Deep Learning for Malware Analysis: A Review

Nitish A[1], Hanumanthappa J[2]
[1,2]*Dept. of Studies in Computer Science, University of Mysore*
(Email: nitish.anantha@gmail.com)

Abstract—Malware Analysis is a process of analyzing an executable for the detection of malicious instructions that could compromise the security (confidentiality, integrity and availability) of a software system; it also helps in the vulnerability assessment of a system. This approach, being data intensive, takes the advantage of mining and learning techniques to produce efficient signatures that can be used against the future attacks. With the advancements in the technology, the efficient, yet resource intensive deep learning methods can be employed in malware analysis. We claim that deep learning methods can provide better performance in malware analysis compared to other approaches; the literatures discussed here also support the claim.

Keywords—*cybersecurity; deep learning; data mining; malware analysis; machine learning; neural networks*

I. INTRODUCTION

Deep Learning (DL) is an area of artificial intelligence (AI) that deals with solving more abstract problems that are not strictly bound by formal mathematical rules. Providing solutions to such intuitive problems requires allowing computers to learn from experience and understand the system in terms of a hierarchy of concepts, with each concept de_ned in terms of its relation to simpler concepts. By gathering knowledge from experience, this approach avoids the need for human experts to formally specify all of the knowledge that the computer needs. The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simple ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. Hence this approach is called deep learning [1].

This approach has proven its potential mostly in computer vision applications since its inception. However, recent literature trends have shown that DL is also applicable for a variety of domains, given its advantages over any other techniques used before.

In this paper, we try to provide a proper scope of DL in the area of Cybersecurity, dealing specifically with the issues of Malware Analysis. A malware is any form of software that is designed to harm a user with malicious purposes, such as leaking data from a computer or deleting files, regardless of his or her intention. The number of softwares with malicious purposes has increased significantly in recent years, reaching almost 800 million, as shown in figure 1.

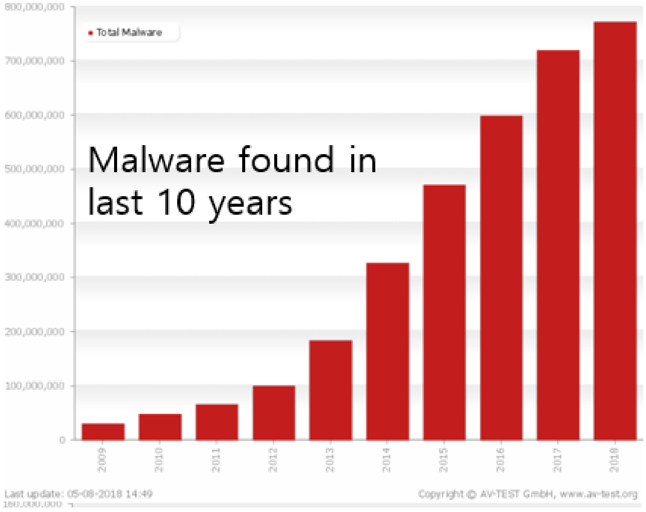

Figure 1: Increasing trend of malware in recent decade

Even though the new malwares that are being discovered each year have shown to be variable, it is safe to say that this will also show an increasing trend in the near future, given the advancements in technology and the increasing number of devices (figure 2) [2].

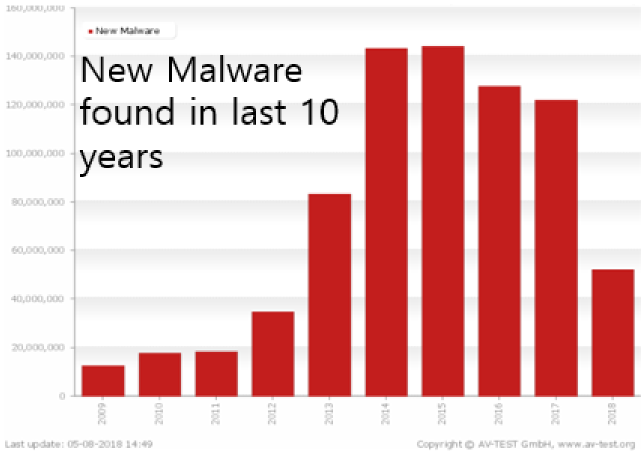

Figure 2: Trend of new malware discovery in recent decade

The prime motivation behind this literature survey is to present the importance of malware analysis and new breakthroughs in this area that has been possible because of the evolution of machine learning and deep learning techniques and the availability of huge audit data and related metadata in recent years.

