COMP 6961

Graduate Seminar Report

Anomaly Detection in IoT Devices using LogBERT

Speaker – Eniela Vela

Date: October 25th, 2022

Submitted To: Denis Pankratov

Rajat Sharma

40196467

ABSTRACT

IoT devices (internet of things devices) term was coined by Kevin Ashton. These devices are computer devices but not standard ones, and they can transmit data such as sensors that collect and exchange data. The number of IoT devices used is around 10 billion, and the number is expected to increase up to 30 billion in 2030. 20% of the companies using IoT devices were victims of cyberattacks in 2020. Some MachineLearning approaches were adopted to solve this issue of cyberattacks; some of them were KNN with an F1 score of 84.82%, SVM with 100% accuracy, and K-means with 80% accuracy. The speaker here is trying to use the LogBERT algorithm to improve the F-1 score for anomaly detection. She further moves on to tell us about the types of anomaly detection and how LogBERT will be effective in that.

She explains that LogBERT is an unsupervised machine learning technique that uses LogParser(DRAIN+SPELL) and the famous BERT algorithm from Google. The BERT algorithm uses Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). A total of six top IoT devices were used to generate data. Then a comparison was performed in IoT devices with PCA, Isolation Forest, One SVM, and Logcluster. After training was done on the LogBERT algorithm was able to distinguish between the anomaly and was able to identify what abnormal part was there which was in deviation from the standard learning path curve.

It was shown from the results that the F1 score drastically increased by about 20% when the abnormal and normal dataset was taken in a 1:9 ratio which is the same used in the LogBERT algorithm. The best performance with LogBERT was shown by ThunderBird, and the IoT device used was an indoor camera. The LogBERT algorithm is better than PCA, OSVM, and Isolation Forest but compared with LogCluster. So it is as good as LogCluster but better than the other three techniques. But the recall value for LogBERT is better in 5 IoT devices. Also, the speaker showed that energy consumption could be somewhat an indication of anomalies in IoT devices.

Lastly, she talked about how only six IoT devices were used for this research, and the cumulative F1 score was around 75%. The study was conducted on a real dataset. The future work includes adding the time data feature or IP and also finding out how well LogBERT performs on data that is gathered for a couple of weeks or months.

TABLE OF CONTENTS

List of Abbreviations

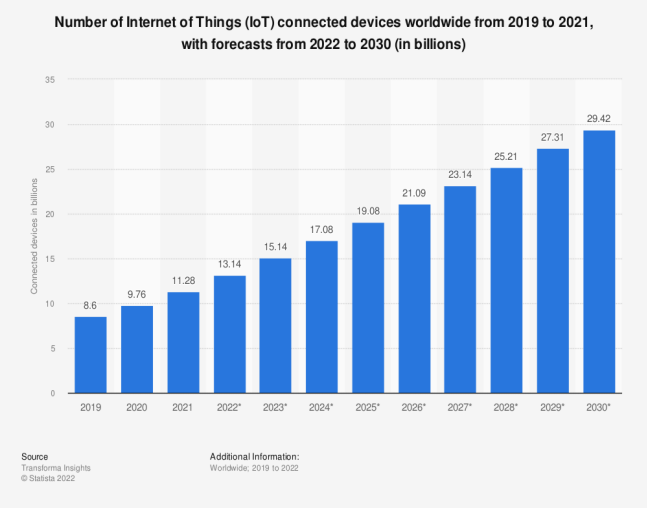
- IoT - Internet of Things
- MLM: Masked Language Modeling
- NSP: Next Sentence Prediction
- LAN: Local Area Network
- XML: Extensible Markup Language
- CSV: comma-separated values
- KNN: K nearest neighbors
- SVM: Support Vector Machine

INTRODUCTION

IoT devices (internet of things devices) term was coined by Kevin Ashton. These devices are computer devices but not standard ones, and they can transmit data such as sensors. They collect and exchange data. The number of IoT devices used is around 10 billion, and the number is expected to increase up to 30 billion in 2030. 20% of the companies using IoT devices were a victim of cyberattacks in 2020. Some MachineLearning approaches were adopted in order to solve this issue of cyberattacks; some of them were KNN with an F1 score of 84.82%, SVM with 100% accuracy, and K-means with 80% accuracy. The speaker here is trying to use the LogBERT algorithm to improve the F-1 score for anomaly detection. She further moves on to tell us about the types of anomaly detection and how LogBERT will be effective in that.

