Artificial Intelligence Assisted Malware Analysis

Mahmoud Abdelsalam
Manhattan College
Riverdale, NY, USA
mabdelsalam01@manhattan.edu

Maanak Gupta
Tennessee Technological University
Cookeville, TN, USA
mgupta@tntech.edu

Sudip Mittal
University of North Carolina
Wilmington
Wilmington, NC, USA
mittals@uncw.edu

ABSTRACT

This tutorial provides a review of the state-of-the-art research and the applications of Artificial Intelligence and Machine Learning for malware analysis. We will provide an overview, background and results with respect to the three main malware analysis approaches: static malware analysis, dynamic malware analysis and online malware analysis. Further, we will provide a simplified hands-on tutorial of applying ML algorithm for dynamic malware analysis in cloud IaaS.

CCS CONCEPTS

• **Security and privacy** → **Malware and its mitigation**; **Intrusion detection systems**; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

Security, Online Malware Detection, Machine learning, Artificial Intelligence

ACM Reference Format:
Mahmoud Abdelsalam, Maanak Gupta, and Sudip Mittal. 2021. Artificial Intelligence Assisted Malware Analysis. In *Proceedings of the 2021 ACM Workshop on Secure and Trustworthy Cyber-physical Systems (SAT-CPS'21), April 28, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3445969.3450433

1 INTRODUCTION AND MOTIVATION

The war between malware analysts and malware writers is an everlasting fight considering the growing complexity and innovative techniques of evolving malware. Security analysts has been struggling with the amount of malware introduced everyday. In the year 2019, around 948 government agencies, educational establishments and health-care providers got hit with a barrage of ransomware attacks at a potential cost of $7.5 billion [1]. We can anticipate that such attacks on mission critical infrastructure will continue to grow in coming years. This is largely due to the techniques like polymorphic malware which is able to change and evolve while preserving code semantics. In addition, malware writers try to complicate the task of security analysts by using techniques such as obfuscation, where binary and textual data is unreadable or hard to understand and packing, where a malware is modified using a run-time compression (or encryption) program.

As such, the need for automated ways to counter such a vast amount of newly developed malware has become necessary. In particular, current research has dominantly focused on the application of Artificial Intelligence (AI) and Machine Learning(ML) techniques for malware detection largely because of their ability to keep pace with malware evolution. This tutorial provides a description of the state-of-the-art approaches in a traditional AI/ML assisted malware detection workflow in cloud IaaS, including malware samples gathering from the cloud testbed, feature identification and collection, and AI/ML models training.

2 TUTORIAL DESCRIPTION

2.1 Outline

Overview and Categorization of AI/ML based Malware Detection: We will begin our tutorial with an overview of the application of AI/ML for malware detection in cloud IaaS, including it's benefits and motivation. We then provide an overview of the broad categories of AI/ML assisted malware detection approaches, mainly *static*, *dynamic* and *online* with respect to their usage, aims, advantages and disadvantages.

Malware Sample Gathering and Feature Identification: In order to combat malware-based attacks, security researchers need to have a databank of executable and workable malware samples to conduct the experimentation work necessary. As such, we will discuss how security researchers, to prevent and detect these malware-based attacks, retrieve malware samples from the "wild." In samples gathering, we will discuss ways of acquiring malware executable including honeypots (active and passive) and malware public databases such as VirusTotal[1] and VirusShare[2]. In system features identification, we will discuss commonly used static features like binary n-grams, Control Flow Graphs (CFGs) and static API calls, along with behavioral features like performance metrics, memory information, and system calls. For the data collection, we will discuss the usage of isolated environments such as sandboxes (e.g., Cuckoo Sandbox) and online virtual machines (VMs) in cloud. We will also discuss the limitations of using isolated environments (referred to as dynamic) and other alternatives including the use of a live testbed for real-world use cases simulation. Further, we will discuss host-based and network-based collecting agents as well as virtual machine introspection.

State-of-the-Art AI/ML assisted Malware Detection Techniques: We will start with file classification techniques including static and dynamic analysis. In static analysis, we will discuss three

[1] https://developers.virustotal.com/reference#public-vs-private-api
[2] https://virusshare.com/

SAT-CPS'21, April 28, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8319-6/21/04.
https://doi.org/10.1145/3445969.3450433

major classes of features including: Binary N-grams [2–5], Control Flow Graphs (CFGs) [6–8] and Static features/Disassembling [9, 10]. Although static analysis techniques are efficient, most recent malware are sophisticated and has polymorphic nature, which hinder the effectiveness of static analysis. To overcome this, we will also discuss AI/ML based dynamic analysis techniques, which focus on behavioural aspects of malware. To that end, we will discuss various tools needed to monitor system processes, filesystem and registry changes and network activity. We will provide a use case that focuses on running executables in a controlled environment and observing their behavior, where system/API calls [11–15] are mainly used. In addition, we will discuss online detection approaches, which will help understand the need and ability to continuously monitor the entire cloud IaaS system for detecting the presence of malicious activities. This includes approaches that rely on different features which are more dynamic and time dependent such as performance metrics [16–21], memory features [22, 23] or run-time system/API calls [24–26].

