Using Artificial Intelligence to Magnify Malware Detection Techniques

Paul Spencer

Brigham Young University

WRTG 316: Technical Communication

Professor Lisa Johnson

December 9, 2020

Abstract

The increase in the complexity of computer systems throughout the world has sparked a growth in malware that can not be handled by overworked cybersecurity professionals. However, the expansion in computing power has also come with increased artificial intelligence capabilities. Because the patterns followed by artificial intelligence algorithms are so similar to patterns used by actual professionals, these algorithms have proven to be highly capable of detecting and tracking down malware. The goal of finding malware as early as possible is achieved through active detection involving scans of computer systems as well as passive detection including programs that listen for signs of malware in the background. Active and passive methods of detecting malware will prove to be an important part of our daily lives.

Using Artificial Intelligence to Magnify Malware Detection Techniques

Malicious software, also known as malware, is any type of program run on a computer without the consent of the owner or administrator of the machine. Commonly, software like this is installed and abused by hackers and criminals to extort victims into relinquishing their money. Because of the momentum carried by the computing industry and the parallel growth in the malware industry, malware has become a harsh reality that affects the daily lives of countless people. As cybersecurity experts try to catch bugs and patch important software on the fly, it has become an almost impossible task to keep up with the expansion of malware. However, enabled by modern improvements in computing, artificial intelligence (AI) has become an extremely successful tool in detecting malware more efficiently than ever.

There is no viable way to detect the wide variety of malware found in the world today without the help of artificial intelligence. Old methods of finding and dealing with malicious software are simply too slow and inefficient for today's standards. With the increasing number of ways we use computers every day, it is essential to neutralize threats as soon as they come up. Identifying and alerting experts of threats has become the role of advanced AI malware detection solutions. AI algorithms recognize patterns in data to predict outcomes. In the case of malware detection, AI algorithms can look through common signs of malware and decide if there is a problem, and then intelligently decide where to look for it. Cybersecurity experts can then assess the predictions and take care of the problems.

In this review, I will analyze the research that has been done concerning the use of artificial intelligence in improving the process of malware detection for cyber experts. This includes recent developments in both passive and active malware detection techniques. I will

also report on what lies in the future of this field and what research should be done further concerning the topic, especially including some specific fields that could benefit from AI malware detection.

Impact of Artificial Intelligence

To fully grasp the incredible impact of artificial intelligence on our ability to detect malware, it is helpful to understand the history of malware as well as old methods of detecting it. An understanding of this timeline makes it clear that efficient and reliable AI malware detection is an enormous milestone. Examining the methods that AI algorithms use also demonstrates how closely modern malware detection algorithms follow old methods under the hood.

Beginnings of the malware problem

According to Karim, Azam, Shanmugam, et. al. (2019), the first malware began circulating in the form of worms. These were fairly harmless and originated as an experiment to see how programs could move between computers. Later on, the same concept was used to create self-replicating ransomware, which was very dangerous and malicious. After the first worms and self-replicating viruses, denial of service viruses like “Wabbit” were used to bog down a computer until it crashed. This was a whole new class of viruses and was very effective because it could do actual damage to a computer. Phishing attempts (including malicious emails) have since surfaced and are perhaps the most dangerous type of malware because they rely on the incompetence of users.

Malware outgrows its cage

This array of attacks was at first manageable for security professionals. According to Kerravala (2020) as well as Karim, Azam, Shanmugam, et. al. (2019), these security professionals were able to stay on top of the threats by putting in long hours. They would manually hunt down copies of self-replicating code, intercept and analyze at least a sample of emails, and otherwise supervise all operations on the computers under their domain. However, Gilbert, Mateu, and Planes (2020) explain that the quantity of malware grew much faster than the capabilities of cybersecurity professionals. This was due to the overall exponential growth in computing. The reason security professionals could no longer keep up with malware was not that the malware changed, but rather that its quantity exploded into exponential growth. This expansion of the malware industry quickly soared past the malware detection capabilities of that time. The old method of detecting malware involved a cybersecurity professional looking for signs of malware, such as heightened resource usage for seemingly no reason, inconsistencies in internet communication, and suspicious emails and communications. Artificial intelligence presents a similar solution to the problem by imitating the abilities of professionals to analyze data and recognize patterns.