Problem Statement:

Anomaly detection in IoT devices is the practice of looking for and identifying unusual patterns in the data generated by connected devices that may be an indicator of malicious behavior or system faults. Finding any potential malicious behavior or system weaknesses is the goal in order to prevent damage to connected devices or network disruption. This necessitates the creation of efficient techniques for spotting abnormalities and acting upon them quickly.



The graph shows the number of IoT devices in the year 2020 and gives a projection of 2030

According to the thesis, there will be roughly 10 billion IoT devices on the market in 2022, and approximately 20 percent of devices will be under cyber attack, which is proved by Gartner in 2020 in his research. According to Vailshery's assumption, the numbers are likely to increase over the ensuing years, rising by three times in 2030. There are currently almost 4.8 billion Internet of Things (IoT) devices in use, and they are permeating every aspect of life and business. Because 57% of these devices

are susceptible to attacks of high or medium severity, they make attractive targets for attackers. The availability of these gadgets leaves huge openings for attackers to capitalize on and pivot to larger targets inside your organization. Maintaining security requires an understanding of the risk that IoT and other endpoints pose to your business. It is not just a matter of how serious these vulnerabilities are. It is most frequently the exploitability that is at issue. Digital criminals can link low-impact attacks to develop footholds in your infrastructure that they can exploit.

Literature Review:

· Pajouh's research in 2016 stated that F1-score with the KNN model is 84.82%, while the Decision tree delivers 81.05% to detect anomalies which is pretty good. However, Eniela noticed that Pajouh used an obsolete dataset from 1999. Hence the relevance to today's problem can be questionable.

· Loannou's research of 2019 showed a 100% f1 score with the SVM model using IoT data generated by blockhole and sinkhole attacks, but the researcher did not publish his dataset, so I can not judge and compare and test its performance. Moreover, the researcher used only two attack data which is entirely unreliable for other attacks to identify the abnormality.

· A MacDermott in 2019 proved that he got 80% accuracy using the K-mean model based on a healthcare IoT dataset. But, they used Synthetic LAN data, and in research, they are claiming IoT dataset.

Objectives:

LogBERT is an unsupervised machine-learning technique that provides good results in computer log collections.

LogBERT is frequently used in computer contexts, such as Hadoop, Bluegene, and ThunderBird, which have respective F1 scores of 82.32%, 90.83%, and 96.64%. No one to date tested LogBERT with IoT devices; therefore, with the goal of some great outcome, she finalized to apply an algorithm to the gadget.

RESEARCH METHODOLOGY

Understanding LogBERT:

LogBERT is a unique algorithm presented by academics at Stanford University. It is a natural language processing (NLP) algorithm that can classify text data into multiple groups. The BERT model and

LogParser serve as the foundation for LogBERT, which use a different method for learning linguistic representations. The BERT model's weights are transformed logarithmically by the algorithm to reduce the number of parameters, which leads to shorter training times and more reliable performance. LogBERT also may capture long-term dependencies in text, which is not achievable with the original BERT paradigm.

As the below example says, the anomaly can be discovered by LogBERT. Red, Blue, Yellow, etc., are a type of colors, whereas the student has no relation with color. Therefore the student is the outcome of anomaly detection.

A tool called LogParser is used to separate structured information from unstructured log data. It is used to build organized data sets from logs, which can then be used for further research. LogParser can parse text files, XML files, CSV files, and other forms of log files. BERT(Bidirectional Encoder Representations from Transformers) is a machine-learning technique for natural language processing (NLP) applications. It is a pre-training strategy that uses unsupervised learning to learn language representation from unlabeled input. BERT can be used to develop strong models for many NLP tasks.

