffic analyzer software. This script check the network traffic to find either connection is to or from the tor network. Firstly, that match the source ip and destination ip of the packet to updated tor server list. If the ip match then alert is generated and send to the network admin for further action. The main goal of this research paper is to find out the tor user in small network and this method is generating large number of false positive[1]

In the authors proposed model for detect tor traffic using supervised machine learning algorithm. For model analysis is done from local observer and bypass the tor protection. Machine learning is whole step by step process. So firstly generating the dataset by fingerprinting top sites on Alexa and some of the tor sites and all this data is gathered using tools like wireshark and tcpdump. There are total 40 feature is selected for classification purpose. They test different algorithm to classify the traffic. The Algorithms are Naïve bayes, C4.5, Random forest and Support vector machine. The main goal of this paper to achieve high accuracy and law false positive rate by using supervised machine learning algorithms. In the comparison result all algorithm achieve accuracy up to 99% and correctly distinguish the HTTPS traffic and Tor traffic[2]

In this paper propose a fingerprinting method to identify the tor and web-Mix networks. Their method uses specific strings, packet length and frequency of the packets. They test their methods on the simulated network obtaining more than 95% accuracy in the both of the system[3]

In this author use Deep Packet Inspection (OpenDPI) to analyze the traffic from group of 6 exit nodes deployed for that purpose. Their result shows that more than 50% of traffic belongs to the bittorrent applications. Although OpenDPI is not able to detect encrypted connection, around 30% of the traffic the author claim that these connection also belongs to P2P,after analyzing the usage of encryption in Bitorrent connections[5]

In the author propose a method based on Hidden Markov model (HMM) to classify the encrypted traffic in the four category:P2P,FTP,IM and web. As classifiers they use burst volumes and directions extracted from Tor flows as a feature. They use HMM to build ingress and egress models of different web application (P2P.FTP.IM and web).They got overall maximum accuracy value of 92%[4]

Fingerprinting method used for analyzing the traffic and detect the user who access which website over network. In this paper author make model using deep learning algorithm. Stacked denoising autoencoder (SDAE) deep learning Method used for this fingerprinting attack. The dataset used in this method is already available with 100 sites as monitored and 900 sites as non monitored websites. The main feature taken as input is time and directional flow of the packet to know the website. The method is compare to other deep learning methods like OSAD and K-nun. So this method achieves the high accuracy than these two methods in open world's environment and accuracy is decreased in closed world environment. The main goal of this paper to achieve high accuracy without selecting features manually[6]

In the authors present an analysis of Tor traffic using an Intrusion Detection System. The papers present the result on analysis done using the Suricata, and commercial IDS rule-set. According to their result 10% of the tot traffic is malicious and it triggers an alert for that. From that 10%,more than 70% of the alerts where triggered by the P2P traffic[9].

Low latency mix networks are vulnerable to traffic analysis due to inherent statistical characteristics of packet data stream and stringent latency requirement incurred by interactive applications. Note that even if the established communication channel and payloads are encrypted and padded to hide payload size, Inter packet arrival time (time differences between consecutive packets) cannot be concealed because of the low latency requirement of the application. One of the papers that focus on the same area of timing analysis is where the author's finds a correlation of inter-packet arrival time and packet flow in order to identify network traffic in mix-networks. By modifying packet flow they were able to fingerprint origin (e.g. browser) and destination (e.g. destination) of IP traffic. To get rid of this privacy attack, authors propose adaptive padding algorithm- where an expected inter-packet interval (EIPI) is randomly chosen in order to destroy natural fingerprints. As the experiments shows the correlation coefficient between two links of the same path based on Inter-packet intervals lies to 0.9 while 0.3 for un- related links. Introducing adaptive padding reduces correlation within the same flow to 0.2-0.4.[14]

