FastPacket: Towards Pre-trained Packets Embedding based on FastText for next-generation NIDS

Khloud Al Jallad (k.jallad.l@gmail.com)
Arab International University https://orcid.org/0000-0001-9474-9204

Research Article

Keywords: Intrusion Detection System(IDS), Packet Embedding, NIDS, MAWI dataset

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-555961/v1

FastPacket: Towards Pre-trained Packets Embedding based on FastText for next-generation NIDS

Khloud Al Jallad*

*Correspondence: k-aljallad@aiu.edu.sy

Faculty of Information Technology, Arab International University, Daraa, Syria

ABSTRACT

New Attacks are increasingly used by attackers every day but many of them are not detected by Intrusion Detection Systems as most IDS ignore raw packet information and only care about some basic statistical information extracted from PCAP files. Using networking programs to extract fixed statistical features from packets is good, but may not enough to detect nowadays challenges. We think that it is time to utilize big data and deep learning for automatic dynamic feature extraction from packets. It is time to get inspired by deep learning pre-trained models in computer vision and natural language processing, so security deep learning solutions will have its pre-trained models on big datasets to be used in future researches. In this paper, we proposed a new approach for embedding packets based on character-level embeddings, inspired by FastText success on text data. We called this approach FastPacket. Results are measured on subsets of CIC-IDS-2017 dataset, but we expect promising results on big data pre-trained models. We suggest building pre-trained FastPacket on MAWI big dataset and make it available to community, similar to FastText. To be able to outperform currently used NIDS, to start a new era of packet-level NIDS that can better detect complex attacks.

Keywords: Intrusion Detection System(IDS), Packet Embedding, NIDS, MAWI dataset.

1. Introduction

A pre-trained model is a saved network that was previously trained on a big dataset, we can either use the pre-trained model as is or use transfer learning to customize this model to a given task. The intuition behind transfer learning is that if a model is trained on a large and general enough dataset, this model will effectively serve as a generic model. We can take advantage of these learned feature maps without having to start from scratch by training a large model on a large dataset.

Although its great success in natural language processing [1] and image processing [2], pre-trained models are not yet used in security. In this paper, we suggest to start a new era of security solutions based on pre-trained models.

Till now, IDS heavily rely on the discrete handcrafted features, while deep learning automatic features based on n-grams of raw pcaps may be better solutions to detect complex attacks.

1.1. IDS types

IDS in general has three basic types based on its location: Host IDS, Network IDS and Hybrid IDS.

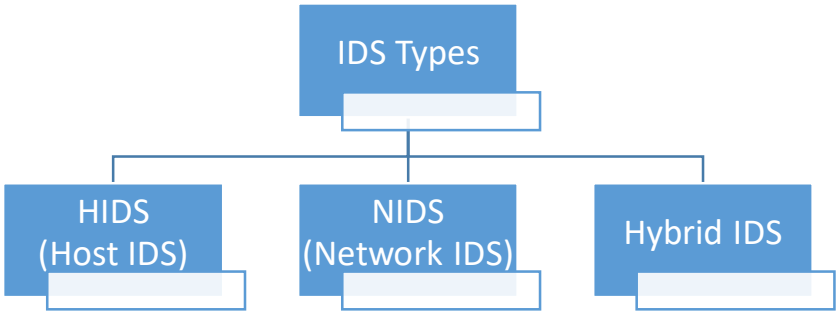

Figure 1 IDS types [3]

Network IDS is the domain of this experiment, so we will talk about in more details.

NIDS Hierarchy shown in figure 2. NIDS has two basic types based on the data source that it is monitoring.

- **Log-based NIDS**: that analyzes logs written by security devices when packets flow.
- **Raw Data-based NIDS** that analyzes the data sent itself, it has two types

- **Traffic-based**: that contains the whole packets' data, headers and bodies.
- **Flow-based**: that contains only headers of packets.

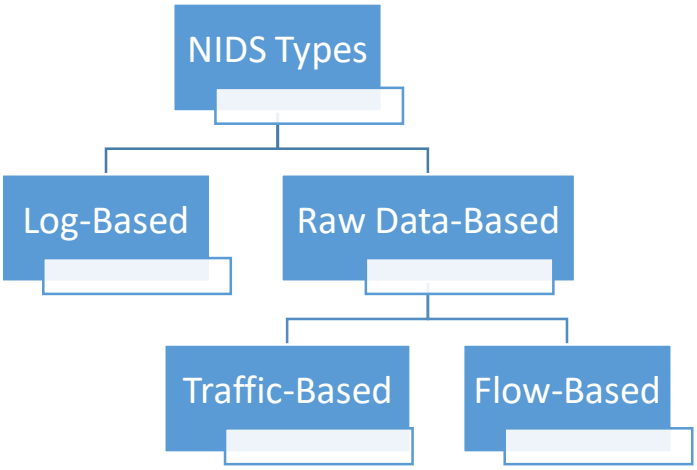

Figure 2 NIDS Types [4]