DRAIN and SPELL are two components of the LogParser component in the LogBERT algorithm. DRAIN stands for Deep Relation Analysis and INference, and it is used to identify the relationships between words in a text. It employs rule-based and deep-learning algorithms to identify items, relations, and events in the input language. SPELL stands for Semantic Parser with Linking and Entailment. It combines a combination of semantic parsing techniques and deep learning techniques to identify things and the relations between them. It also employs techniques for natural language inference to find entailment connections between entities. For instance, DRAIN would recognize the terms "cat" and "mouse" as well as the relation "ate" between them if the input text was "The cat ate the mouse." SPELL would then use natural language inference techniques to infer that the cat is consuming the mouse.

Masked Language Modeling (MLM) is a technique employed by the LogBERT algorithm to assist the model in better grasping natural language by guessing a missing word or phrase in a sentence. This is done using masked language modeling, where some words in the sentence are replaced by a unique token (e.g., [MASK]) (e.g., [MASK]). Instead, they use unsupervised techniques to learn from data. The LogBERT architecture employs self-supervised tasks to learn representations from unlabeled data. For instance, LogBERT learns from text using a self-supervised task known as Masked Language Model (MLM). In this task, some words in the text are randomly masked, and the model is then asked

to predict what those words are. This allows the model to learn from the context of the words that are still visible and build representations for the words that are masked.

The Transformer encoder in LogBERT architecture is a sort of deep learning method that helps to capture long-term dependencies in natural language processing models. This type of encoder is based on transformer design and is used to process input sequences. The transformer encoder consists of numerous layers, each layer consisting of a self-attention mechanism, a feed-forward neural network, and a residual connection. The self-attention tool allows the model to recognize which words are most relevant to each other, while the feed-forward neural network helps to learn the non-linear structure of the input sequences. The residual link works as an information flow that allows the model to communicate information between layers. For instance, in LogBERT, the log data is processed using the transformer encoder to look for patterns and abnormalities. The model analyses each log entry as a sequence of words and then uses the transformer encoder to learn the long-term dependencies between the words. The model then employs the self-attention mechanism to select which words are most significant to each other and the feed-forward neural network to learn the non-linear structure of the input sequences. Finally, the residual link helps to transfer information across the levels of the transformer encoder, allowing the model to learn from the data and produce more accurate predictions.

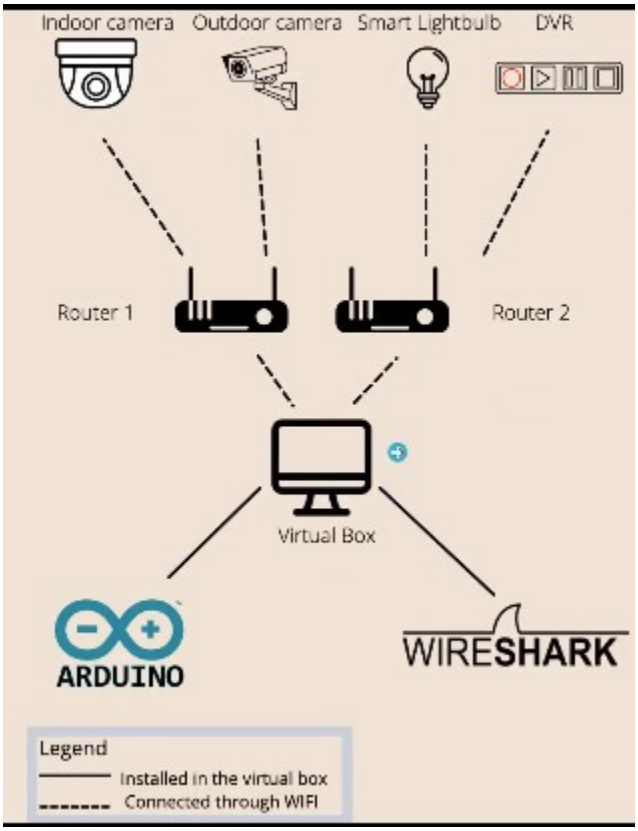
The Log key sequence in LogBERT is a memory-augmented neural network that enables the model to store, retrieve, and update the knowledge it has learned over time. This approach allows the model to preserve an internal representation of the data, which can be utilized to boost its performance. For example, the Log key sequence can be used to assist LogBERT in learning the associations between objects in a digital image. In this scenario, the Log key sequence will retain the associations between the things in the image and allow LogBERT to recall and update them as it analyses more data. This will help the model recognize complicated patterns and enhance its accuracy.