II. BACKGROUND

The purpose of malware analysis is usually to provide the information required to respond to a network intrusion. The process involves determining the cause of an intrusion, locating all the infected system resources. While analyzing a resource for a suspected malware, understanding what the particular binary can do is of prime importance so that, the corresponding vulnerability exploited by the malware can be patched and the same signature can be stored for defense against future malwares, bearing same or similar signatures [3].

A signature is a sequence of operations involving all the resources (software and hardware) and system calls accessed in a particular order at a particular time, at varied levels of hierarchy, in order to perform an operation. These details are usually captured and stored as logs by the operating system itself. However specific signatures can be captured through customized rules according to the user requirement.

Malware analysis can be performed to obtain host based (at node level) and network based (at network level) signatures to obtain wide range of information, since the complete information required in handling malwares is spread across different levels. Host-based signatures (indicators) are used to detect malicious code on victim computers. These indicators often identify _les created or modified by the malware or specific changes that it makes to the registry. Unlike antivirus signatures, malware indicators focus on what the malware does to a system, not on the characteristics of the malware itself, which makes them more effective in detecting malware that changes form or that has been deleted from the hard disk. Network signatures are used to detect malicious code by monitoring network traffic. Network signatures can be created without malware analysis, but signatures created with the help of malware analysis are usually far more effective, offering a higher detection rate and fewer false positives.

After obtaining the signatures, the final objective is to understand how the malware works and take the necessary counter-measures.

A. Malware Analysis Techniques

While analyzing a malware, the available file will mostly be in the form of an executable, which will not be in human readable format. Hence it needs to be pre-processed to obtain a definite signature out of it. The malware analysis techniques can be broadly classified into four types based on the extent and the stages of analysis [3].

1. **Basic Static Analysis:** It involves examining the executable file without viewing the actual instructions. This approach can confirm whether a file is malicious, provides information about its functionality and sometimes produce simple network signatures.
2. **Basic Dynamic Analysis:** Involves running the malware and observing its behavior on the system to remove the infection and produce effective signatures. Malware is usually run in a safe environment called Sandbox, which mimics to a certain degree, the system environment under study (similar to a virtual machine) to make sure that the actual system remains unaffected by the malware.
3. **Advanced Static Analysis:** It involves reverse engineering the malwares internals by loading the executable into a disassembler in order to extract its effects on a sandbox environment to build an effective signature involving various levels of system hierarchy.
4. **Advanced Dynamic Analysis:** Uses a debugger to examine the internal state of running malware, providing much detailed information for producing signatures.

Even though the classification provides definite boundaries between the functions to be performed, analysis of recent advanced malwares, based on the functions they perform at different levels and systems they target, involves performing multiple, if not all the techniques discussed above (hybrid), utilizing the advantages of each approach.

The hybrid approach can be generalized as depicted in figure 3.

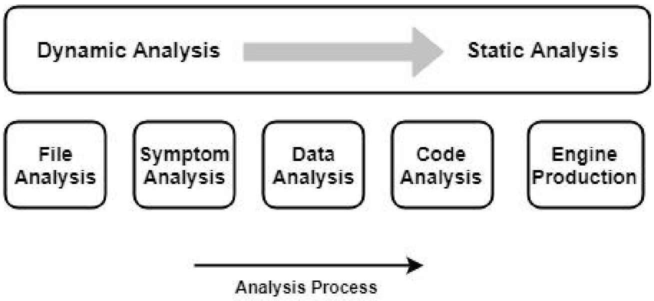

Figure 3: Malware analysis process

Since malware analysis is primarily data analysis, the same approach can be employed to the huge historical data that is already available. Most of the available big data is highly unordered, unlabeled, imbalanced, and has no uniform structure.