Advantages of AI methods

In their evaluation of specific AI algorithms applied to cybersecurity, Shaukat, Luo, Varadharajan, et. al. (2020) recognize neural networks and decision trees as some of the most efficient and cost-effective methods, especially when applied well. Because neural networks and decision trees are the most common techniques used in AI algorithms for detecting malware, it is apparent that these methods fall under that category. Also, as discussed by Gilbert, Mateu & Planes (2020), the growth of malware corresponds with the growth of computing, so manually searching for malware quickly became ineffective. However, because the efficiency of AI

algorithms also grows with the field of computing, it will remain a viable solution for quickly finding malware in complex systems for a long time, if not indefinitely. Shaukat, Lue, Varadharajan, et. al. (2020), and Gilbert, Mateu & Planes (2020) agree that using AI to detect malware is the best solution to the problem of lacking cyber protection in an increasingly dangerous cyber world.

Specifics on neural networks and decision trees

Some of the most common and effective tools used in intelligent malware detection are neural networks and decision trees. These are fairly advanced computer science topics, which may be unknown to those working in the cybersecurity field, but I attempt to explain them here. In mathematics, functions are used to convert an input to an output using a defined set of instructions. In a neural network, a large group of inputs is fed into the algorithm along with their known corresponding outputs. The algorithm then tweaks itself little by little with each input until it can predict an output that is usually correct for a given input. Therefore, the algorithm teaches itself the set of instructions that convert inputs to outputs, much like a human brain. In the case of malware detection, different information such as computer metrics, email contents, and network traffic are provided as inputs to a custom neural network and the algorithm provides some sort of output, such as the probability of the presence of malware. Along with neural networks, decision trees can be used to further narrow down the actual location of malware when its presence is detected. A branch of a decision tree consists of a sequence of information gathering, analysis, and progression to another level in the tree. The branches of the tree are then followed recursively to a conclusion. For example, a computer could be told to look at the sender of an email, then check if it is on a blacklist, then decide to proceed with analyzing it for malware if it is, or if not, decide the email is clean. This is a simple decision tree. However,

much larger decision trees consist of many layers. Neural networks and decision trees closely mimic the way the human brain functions, which is largely the reason they are capable of working together to accomplish the task of detecting malware in ways that mimic the methods of cybersecurity professionals.

Passive Malware Detection

Kerravala (2020) explains that the need for a huge range of malware detection methods justifies grouping the methods into two main categories: passive and active malware detection. Inside each of the two groups, the techniques share common properties. For example, passive malware detection algorithms require a constant inflow of data. The algorithm then filters through the data as it arrives to listen for patterns that could signify the presence of malware. The data provided as input constitutes the difference between the algorithms of this type. Data center metrics, binary analytics, and autonomous vehicle metrics are specific examples of types of inputs to intelligent passive malware detection algorithms, each with different strengths and weaknesses.

Data Center Monitoring

In data centers, there is an enormous amount of valuable data available for malware detection. The trick is deciding which pieces of data are the most useful and efficient enough for low-cost passive detection algorithms. Bartolini, et. al. (2020) introduces one cutting-edge tool for this exact purpose called pAElla. This AI-powered tool for passively finding breaches in a datacenter uses a handful of the available inputs. It reads in network traffic which includes quantities, directions, throughputs, hotspots (i.e. which machines are using the most network bandwidth), and frequent external sources and destinations. The pAElla tool also uses the

information provided by the physical metrics of the machines. It takes in the amount of power being consumed by each machine and the temperatures read by various sensors throughout the data center. Additionally, the actual workloads of different machines including CPU, GPU, and memory utilization levels, along with task quantities and other simple metrics are monitored. It then feeds these inputs through a trained neural network to recognize any abnormalities that commonly signify the presence of malware. Next, if there are any resulting signs of malware, the intelligent detection algorithm will report on potential sources of the problem.

Because of its primary dependence on information about the physical state of the datacenter as a whole, the approach discussed by Bartolini et. al. (2020) is not as effective at finding specific details about the malware it finds. Methods like the one discussed by Alazab, Akram, et. al. (2020) that use more internal information about unique devices in a system are more capable of narrowing down problems but would be more costly because of their need to accommodate such a variety of devices. Additionally, Bartolini et. al. (2020)'s data center management technique is less customizable because it only accommodates large numbers of one type of technology (i.e. large numbers of similar servers instead of unique IoT devices).