As for Traffic-based (packet-level NIDS), also called Deep Packet Inspection (DPI) or traditional packet-level NIDS, it is considered time-consuming when it comes to big data networks (more than 1 TB in second) or it will need a high cost of needed servers for just a very small optimization in performance, so we have to decide a tradeoff cost and accuracy. Some researchers choose to filter some packets to reduce costs [5]

As for Flow-based (flow-level NIDS), also considered Behavioral Analyzer NIDS, the body of each packet is ignored, only headers of packets are used to extract tuples. Each tuple has five values Source IP, Destination IP, Source Port, Destination Port, Protocol. Flow-level is better than packet-level in big networks when it comes to the cost of processing and storage, as it has very less cost because it processes only headers without bodies. Flow-based data approach is very lightweight. Storage issues that appeared in the packet-based approach are almost disappeared, but some types of attacks that are injected in bodies cannot be detected by analyzing headers only. [5]

In general, Flow-level NIDS uses anomaly-based detection methods and packet-level NIDS uses signature-based detection methods. Each type has its advantages and disadvantages. Therefore, we need to tradeoff high cost or high false positive. Some researchers believe that a combination of both is the best solution. [5]

Flow-Level NIDS needs less performance cost than Packets-Level NIDS, However, lots of critical information are dropped when body of packet is ignored, especially information related to some

critical attacks, such as, SQL injection attacks, phishing attacks and Trojans that seems normal when only analyzing header information.

In this research we propose to combine them in a pre-trained model to overcome the cost and time-consuming process of training and get the advantage of better accuracy.

This paper is organized as follows, we will talk about related works in "Related Work Section". The proposed method is explained in detail in "Method Section". "Data Section" contains detailed information about used dataset. We will discuss results in "Results and Discussion Section". We will talk about conclusion and future vision in "Conclusion and Future Work Section".

2. Related Work

Most recent previous works on NIDS used SVM and LSTMs on hand-crafted features for anomaly detection NIDS. [6]

Few recent articles trying to embed packets inspired by famous NLP Word-Embedding.

LogBERT [7] for analyzing anomaly logs based on BERT for text. They have extracted some regular expressions from logs then embed results using BERT embeddings.

Packet2Vec [8] utilized w2vec for packets embedding on DARPA2009 dataset and their results outperformed state-of-the-art results on same dataset.

Both papers are great and inspired us to do this article. But, although both used embedding, no encoding was done before embedding to reduce alphabet size thus processing cost, Moreover Applying NLP famous embedding is a great idea but we think that it is very expensive to always start from scratch when it comes to the cost of processing and storage. We argue that it is essential to have security special pre-trained anomaly models, not only train NLP embedding on a dataset used for research.

3. Data

The dataset used for this experiment is CIC-IDS-2017 dataset [9] that contains benign and some common attacks, which resembles the true real-world data (PCAPs).The data capturing period started at 9 a.m., Monday, July 3, 2017 and ended at 5 p.m. on Friday July 7, 2017, for a total of 5 days. Monday is the normal day and only includes the benign traffic. The implemented attacks include Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and

DDoS. They have been executed both morning and afternoon on Tuesday, Wednesday, Thursday and Friday. Some basic information of data is shown in table1

Table 1 CIC-IDS-2017 Dataset Basic Information

Day	Description	Size (GB)
Monday	Normal Activity	11.0G
Tuesday	Attacks + Normal Activity	11G
Wednesday	Attacks + Normal Activity	13G
Thursday	Attacks + Normal Activity	7.8G
Friday	Attacks+ Normal Activity	8.3G

However, the dataset we propose to be used in pre-trained model is MAWI dataset [10]MAWI stands for Measurement and Analysis on the WIDE Internet [11] [12] [13]. It is the biggest online available and most real-life dataset that is publically available for free. It contains real-life traffic data of Japan-US cable. It is collected and preprocessed by a sponsor of the Japanese ministry of communication

MAWI is labeled by MAWILAB project [14] Which is a project done on top of MAWI archive that contains labels of data, and it is updated automatically every day. Labeling data is done by community of four classifiers. Classifiers are Principal component analysis (PCA), Gamma Distribution, Hough Transform, Kullback–Leibler (KL). Labels are tagged according to class of majority classifiers detection. That help reducing false positive rate. Labels of MAWILAB are done according to taxonomy of anomalies in network traffic [15]

4. Proposed Method

Instead of creating hand-crafted features for each packet, we proposed to encode raw packet data then embed it using FastText vectorization for each packet, then perform classification.