2.2 Live Demo

We will do a live demo that will be part of the tutorial and include training a machine learning model for malware detection. This particular model will focus on dynamic analysis by extracting behavioral features from malware analysis reports generated using Cuckoo sandbox. The demo will go through the steps of gathering malware samples, using Cuckoo sandbox to generate the reports, parsing the reports to acquire the data needed, and finally, preprocessing and training the ML model.

2.3 Target Audience

This tutorial aims to target and spark the interest of computer science audience at the introductory and intermediate levels. This includes students, faculty, industry representatives and researchers who are interested in the intersection of malware analysis and AI/ML. In addition, this will encourage cybersecurity professionals who are interested in expanding their skill set by including the application of AI/ML, and help towards producing next generation of cyber warriors.

2.4 Learning Outcomes

After attending this tutorial session, we expect the attendees will be able to:

- Explain the importance and need for AI/ML skill set for malware analysis.
- Understand the broad spectrum of AI/ML based malware detection approaches and their categorization.
- Describe the overall common steps required for researchers and professionals to develop AI/ML assisted malware detection techniques.
- Understand and deploy various data collection and feature identification techniques for malware analysis.
- Use Cuckoo sandbox to generate dynamic analysis reports.
- Train simple ML model for malware detection using behavioral malware data.

3 PRESENTERS BIOGRAPHY

Mahmoud Abdelsalam received his B.Sc degree from the Arab Academy for Science and Technology and Maritime Transportation (AASTMT) in 2013, his M.Sc from the University of Texas at San Antonio (UTSA) in 2017 and his Ph.D. from UTSA in 2018. He is currently an assistant professor with the Department of Computer Science, Manhattan College. Prior to joining Manhattan College, he was working as a Post doctoral research fellow in the Institute for Cyber Security (ICS) at UTSA. His research interests include computer systems security, anomaly and malware detection, cloud computing security and monitoring, cyber physical systems security and applied machine learning.

Maanak Gupta is currently an assistant professor in computer science at Tennessee Technological University, Cookeville, USA. He received M.S. and PhD degree in computer science from the University of Texas at San Antonio (UTSA). He has also worked as post doctoral fellow at the Institute for Cyber Security (ICS) at UTSA. His primary area of research includes security and privacy in cyber space focused in studying foundational aspects of access control and there application in technologies including cyber physical systems and cloud computing. He is also interested in malware analysis and AI assisted cyber security solutions. He holds a B.Tech degree in computer science and engineering from Kuruskhetra University, India, and M.S. degree in information systems from Northeastern University, Boston. He is a reviewer and a technical committee member for journals and conferences.

Sudip Mittal is an assistant professor of computer science at University of North Carolina Wilmington (UNCW). He received a Ph.D. in computer science form University of Maryland Baltimore County. His primary research interests are cybersecurity and artificial intelligence. His goal is to develop the next generation of cyber defense systems that help protect various organizations and people. He holds a M.Tech and a B.Tech degree in computer science from IIIT Delhi. He has also previously worked with Accelerating Cognitive Cyber Security Research Lab (ACCL), Ebiquity Research Lab, Center for Hybrid Multicore Productivity Research (CHMPR) and Cybersecurity Education and Research Centre (CERC@IIITD).

ACKNOWLEDGEMENT

This work and tutorial is partially supported by National Science Foundation awards 2025685, 2025682, and 2025686.

REFERENCES

[1] CRN. The 10 Biggest Ransomware Attacks of 2019 . https://www.crn.com/slide-shows/security/the-10-biggest-ransomware-attacks-of-2019, 2019.

[2] Gil Tahan, Lior Rokach, and Yuval Shahar. Mal-ID: Automatic malware detection using common segment analysis and meta-features. *Journal of Machine Learning Research*, 13, 2012.

[3] Tony Abou-Assaleh and et al. N-gram-based detection of new malicious code. In *COMPSAC*, volume 2. IEEE, 2004.

[4] J Zico Kolter and Marcus A Maloof. Learning to detect and classify malicious executables in the wild. *Journal of Machine Learning Research*, 7, 2006.

[5] Asaf Shabtai and et al. Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey. *information security technical report*, 14, 2009.