Tor traffic can be detected by analyzing the traffic packets. The analysis can be done on the Tor node or, in between client and entry node The analysis can be done on a single flow of packet .Each flow constitutes a tuple of source address, source port, destination address and destination port. Network flows for different time intervals are extracted and analysis carried on them. In this paper traffic characterization is based on the time based features. At the different time interval all the features are extracted and create huge tor-nontor dataset by UNB-CIC. That is not detect only the tor-non traffic but also find application behind the traffic. There are 7 type of traffic categorized as: VOIP, mail, P2P, FTP, Video, Audio. There are total 23 features is selected from all features. For traffic classification different machine learning algorithm is used given by tool Weka. They give comparison table of all algorithm and calculated accuracy of each algorithm [7]

III TOR ARCHITECHURE

Tor is free open source software that work almost on every platform, once tor is installed, users can use this web browser to anonymous their traffic .Traffic passes between Tor Nodes and users are via strong encryption. Tor is the most popular anonymity network nowadays that protects user communication and attains anonymity and privacy[13]. Tor is used for online anonymity; it's heavily used by the attacker and hacker to avoid their traceability.

Tor allows people to access and publish content on the internet without being identified or cleared to authorities. The usage of Tor by various type of people the risk is varied from a child accessing forbidden sites to other type of risk such as employee or political activist accessing tor where the risk is higher. The main purpose of Tor to protect the user privacy but we can't deny the fact that used by criminal to commit their crimes with impunity[12].

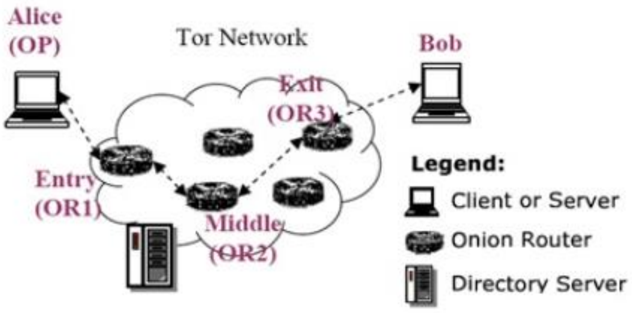

Figure 1. Tor client-server Architechaure

In the figure1. Alice is client and bob is a server. The data is passing through three server node and after that passing to destination node. There are three onion router and each router knows ip address of only before or after router and original content is not revel at any router. The destination node can only see the original content that coming from the Alice. Here, Directory server maintain all the list of active router in the relays and give that information to the client. Client can choose the relays where data is passing through and reach to the destination node.

IV.BRIEF OVERVIEW OF FEED FORWARD NEURAL NETWORK

Deep learning is an advanced model of traditional machine learning. This has the capability to extract optimal feature representation from raw input samples. This has been applied towards various use cases in cyber security such as intrusion detection, malware classification, android malware detection, spam and phishing detection and binary analysis. Neurons are the atomic unit of a biological neural network. Each neuron consists of dendrites, nucleus, and axons. It receives signals through dendrites and is carried out through axons (Figure 1 below). The computations are performed in the nucleus. The entire network is made up of a chain of neurons.

AI researchers borrowed this idea to develop the artificial neural network (ANN). In this setting, each neuron accomplishes three actions:

1. it accumulates input from various other neurons or inputs in a weighted manner
2. it sums up all input signals
3. based on the summed value, it calls an activation function

Each neuron thus can classify whether a set of inputs belong to one class or another. This power is limited when only a single neuron is used. However, coining a set of neurons makes it a powerful machinery for classification and sequence labeling tasks.

A set of neuron layers can be used to create a neural network. The network architecture differs based on the objective it needs to achieve. A common network architecture is a Feed Forward Neural Network (FFN). Neurons are arranged linearly without any cycles to form a FFN. It is called feed forward because information travels in the forward direction inside the network, first through the input neurons layer, then through the hidden neuron layers, and the output neurons layer. The figure below shows the feed forward network with two hidden layer.