Passive Binary Analysis

In addition to data centers full of servers, regular home computers also require some level of passive malware detection. Personal computers and laptops do not have nearly the monitoring capabilities of a data center, but they do lack the element of distribution that makes detection more difficult. One common method of listening for trouble inside of a computer, as introduced by Gilbert, Mateu & Planes (2020), is through binary analysis. As a program is loaded into memory, the operating system kernel (a program that provides a low-level interface to computer

hardware for the programs of the computer) can read off the opcodes, or binary instructions, into an artificial neural network. This network is trained to recognize common patterns of malware, much like the neural network portion of the pAElla algorithm. Using the output of the neural network, the operating system reports to the user that a program may be dangerous, or else continues to run the program. Alnumay, et. al. (2021) explains that because this process happens each time a program is loaded into memory for execution, the technique is classified as passive, unlike the active methods of dynamic binary analysis which I will discuss later in this review.

Autonomous Vehicles

There are countless examples of computers beyond the obvious servers, PCs, laptops, and phones. For instance, self-driving cars consist of far more sensors and processors than one might think. These systems are all independently capable of being infected with malware and causing drastic damage to the vehicle or its passengers. However, artificial intelligence is highly applicable due to the high number of sensors and different pieces of data that are collected. Most self-driving cars come equipped with chips enhanced with machine learning acceleration. Choi & Park (2020) explains how AI is commonly used in detecting malware in self-driving cars. The car can read a wide range of sensors (tire pressure, gas temperature, engine noise levels, etc.) and run them through a neural network to detect malware while strengthening the algorithm itself. Because of the extreme importance of the safety of self-driving vehicles, malware must be quickly discovered and removed. This way, hackers have little to no time to sabotage the vehicle.

In the studies done by Choi & Park (2020), Alnumay, et. al. (2021), Gilbert, Mateu & Planes (2020), Bartolini et. al. (2020), and Alazab, Akram, et. al. (2020), AI is used to passively detect malware of some type. In each malware detection system, neural networks are used to

recognize patterns in critical data that could signal malware, and decision trees are used to further hunt down specifics about the detected malware. However, each application depends on different inputs that cater to the specific dangers of each system. For example, Choi & Park (2020)'s system for detecting malware on the fly in self-driving cars depends on reading a variety of sensors along with information related to the internal computer systems of the car while Bartolini et. al. (2020)'s pAElla system for recognizing malware in datacenters relies more on physical metrics. This and other differences demonstrate the importance of continuing to research the best ways to detect malware in unique environments.

Clearly, artificial intelligence plays a huge role in modern passive malware detection methods. Using pre-trained neural networks for these applications also means that the techniques are efficient. In-depth programmatic approaches to discovering potential malware require heavy computational costs, while simply running some inputs through a trained neural network is much quicker and can be just as accurate.

Active Malware Detection

In addition to constantly checking for breaches through passive malware detection, cybersecurity experts frequently need to perform one-time scans of a target to determine if malware is present. After this analysis, they decide how to fix any detected problems. Scans like these are useful for checking devices known to possibly be infected with malware. For example, Keller (2018) explains that the United States Air Force frequently needs capabilities like this to clean and check devices that were used in dangerous areas in the world as well as devices that are about to be deployed on critical missions. AI has recently made an enormous impact in active device scans such as those used by the Air Force, which enables cybersecurity experts to analyze

targets faster than ever. Some of the most interesting examples of AI in active malware detection are in locating malicious cryptocurrency mining software, performing dynamic binary analysis of programs, and scanning files in a device as a check for malware.

Malicious cryptocurrency mining detection

In the field of cryptocurrency, it is possible to generate Bitcoin or another unit of cryptocurrency via mining. Cryptocurrency mining requires an enormous amount of computational power (either on a single machine or distributed across many) to make any profit but can be highly lucrative. In a study on cryptocurrency mining malware, Canavese, Lopez, et. al. (2020) investigates hackers who have begun to include cryptocurrency mining software in malware packages. These hackers are then able to remotely harness the power of a victim's computer to mine cryptocurrency for themselves and therefore generate a profit without having to pay for the means (the computer hardware and the power bill). New advancements in using artificial intelligence for detecting malware make it possible to quickly scan a device and determine if malware of this class is present. Checking hardware metrics, load balancing, and network traffic analysis is essential to this malicious cryptocurrency mining software detection method. Canavese, Lopez, et. al.'s (2020) tool takes these as inputs, feeds them to a neural network, and is then able to efficiently get an accurate answer as to whether malware is present or not. The neural network is trained before distributing the tool so that the network is ready to use when a user decides to scan their device. This tool is highly effective and relies almost entirely on recent advancements in AI.