Because of this, data mining approaches can be applied to the security related data to provide proper structure, handle their associations and adjust the missing values accordingly through calculated forecasting [4], [5], [6]. Researches using this approach have proven to be efficient compared to traditional knowledge oriented approaches. These approaches also reduce human intervention by automating the analysis process, considering all possibilities, thus reducing human induced errors.

B. Importance of Data in Malware Analysis

Mining the data in order to obtain signatures of existing or potential malware can be done in three phases:

- **Phase 1:** This takes the advantage of available labeled data by learning the associations to form rules (signatures), and then compare those signatures to new malware signatures encountered in real-time. This method is also called misuse detection. It produces lower false positives, but it is not effective for malwares with signatures that are not yet detected (Zero day attacks).

- **Phase 2:** In case of unlabeled and unstructured data, the constituents and their associations need to be understood in order to devise efficient signatures. This phase is very important, since it performs calculated prediction of associations of attributes of given data. This phase takes into account the variable dependencies at different hierarchy to produce association rules [5]. This can be done by monitoring the system behavior for a specific time period, understanding the variables at play along with their associations and drafting a normal profile of the system. Any signatures detected after the analysis, are compared against the normal profile; resulting deviations are considered as potential intrusions. This approach is also called as Anomaly detection and has the advantage of detecting new attacks. As new signatures can also be normal operations, it also results in higher false positives.

- **Phase 3:** Some of the attacks found in the literature show temporal characteristics. It has been proved that a given signature or a set of signatures occurring at certain time intervals lead to different kinds of intrusions, when captured over certain period of time. Such attacks are handled in this phase. This process is called time-series analysis.

C. Machine Learning and Deep Learning Techniques in Malware Analysis

The learning and prediction processes discussed previously, are usually performed using machine learning or deep learning algorithms. These learning models help automate the decision making processes (in case of rule building) and signature or missing values prediction (upon system model analysis) efficiently. They also adapt well with the dynamics of both, system topology and data distribution, improvising the rules accordingly.

Machine and Deep learning techniques are also used during post-processing. In other words, after performing malware analysis on the available labeled data (training), the analyzer (intrusion detection system) can be tested upon new data in real-time for performance assessment.

With the increasing demand for time and mission critical applications, the security provided, should be top notch and hence there is no compromise with respect to system performance. Even though machine learning techniques handle the objective very well, deep learning came into popularity through Neural Networks (NN), with its _ne grained performance, compared to regular machine learning techniques. Further advancements in neural networks (like convolution NN, recurrent NN) made them useful for a variety of applications, including Cybersecurity [7].

The advantage of neural networks over other learning techniques is that, it builds models, considering constituent atomic entities over higher abstractions, which results in missing out some details. But the _ne grained performance comes with a performance cost. Deep learning techniques are highly resource and data intensive. They demand sophisticated hardware resources (like graphical processing unit and high speed high capacity primary memory). A typical deep learning neural network is as shown in the figure below.

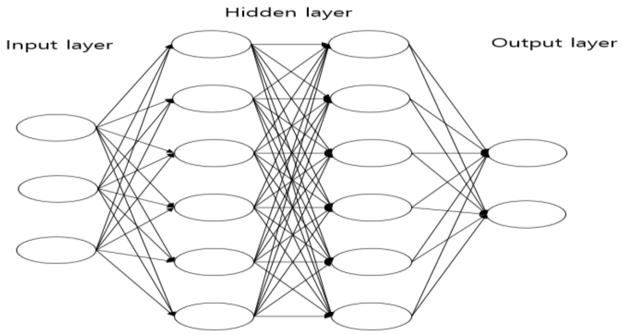

Figure 4: Representation of a Neural Network

III. REALTED WORK

Saif et. al, [8] have proposed an efficient computational framework for malware detection based on Deep Belief Networks has been developed for malware detection in android based systems, performing static malware analysis. The proposed framework achieves an efficiency of 99.1%. It also performs similarity checking between same applications downloaded from different sources.