[6] Shun Tobiyama, Yukiko Yamaguchi, Hajime Shimada, Tomonori Ikuse, and Takeshi Yagi. Malware detection with deep neural network using process behavior. In *COMPSAC*, volume 2. IEEE, 2016.

[7] Zeliang Kan, Haoyu Wang, Guoai Xu, Yao Guo, and Xiangqun Chen. Towards light-weight deep learning based malware detection. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, pages 600–609. IEEE, 2018.

[8] Mojtaba Eskandari and Sattar Hashemi. Ecfgm: enriched control flow graph miner for unknown vicious infected code detection. *Journal in Computer Virology*, 8:99–108, 2012.

[9] Joshua Saxe and Konstantin Berlin. Deep neural network based malware detection using two dimensional binary program features. In *10th MALWARE*. IEEE, 2015.

[10] Seonhee Seok and Howon Kim. Visualized malware classification based-on convolutional neural network. *Journal of the Korea Institute of Information Security and Cryptology*, 26, 2016.

[11] Rakshit Agrawal, Jack W Stokes, Mady Marinescu, and Karthik Selvaraj. Robust neural malware detection models for emulation sequence learning. *arXiv preprint arXiv:1806.10741*, 2018.

[12] Ben Athiwaratkun and Jack W Stokes. Malware classification with LSTM and GRU language models and a character-level cnn. In *ICASSP*. IEEE, 2017.

[13] George E Dahl, Jack W Stokes, Li Deng, and Dong Yu. Large-scale malware classification using random projections and neural networks. In *ICASSP*. IEEE, 2013.

[14] Wenyi Huang and Jack W Stokes. MtNet: a multi-task neural network for dynamic malware classification. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2016.

[15] Dhilung Kirat and Giovanni Vigna. Malgene: Automatic extraction of malware analysis evasion signature. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 769–780. ACM, 2015.

[16] John Demme and et al. On the feasibility of online malware detection with performance counters. In *ACM SIGARCH Computer Architecture News*, volume 41. ACM, 2013.

[17] Mahmoud Abdelsalam, Ram Krishnan, and Ravi Sandhu. Clustering-based iaas cloud monitoring. In *In Proceedings 10th IEEE International Conference on Cloud Computing (CLOUD)*, 2017.

[18] Mahmoud Abdelsalam, Ram Krishnan, Yufei Huang, and Ravi Sandhu. Malware detection in cloud infrastructures using convolutional neural networks. In *2018 IEEE 11th International Conference on Cloud Computing (CLOUD)*, pages 162–169. IEEE, 2018.

[19] Mahmoud Abdelsalam, Ram Krishnan, and Ravi Sandhu. Online malware detection in cloud auto-scaling systems using shallow convolutional neural networks. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 381–397. Springer, Cham, 2019.

[20] Andrew McDole, Mahmoud Abdelsalam, Maanak Gupta, and Sudip Mittal. Analyzing CNN based behavioural malware detection techniques on cloud IaaS. In *International Conference on Cloud Computing (CLOUD)*, pages 64–79. Springer, 2020.

[21] Andrew McDole, Maanak Gupta, Mahmoud Abdelsalam, Sudip Mittal, and Mamoun Alazab. Deep learning techniques for behavioral malware analysis in cloud iaas. In *Malware Analysis using Artificial Intelligence and Deep Learning*, pages 269–285. Springer, 2021.

[22] Meltem Ozsoy, Caleb Donovick, Iakov Gorelik, Nael Abu-Ghazaleh, and Dmitry Ponomarev. Malware-aware processors: A framework for efficient online malware detection. In *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*, pages 651–661. IEEE, 2015.

[23] Zhixing Xu, Sayak Ray, Pramod Subramanyan, and Sharad Malik. Malware detection using machine learning based analysis of virtual memory access patterns. In *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2017.

[24] Joel A Dawson, Jeffrey T McDonald, Lee Hively, Todd R Andel, Mark Yampolskiy, and Charles Hubbard. Phase space detection of virtual machine cyber events through hypervisor-level system call analysis. In *Data Intelligence and Security (ICDIS), 2018 1st International Conference on*, pages 159–167. IEEE, 2018.

[25] Patrick Luckett, J Todd McDonald, and Joel Dawson. Neural network analysis of system call timing for rootkit detection. In *2016 Cybersecurity Symposium (CYBERSEC)*, pages 1–6. IEEE, 2016.

[26] Mamoun Alazab, Sitalakshmi Venkatraman, Paul Watters, and Moutaz Alazab. Zero-day malware detection based on supervised learning algorithms of api call signatures. In *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, pages 171–182. Australian Computer Society, Inc., 2011.