The fact that it is used as a scanner when there is already suspicion of cryptocurrency mining software is what qualifies this technique active and not passive. Canavese, Lopez et. al.

(2020)'s tool specifically hunts down cryptocurrency traffic, then performs an analysis on it. Professionals use it in their toolkit for identifying malware. This is unlike cases of listening detection mechanisms like Choi & Park's (2020) self-driving car malware detection system, because it is more computationally intensive in an instant, and because it runs once and returns a result rather than sporadically notifying the presence of malware. The self-driving car technique constantly monitors and alerts of the presence of malware.

Dynamic Binary Analysis

Dynamic binary analysis tools also use artificial intelligence to increase efficiency while still reporting accurate results. Many programs stored on the hard drives of computers rely on other programs. This is the concept of dependency. Program dependencies can be swapped out and modified, which hackers sometimes exploit to insert malicious pieces of code. The process of checking all these dependencies of a program for potential malware becomes a very large problem, which can make it difficult to find accurate results. However, as shown in research done by Alnumay, Imtiaz, Jalil, et. al. (2021), AI has recently made it much more feasible to obtain quick and accurate results. They present a method for performing dynamic binary analysis with AI. A program is read and checked for dependencies. Then, the dependencies all go through the same process recursively until all the code that may be used by a program has been read. Next, a pre-trained neural network quickly analyzes the memory addresses, opcodes, strings, and other pieces of data contained in the program to determine if anything has been tampered with. The network is trained to recognize common dependencies and understand their inner workings, so it is very capable of recognizing fraudulent code. Performing these steps without the neural network could take enormous amounts of time, so AI has made a large impact on the industry. This method can be used across a very wide variety of devices because it doesn't require heavy

computation. For example, it is very common in Android devices, which are usually fairly low powered.

Filesystem scanning

Lastly, performing a simple scan of files on a device and then heavily using AI to decide if they are safe or not has become a common technique with successful results. One powerful attribute of artificial intelligence algorithms is their capability of being applied to a variety of levels of a problem. Kerravala (2020) points out that you can either train a neural network to perform one specific step of a process, or you can use it as a "black box" and allow it to perform multiple, or even all, of the steps of a process. For example, as explained by Alnumay, Imtiaz, Jalil, et. al. (2021), in a dynamic binary analysis, dependencies are programmatically tracked down for analysis, and then the opcodes are fed into a neural network to detect specific malicious patterns. After that, other tools including decision trees are used to further investigate potential malware. But, on the "black box" end of the spectrum, we can simply read the files on a hard drive into a neural network to determine if any are dangerous. The algorithm is trained by using a variety of files, some of which are safe and some of which are dangerous. Based on the algorithm's "experience" with the files it was trained on, it then is capable of deciding fairly accurately if other files are safe or dangerous.

The tradeoff between this "black box" technique and other techniques is the difficulty of training neural networks to the point where they give accurate results. As explained by Shaukat, Luo, Varadharajan, et. al. (2020), the smaller the step performed by a neural network, the easier it is for distributors of the tools to pre-train it. On the easy end of the spectrum lies Choi & Park's (2020) study of self-driving cars. Specific neural networks are assigned to specific sensors

to determine if the part controlled sensor is worth investigating. This is a small task for a neural network, which can therefore be trained at little cost. On the other end of the spectrum is Kerravala's (2020) example of scanning filesystems. Jumping directly from the contents of a file to a prediction of its safety is extremely difficult and requires a very well trained neural network.

In all, artificial intelligence has made the process of performing active malware scans far more efficient and accurate. It cuts down on complex analysis times that were formerly required for any success in the area by using neural networks and other intelligent methods. Now, thanks to these powerful tools, we can very accurately decide if a target is infected with malicious software of many types, both based on information about the hardware and operating system, as well as using information gathered directly from the programs and files stored on the device.

Conclusion and Further Research

Malware detection is critical to the future of the world. The rise in technology usage has increased so much in the last 20 years that computer systems have become too complex for cybersecurity professionals to manually search for malware and handle all vulnerabilities. At the same time that computer systems and malware have become more common, the capabilities of artificial intelligence have skyrocketed.