Sang Ni et. al, [9] proposed a malware classification algorithm (MCSC), extracting the opcode sequences from malware and encoding them to equal lengths to convert to gray images and classify using CNN with an average accuracy of 98.862%.

Kim et. al, [10] proposed a system with a capability of detecting a Zero day malware using transferred generative adversarial networks (GANs) by generating virtual data and train the detector, making it robust to data deformation. The achieved average classification accuracy is around 95.74%.

Le et. al, [11] proposed a malware classification using deep learning data-driven approach using CNN and LSTM. It has a working accuracy of 98.8% with a better time efficiency.

Karbab et. al, [12] proposed a malware detection framework for android environment for sequence classification using neural networks.

Vinaykumar and Soman [13], proposed a static PE malware detection approach, using EMBER malware benchmark data with classification efficiency of 98.9%.

IV. CONCLUSION

This paper provides a brief overview of deep learning and its scope in cybersecurity. It also signifies the role of audit data in malware analysis, generation and validation of new data, and introduces the methods to handle unordered and unlabeled data. Finally, this paper discusses the recent research trends in malware analysis using deep learning.

REFERENCES

[1]. I. Goodfellow et. al, Deep Learning, MIT Press, 2016.

[2]. Y. S. Lee et. al, Trend of Malware Detection Using Deep Learning, ICMET, ACM, 2018.

[3]. M. Sikorski and A. Honig, Pratical Malware Analysis, ISBN: 978-1-59327-290-6, 2012.

[4]. Niranjan A et. al, Security in Data Mining – A Comprehensive Survey, GJCST, vol. 16, no. 5, pp. 51-72, 2016.

[5]. W. Lee and S. J. Stolfo, A framework for Constructing Patterns and Models for Intrusion Detection Systems, ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 227-261, 2000.

[6]. Z. Yan et. al, A Survey on Security Related Data Collection Technologies, IEEE Access, vol. 6, pp. 18345-18365, 2018.

[7]. Z. Liu et. al, Machine Learning and Deep Learning Methods for Cybersecurity, IEEE Access, vol. 6, pp. 35365-35381, 2018.

[8]. D. Saif et. al, Deep Belief Networks-based framework for malware detection in Android systems, Alexandria university Journal, vol. 57, pp. 4049-4057, 2018.

[9]. S. Ni et al, Malware identification using visualization images and deep learning, Computers & Security, Elsevier, pp. 1-15, 2018.

[10]. J.Y. Kim et. al, Zero-day malware detection using transferred generative adversarial networks based on deep auto encoders, Information Sciences 460-461, pp. 83-102, Elsevier, 2018.

[11]. Q. Le et. al, Deep Learning at the shallow end: Malware classification for non-domain experts, in proc. DFRWS, USA, pp. 118-126, Elsevier, 2018.

[12]. E.B. Karbab et. al, MalDozer: Automatic framework for android malware detection using deep learning, in proc. DFRWS, Europe, pp. 48-59, Elsevier, 2018.

[13]. Vinaykumar R and Soman K.P, DeepMalNet: Evaluating shallow and deep networks for static PE malware detection, ICT Express, KICS, vol. 4, pp. 255-258, 2018.

Nitish A is currently pursuing his PhD under the guidance of Dr. Hanumanthappa J, in the Dept. of Studies in Computer Science, University of Mysore.

His research interests include Cybersecurity, Malware detection, Machine learning and Wireless Sensor Networks.

Hanumanthappa J is currently working as an Associate Professor in the Dept. of Studies in Computer Science, University of Mysore.

He has over ten years of teaching and research experience and has published over 50 papers in various national and international journals and conferences.

His areas of research include Wireless Routing, Wireless Sensor Networks, Intrusion Detection and Cryptography.