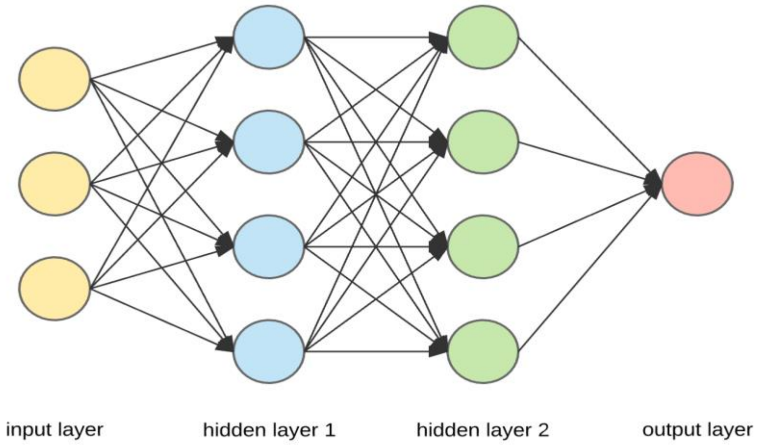

Figure 2.Feed Forward Network with Two hidden layer

Like any supervised machine learning model, the FFN needs to be trained using labeled data. The training is in the form of optimizing the parameters by reducing the error between the output value and the true value. One such important parameter to optimize is the weight each neuron gives to each of its input signals. For a single neuron, the weight can be easily computed using the error.

However, when a set of neurons are collated in multiple layers, it is challenging to optimize the neuron weights in multiple layers based on the error computed at the output layer. The backpropagation algorithm helps to address this issue [15]. Backpropagation is an old technique which comes under the branch of computer algebra. Here, automatic differentiation is used to calculate the gradient that is needed in the calculation of the weights to be used in the network.

In a FFN, based on activation of each linked neuron, the output is obtained. The error is propagated layer by layer. Based on the correctness of the output with the final outcome, the error is calculated. This error is then in turn back propagated to fix errors of internal neurons. For each data instance, the parameters are optimized by going through multiple iterations.

V. FEATURE SELECTION

The first part of any deep learning model is gathering data and dataset generation. Here, We obtained the data from Habibi Lashkari et al. at the university of New Brunswick for data experiment in this paper. From that data number of feature is extracted and from that feature we are selecting some of the features for deep learning model. That contain extracted feature from traffic analysis os the university internet traffic. Extracted features from the PCAP file given in this table below:

Table 1 .Feature information and Explanation

Feature Name	Explanation of Feature
FIAT	Forward Inter Arrival Time, the time between two packets sent in forward direction(mean, min ,max, std)
BIAT	Backward Inter Arrival Time,the time between two packets sent backwards(mean, min, max, std)
FLOWIAT	Flow Inter Arrival Time, the time between two packets send in either direction(mean, min, max, std)
ACTIVE	The amount of time flow was active before going idle(mean, min, max, std)
IDLE	The amount of time flow was idle before becoming active(mean, min, max, std)
FB PSEC	Flow Bytes per second. Flow packets per second duration The duration of the flow.

Apart from this features all the flow based features are also selected and remove the common features like source ip/port ,destination ip/port that does not effect on the model' accuracy and take more processing time. After creating dataset of this features using this features in deep feed forward neural network with the N number of hidden layers. The hidden layer is varied from the 2 to 10.we found N=8 optimal. For activation of the function Relu is used each layer of hidden layer is dense in nature of dimension 100.

The output node is activated by sigmoid function. This was used the output is binary classification tor-nontor. Here, Keras and Tensorflow is used in the backend to train this model. The model trained for different epoch and this epoch is directly proposal to the accuracy of classification. Binary cross entropy loss was optimizing the FFN.

This below two graph shows that model performance is increased and loss value is decreased if we increase number of epoch.