Because the patterns that AI algorithms use (i.e. neural networks and decision trees) are so similar to the patterns that professionals follow in detecting malware, it has become clear that AI is a viable solution to the overall inability of security professionals to keep up with malware. Passive malware detection through AI involves constantly running a low-intensity program that listens to processes and I/O on a computer or system of computers. This information is then run through a neural network and decision tree to detect malware and track down its source to

ultimately reveal it to security professionals. Similarly, active detection involves running inputs through a neural network and decision tree, but the inputs are instead collected by a one-time intensive program that scans a device or system of devices. As discussed previously, AI has improved the effectiveness, cost-efficiency, and maintainability of malware detection. Whether it is through using a well-trained black box neural network for analyzing big picture data, or a sequence of networks and decision trees to perform small steps towards a solution, AI has enormous capabilities in this application. AI greatly increases the amount of malware that can be detected and tracked down without weighing down computers with computationally intensive programs that are not resistant to changes in the malware industry.

Future Research

It is clear that we will soon begin to see far more AI malware detection tools and software used daily. The applications of AI in this field are extensive and still growing. Some areas that should soon be researched in this field include overall internet safety and integrity as well as air and space defense. Because the internet consists of countless sites that may or may not be legitimate, and the fact that large companies commonly use internet crawlers and scrapers to gather information about the internet, it would make sense for them to use AI to detect the legitimacy of websites. Then, they could report and potentially shut down malicious websites and scams. Along with internet safety, future research should include cyber defense in air and space travel. The quantity of computer systems involved in a safe flight is incredibly large. The security of these systems is critical to the safety of the passengers, so detecting malware before any damage is done is a necessity. AI could revolutionize this field and greatly increase the safety of passengers on flights. These two topics are logical and important directions for research to continue in the world of AI malware detection.

References

Alazab, M., Akram, J., Vasan D., Venkatraman, S., et. al. (2020). *MTHAEL: Cross-architecture IoT malware detection based on neural network advanced ensemble learning* doi:10.1109/TC.2020.3015584

Alnumay, W. S., Imtiaz, S. I., Jalil, Z., Javed, A. R., Liu, X., Rehman, S. U. (2021). *DeepAMD: Detection and identification of Android malware using high-efficient Deep Artificial Neural Network. Future Generation Computer Systems*, 115, 844-856. https://search-lib-byuedu.erl.lib.byu.edu/byu/record/edsbyu.edselp.S0167739X2032985X

Bartolini A., & Benini, L., & Libri, A. (2020). *pAElla: Edge AI-Based Real-Time Malware Detection in Data Centers* https://10.1109/JIOT.2020.2986702

Canavese, D., López, D. R., Mozo. A., Pastor, A., Regano, L., Vakaruk, S., et al. (2020). *Detection of encrypted cryptomining malware connections with machine and deep learning* doi:10.1109/ACCESS.2020.3019658

Choi, J. & Park, S. (2020). *Malware detection in self-driving vehicles using machine learning algorithms. Journal of Advanced Transportation, 1-10*. https://search-lib-byuedu.erl.lib.byu.edu/byu/record/edsbyu.asn.141398720

Gibert, D., Mateu, C., & Planes, J. (2020). *The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. Journal of Network and Computer Applications*, 153, 102526. doi:https://doi-org.erl.lib.byu.edu/10.1016/j.jnca.2019.102526

Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). *A Comprehensive Survey for Intelligent Spam Email Detection*. IEEE Access, Access, IEEE, 7, 168261-168295. https://search-lib-byu-edu.erl.lib.byu.edu/byu/record/edsbyu.edseee.edseee.8907831

Keller, J. (2018). *DISA looks for ways of using artificial intelligence (AI) to detect malware.* Military and Aerospace Electronics https://www.militaryaerospace.com/trusted-computing/article/16707238/disa-looks-for-ways-of-using-artificial-intelligence-ai-to-detect-malware

Kerravala, Z. (2020). *How Using AI Vastly Improves Threat Detection.* eWeek https://www.eweek.com/security/how-using-ai-vastly-improves-threat-detection

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., Chen, S., Liu, D., et al. (2020). *Performance comparison and current challenges of using machine learning techniques in cybersecurity.* Energies, 13(10), 2509. doi:http://dx.doi.org.erl.lib.byu.edu/10.3390/en13102509