Specifically, our approach has the following steps: We proposed encoding packets' raw content in hex decimal encoding to have limited character alphabets for character-level embeddings. Then we do FastText supervised learning on hex-encoding to build a model that can take n-grams of raw content into consideration as many of attacks are injected in packet payload, such as SQL injections, Phishing attacks and Trojans. After applying Fasttext embedding, we have applied several traditional machine learning algorithms on subsets of used dataset. We chose traditional machine learning approaches as our experiments are done on small datasets only because of hardware limitations. But we think that pre-trained models will be better used with convolutional or sequential deep models

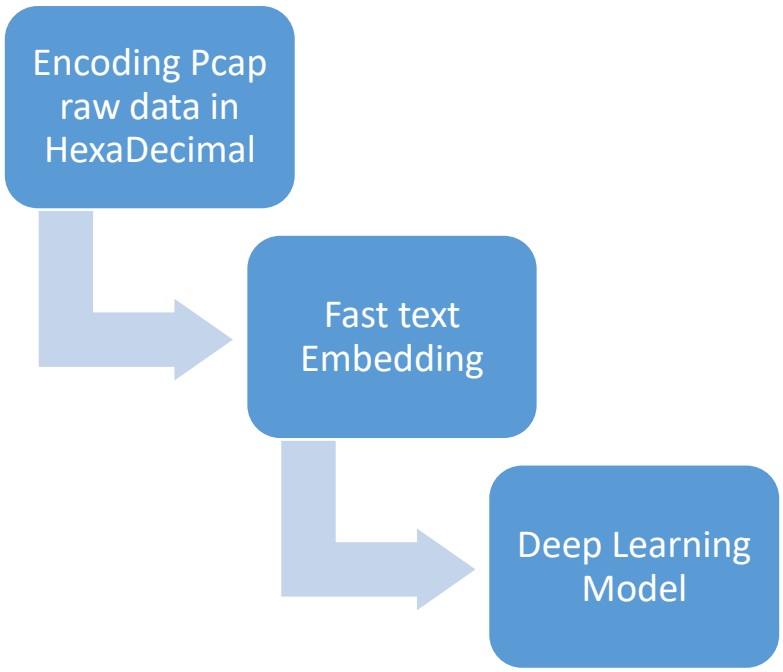

Figure 3 Proposed Solution

5. Proposed Framework and Libraries

5.1. Colab

Colaboratory is a free research tool offered by Google, for machine learning education and research. It's a Jupyter notebook environment that requires no setup to use. Code is written on browser interface. Code is executed in a virtual machine dedicated to user account(options available now are CPU, Graphical Processing Unit GPU, Tensor Processing Unit TPU) [16]

5.2. FastText

we build our code based on FastText [17] [18] FastText is an open-source, free, lightweight library that allows users to learn text representations and text classifiers. We have used supervised learning approach; thus an attack pattern can be learned in embeddings.

5.3. Scapy

we have used scapy [19] for reading and manipulating packets' raw data. Scapy is a powerful interactive packet manipulation program written in python language.

6. Results and Discussion

We chose some traditional machine learning algorithms such as random forest and SVM as it is the most widely used classifier in security researches. We think that building pre-trained model will enable us to get the advantage of higher accuracy by using packet-level NIDS, with avoiding the packet-level NIDS disadvantage of high time and cost needed to train model on each dataset. But our results are not accurate as we compare random subsets of dataset, thus we will not share numbers or charts. But, based on basic research and analysis done on data, we expect promising results on big datasets, therefore we share our insights so other researchers who have better hardware infrastructure can create a pre-trained model dataset and make available to community to be used in future researches.

Raw payload is very important to take into consideration and better than hand-crafted features because many attacks are injected in packet payload, such as SQL injections, Phishing attacks and Trojans.

Encoding raw packets in Hex Encoding can be useful to limit character alphabet size instead of ASCI or UNICODE basic encoding of packets.

We chose traditional machine learning approaches as our experiments are done on small datasets only because of hardware limitations. But we think that big pre-trained models will be better used with convolutional or sequential deep classifiers.