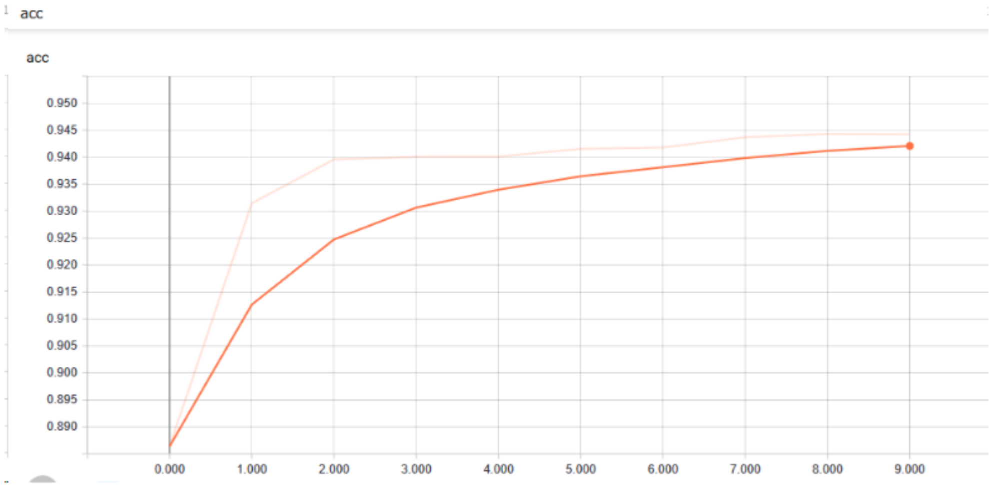

Figure 3 Graph of accuracy opposite number of epoch

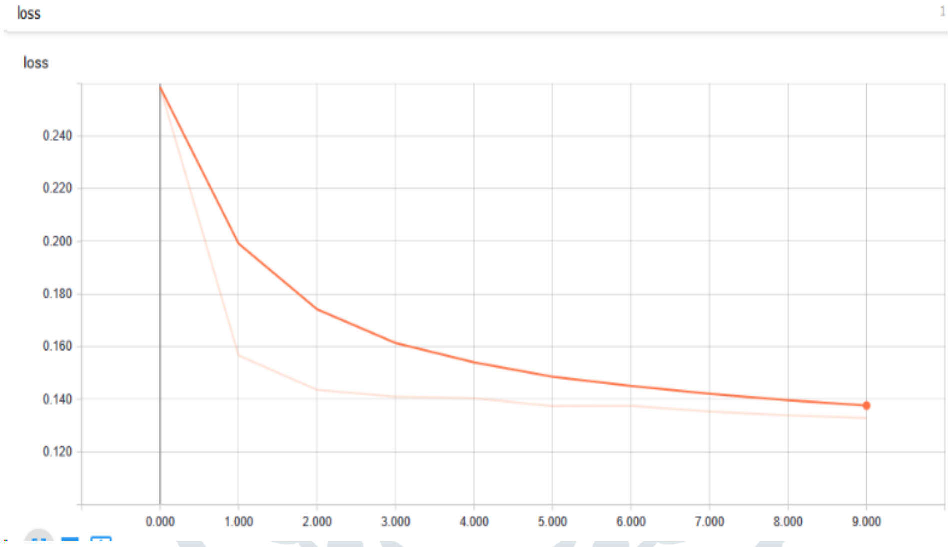

Figure 4.Graph of loss opposite number of epoch

The result of the deep learning algorithm is compared with various other estimators. From all the estimator DL based system work and detect TOR class well. However, it is the Non-tor class that we need to give more priority. DL based system reduce the false positive of Non-Tor category. The result are shown in below table:

Table 6.4 Comparison of algorithm

Algorithm	Accuracy
Naive Bayes	81.02%
Logistic Regression	93.86%
Support Vector Machine(SVM)	93.40%
Random Forest	98.26%
Deep Learning	98.27%

Among this all classifiers Random forest and Deep learning perform better than the rest. The result is shown based on 55k instances. The dataset used in this experiment is comparatively a smaller a dataset for the DL-based systems .As the training dataset increased performance of the model increased for both DL based system and Random Forest classifier. However, DL based classifier performs very well for larger dataset where ML based system takes larger resources to retrain the dataset.