Data Generation:

The speaker chose to conduct a test using the top 6 IoT devices purchased from BestBuy in order to provide findings that were as realistic as possible. These devices were an Indoor camera, Outdoor camera, Smart light bulb, DVR, and two routers of different companies. All the attacks were performed in the virtual box connected to ARDUINO(to calculate the electric current for each IoT device) and WIRESHARK(to network traffic data), shown in the picture below.

She enlisted the aid of the RouterSploit, UFONet, and Mirali Botnet, which are essentially attack types that this framework offers, such as a massive amount of data packets to exhaust the device, exploit the vulnerable device, and Dos and DDos attacks appropriately.

Finally, the dataset was constructed by combining network traffic data and energy data with two kinds of abnormal state and normal state.



Data Analysis:

If I look at an example of log data, I get a network traffic dataset. Since LogBERT will use the information section, it appears comparable in this experiment. The speaker employed data packets from network traffic.

Data analysis in machine learning entails utilizing several approaches to extract useful information from data collections. In addition to building predictive models, it may be used to identify patterns and trends in data. Data analysis in machine learning often encompasses preprocessing, data preparation, feature engineering, model selection, hyperparameter tuning, evaluation, and deployment.

In my situation, she discovered that LogBERT uses information from log data, which are packets of network traffic. I would use Network Traffic Data packets for my anomaly detection.

Data Cleaning:

Data cleaning in machine learning is preparing data for modeling and decision-making. This includes locating and fixing erroneous or incomplete records, eliminating excessive or redundant data, and formatting data consistently. An example of an empty field is a column in a database table that has no values. To fully clean the data, the open field must be discovered, and remove the whole row from the dataset. By doing this, you can help verify that the data is accurate for every record and avoid mistakes when using machine learning algorithms.

In my dataset, the researcher detected 0.1% empty fields and eliminated rows that contained empty fields.

Data Grouping:

Data grouping in machine learning is the process of organizing data into groups or clusters that share similar characteristics. The goal is to identify patterns in the data and use those patterns to make predictions or decisions. Clustering algorithm can discover groups in this data and assign each data point to one cluster. This enables machine learning algorithms to learn more effectively since the data points within a cluster share similar characteristics. Data grouping can be done manually or automatically with various clustering algorithms.

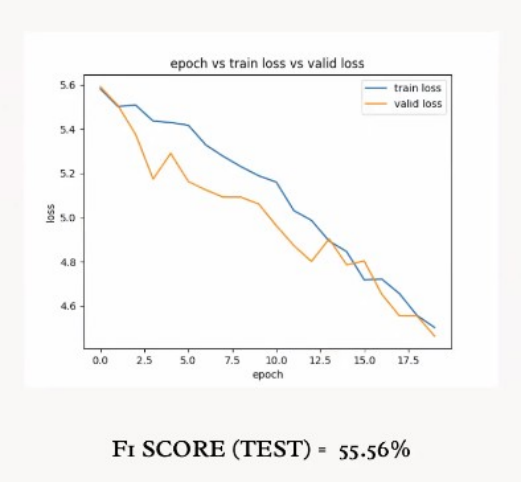
For example, if the dataset contains records with the same source and destination IP, clustering algorithms can group those records. This can help identify patterns in the data and gain insights into network traffic.

Data Process:

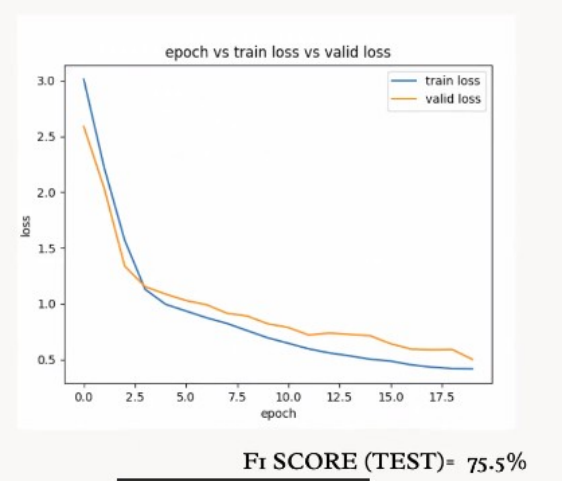
From the Processed data, she developed an Anomaly label partition and Event sequence. Event Template provides a contribution to developing Event sequence. Finally, Structured data was produced with the help of Event sequence and Event Template.