Although the number of hacking attacks is growing exponentially every few months, a general pattern can be automatically extracted by pre-trained anomaly models on big datasets. Same as we see in computer vision and natural language processing that have a great accuracy detection

rates despite the wide diversity of texts and images that are growing also exponentially thank to big data pre-trained models.

We recommend using MAWI dataset [10] for pre-trained model as it is the biggest and most organized available online PCAP dataset for anomaly-IDS tasks.

7. Conclusion & Future Work

Similar to the huge impact that anomaly-based IDS outperform signature-based IDS in detecting new threats, Automatic deep anomaly pre-trained models are promising to outperform deep hand-crafted anomaly-IDS, because of its ability to detect new threats injected in raw packet contents. We proposed building pre-trained FastPacket models on big datasets, so we can get the advantage of higher accuracy by using packet-level NIDS, with avoiding the packet-level NIDS disadvantage of high time and cost needed to train model on each dataset.

We discuss our claim the reason behind our hypothesis but we were not able to do a complete experiment on a big dataset and create a pre-trained model because of hardware limitations as this type of experiments needs big companies' hardware capabilities to be implemented. The experiments we did on random small subsets of dataset were promising but not enough to prove our hypothesis. Therefore, we share experiment and our future vision thoughts and we wish that complete experiment will be done in future by other interested researchers who have better hardware infrastructure than ours. We wish that a university or a big data company create a pre-trained anomaly model on PCAP files of MAWI big dataset and make it available to community for future research to start a new era of security intelligent solutions.

Abbreviations

IDS: Intrusion Detection System,

Authors' contributions

KHJ took on the main role so she performed the literature review, conducted the experiments, analyzed results and wrote the manuscript

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All datasets in this survey are available online, you can find links in references.

Funding

The authors declare that they have no funding.

References

[1] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai and X. Huang, "Pre-trained Models for Natural Language Processing: A Survey," *arXiv.org,* 2020.

[2] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan and M. Shah, "Transformers in Vision: A Survey," *arxiv.org,* 2021.

[3] B. M, "A survey on secure network: intrusion detection & prevention approaches," *Am J Inf Syst,* vol. 4, no. 3, p. 69–88, 2016.

[4] A. J. Khloud, A. Mohamad and M. S. Desouki, "Big data analysis and distributed deep learning for next-generation intrusion detection system optimization," *Journal of Big Data,* vol. 6, no. 88, 9 9 2019.

[5] A. H, M. M and M. AA., "An overview of flow-based and packet-based intrusion detection performance in high speed networks.," 2017.

[6] K. A. Jallad, M. Aljnidi and M. S. Desouki, "Anomaly detection optimization using big data and deep learning to reduce false-positive," *Journal of Big Data,* vol. 7, no. 68, 2020.

[7] H. Guo, S. Yuan and X. Wu, "LogBERT: Log Anomaly Detection via BERT," *arXiv,* 7 March 2021.

[8] E. L. Goodman, C. Zimmerman and C. Hudson, "Packet2Vec: Utilizing Word2Vec for Feature Extraction in Packet Data," *arxiv,* April 2020.

[9] C. I. f. Cybersecurity, "Intrusion Detection Evaluation Dataset (CIC-IDS2017)," 2017. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html.

[10] C. K. M. W. G. T. Archive., "WIDE Project," 2011. [Online]. Available: http://mawi.wide.ad.jp/mawi/. [Accessed 2021].

[11] "WIDE Project," 1988–2018.. [Online]. Available: http://www.wide.ad.jp/.. [Accessed 2021].

[12] J. a. b. t. E. U. S. F. P. International Collaborative R&D Promotion Project of the Ministry of Internal Affairs and Communication, "necoma-project (Nippon-European Cyberdefense-Oriented Multilayer threat Analysis)," 2013. [Online]. Available: http://www.necoma-project.eu/.

[13] M. Wählisch, "Measuring and implementing internet backbone security: current challenges, upcoming deployment, and future trends," 2016.

[14] "MAWIlab," [Online]. Available: http://www.fukuda-lab.org/mawilab/.

[15] M. J, F. R and F. K. A, "taxonomy of anomalies in backbone network traffic," in *5th international workshop on TRaffic analysis and characterization (TRAC)*, 2014.

[16] Google, "Colab," [Online]. Available: https://colab.research.google.com/.. [Accessed 2021].

[17] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching Word Vectors with Subword Information," *arXiv,* 2016.

[18] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *arxiv,* Aug 2016.

[19] P. B. a. t. S. community, "scapy," 2008-2021. [Online]. Available: https://scapy.readthedocs.io/en/latest/introduction.html. [Accessed 2021].