VII. CONCLUSION

In this paper we present tor traffic classification using deep learning based techniques and use time based features to detect the traffic .current approaches use ML based system that is not more efficient in some kind of situations and not give accurate result every time. The result obtained prove that DL based system is more accurate than the ML based system. DL based system take less significant resources to retrain the model while ML based system required more resources to retrain the parameter.

REFRENCES

[1] Ghafir, Ibrahim, Jakub Svoboda, and Vaclav Prenosil. "Tor-based malware and Tor connection detection" *IEN* (2014): 19-6

[2] Almubayed, Alaeddin, Ali Hadi, and Jalal Atoum. "A model for detecting tor encrypted traffic using supervised machine learning." *International Journal of Computer Network and Information Security* 7.7 (2015): 10-23.

[3] Bai, X., Zhang, Y., and Niu, X. (2008). Traffic identification of tor and web-mix. In2008 Eighth *International Conference on Intelligent Systems Design and Applications*, volume 1, pages 548–551.

[4] Chakravarty, S., Barbera, M. V., Portokalidis, G., Poly- chronakis, M., and Keromytis, A. D. (2014). On the effectiveness of traffic analysis against anonymity networks using flow records. *PAM2014*,pages 247–257, NewYork, NY, USA. Springer-VerlagNewYork,Inc

[5] Chaabane, A., Manils, P., and Kaafar, M. A. (2010). Dig- ging into anonymous traffic: A deep analysis of the tor anonymizing network. In Proceedings of the 2010 Fourth *International Conference on Network and System Security*, NSS '10, pages 167–174, Washington, DC, USA. IEEE Computer Society.

[6] Abe, Kota, and Shigeki Goto. "Fingerprinting attack on tor anonymity using deep learning" *Proceedings of the Asia-Pacific Advanced Network* 42 (2016): 15-20.

[7] Lashkari, Arash Habibi, Gerard Draper-Gil, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. "Characterization of Tor Traffic using Time based Features." In *ICISSP*, pp 253-262. 2017.

[8] Cuzzocrea, Alfredo, Fabio Martinelli, Francesco Mercaldo, and Gianni Vercelli. "Tor traffic analysis and detection via machine learning techniques" In *Big Data (Big Data), 2017 IEEE International Conference on*, pp. 4474-4480. IEEE, 2017

[9] Murdoch, Steven J.,and George Danezis."Low-cost traffic analysis of Tor." *Security and Privacy, 2005 IEEE Symposium on*. IEEE, 2005

[10] Jansen, Rob, Marc Juarez, Rafael Galvez, Tariq Elahi, and Claudia Diaz. "Inside Job: Applying traffic analysis to measure Tor from within." In *Network and Distributed System Security Symposium (NDSS)*. 2018

[11] Overdorf, Rebekah, Mark Juarez, Gunes Acar, Rachel Greenstadt, and Claudia Diaz. "How Unique is Your. Onion?: An Analysis of the Fingerprintability of Tor Onion Services." In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp . 2021-2036. ACM, 2017.

[12] Pineda, Justin David, Arianne Wisdom Abinal, and Aliana Marie Lachica. "Proactive Response and Detection of TOR Anonymizers through Signature and Heuristic-based Mechanisms"

[13] Velan, Petr, et al. "A survey of methods for encrypted traffic classification and analysis." *International Journal of Network Management* 25.5 (2015): 355-374.

[14] Springer Berlin Heidelberg, Berlin, Heidelberg. Shmatikov, V. and Wang, M.-H. (2006). Timing analysis in low-latency mix networks: Attacks and defences. In Proceedings of the 11th *European Conference on Research in Computer Security, ESORICS'06,* pages 18–33, Berlin, Heidelberg. Springer-Verlag.

[15] "Deep Learning," Ian Good fellow, Yoshua Bengio, Aaaron Courville; pp 196, MIT Press, 2016.