DRAIN and SPELL are done on data to get the pattern. The event sequence was done by the window sliding method. Event sequence is the most important factor for the LogBERT method because if I give it row data, it might not grasp and produce incorrect results.

TECHNICAL CONTRIBUTIONS AND RESULTS



F1 SCORE (TEST) = 55.56%



F1 SCORE (TEST)= 75.5%

Results:

- The aforementioned photos demonstrate how the f1-score increases as the dataset's normal data increases. This try-and-error strategy I used to discover the best-balanced dataset. A pass over the complete training dataset in machine learning is known epoch.
- After each epoch, the model is trained, and its weights and biases are modified to reduce error for the current epoch. Until the error is minimized or until a predetermined number of epochs have been completed, this process is repeated.
- In machine learning, the training loss represents the model's performance on a training dataset. It is determined by comparing the outputs expected and predicted by the model for each sample in the training dataset. The model performs better on the training dataset, the smaller the training loss value is.
- A model's ability to correctly forecast brand-new data that it has never seen before is measured by something called validation loss in machine learning. It is obtained by calculating the difference between projected values and the actual values from a validation dataset. It is a crucial indicator for evaluating the performance of a model since it gives an indication of how effectively the model generalizes to unknown data.
- Hadoop, BGL, and ThunderBird each received F1 scores of 82.32, 90.83, and 96.64, respectively. IoT devices, however, have F1 ratings of roughly 80, 72, 75, 80, 72, and 69. It has a lot of similarities to computer hardware.

The researcher presented the same dataset to many models to check to compare with each other. As a consequence, LogBERT seems better compared to another model in the above graphic.

Energy usage is roughly twice as when an IoT device is under an attack state compared to normal conditions.

I can demonstrate if the equipment is under cyber attack or not with energy consumption. Energy consumption is double compared to usual consumption while under cyber assault.

CONCLUSION AND FUTURE WORK

CONCLUSION

The increased complexity of IoT systems and their requirement for strong security can be managed through anomaly detection in IoT devices using LogBERT. LogBERT is a robust tool for finding anomalies in IoT logs and gives an excellent solution to defend IoT-based systems from malicious attackers. By harnessing the power of transfer learning and pre-trained transformer models, LogBERT can instantly detect anomalies in IoT logs while allowing for customization to cater to the particular problems of each IoT system. LogBERT has the potential to turn into a crucial tool for overseeing IoT system security with more work.

FUTURE WORK

Add IP or time-related data to LogBERT to make it better. Additionally, construct a different machine learning assessment technique for energy utilization and execute it in parallel with an anomaly detection method as additional protection. Examine LogBERT's performance over a longer time frame, like a day or month.

REFERENCES:

1.) IoT anomaly detection methods and applications: A survey Author links open overlay panelAyan,ChatterjeeaBestoun S.Ahmed (https://www.sciencedirect.com/science/article/pii/S2542660522000622)

2.) A Comprehensive Study of Anomaly Detection Schemes in IoT Networks Using Machine Learning Algorithms by Abebe Diro 1ORCID,Naveen Chilamkurti 2ORCID,Van-Doan Nguyen 2,*ORCID andWill Heyne 3 (https://www.mdpi.com/1424-8220/21/24/8320)

3.) Machine learning for Internet of things anomaly detection under low-quality data

Shangbin Han https://orcid.org/0000-0002-1976-7856 hanshangbin@buaa.edu.cn, Qianhong Wu, and Yang Yang

4.) Deep Learning Anomaly Detection for Cellular IoT with Applications in Smart Logistics Milos Savic, Milan Lukic, Dragan Danilovic, Zarko Bodroski, Dragana Bajovic Member, Ivan Mezei Senior Member, Dejan Vukobratovic Senior Member, Srdjan Skrbic and Dusan Jakovetic

5.) IoTDefender – IoT Anomaly detection Sustainable Security and Safety for Future Technologies (https://opens3-lab.com/projects/iotdefender-iot-anomaly-detection/)