VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

UNIVERSITY OF INFORMATION TECHNOLOGY

FACULTY OF SOFTWARE ENGINEERING

THESIS PROPOSAL

Title: **Deep Learning in Windows Malware Detection**

Advisor: Assoc Prof. Dr. Vũ Thanh Nguyên

Student: Phạm Hữu Danh - 14520134

Degree: Bachelor of Software Engineering

Introduction

Malware is short for malicious software and is typically used as a catch-all term to refer to any software designed to cause damage to a single computer, server, or computer network [4]. In general, malwares are classified into the following categories [2]: worms, viruses, backdoors, Trojan horses, bots, rootkits, and spyware. A single incident of malware can cause millions of dollars in damage, e.g., zero-day ransomware WannaCry has caused world-wide catastrophe, from knocking U.K. National Health Service hospitals offline to shutting down a Honda Motor Company in Japan [1]. Furthermore, malware is getting more sophisticated and more varied each day [9]. Therfore, The detection of malicious software is

an important problem in cyber security, especially as more of society becomes dependent on computing systems.

The current generation of malware detection products typically uses rule-based or signature-based approaches, which require analysts to handcraft rules that reason over relevant data to make detections. This approach has high accuracy, however, these rules are generally specific, and usually unable to recognize new malware even if it uses the same functionality. That is why the need for machine learning-based detection arises.

Machine learning algorithms learn the underlying patterns from a given training set, which includes both malicious and benign samples. These underlying patterns discriminate malware from benign code. Schultz et al. [8] first applied machine learning methods to malware detection. They showed that compared with signature-based methods, machine learning methods yield more accurate classification results.

However, the main drawback of classical machine learning-based approaches is to accomplish accurate detection, since it is difficult to analyze complex and longer sequences of malicious behaviors, especially when malicious and benign behaviors are interposed. In contrast, deep learning models are capable of analyzing longer sequences of system calls and making better decisions through higher-level information extraction and semantic knowledge learning [11].

Motivation

Many deep learning-based malware detection methods was presented, i.e., Tian et al. [10] recorded the API sequences of binaries using the tool HookMe and proposed a scalable approach for distinguishing malicious files using the features extracted from logs of various API calls; Saxe and Berlin [7] used a deep feed-

forward neural network consisting of four layers with binary features of Windows portable executable (PE) files to detect malware; Yuxin et al. [12] represented malware as opcode sequences and detect it using a deep belief network which can use unlabeled data to pretrain a multi-layer generative model.

Understanding how machine learning works in general and keeping track of state-of-the-art approaches emerging in the cybersecurity field can help organizations cope well with the increased sophistication and complexity of cyber attacks, especially those performed by advanced persistent threats (APT), which are multi-module, stealthy, and target- focused [11].

Objective

The main objective of this thesis is to build a malware detection system for Windows platform based on the recommended deep learning methods, as well as the guidelines for its implementation. Specifically, I will build the system with a deep learning model which uses both features extracted from the Cuckoo Sandbox [3] and raw byte sequences [5].

The dataset used for this thesis was published in Microsoft Malware Classification Challenge 2015 [6], which is almost half a terabyte when uncompressed and consists of a set of known malware files representing a mix of 9 different families.

Besides, the study performed can be useful as a base for further research in the field of malware analysis with machine learning methods.

Research timelines

05 Mar - 15 Apr 2018	Research about machine learning / deep learning methods in malware detection.
16 Apr - 15 May 2018	Conduct experiments on various datasets and configurations.
16 May - 15 Jun 2018	Build applications.
16 Jun - 30 Jun 2018	Completing experiments and writing thesis.

Approved by the advisor

Signature of advisor

Assoc Prof. Dr. Vũ Thanh Nguyên

Ho Chi Minh City, March 26, 2018

Signature of student

Phạm Hữu Danh

References

[1] Chen, Q., and Bridges, R. A. Automated behavioral analysis of malware A case study of wannacry ransomware. *CoRR* (2017).

[2] Egele, M., Scholte, T., Kirda, E., and Kruegel, C. A survey on automated dynamic malware-analysis techniques and tools. *ACM computing surveys (CSUR)* (2012).

[3] Guarnieri, C., Schloesser, M., Bremer, J., and Tanasi, A. Cuckoo sandbox-open source automated malware analysis. *Black Hat USA* (2013).

[4] Moir, R. Defining malware: Faq. *Microsoft Windows Server* (2003).

[5] Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., and Nicholas, C. Malware Detection by Eating a Whole EXE. *ArXiv e-prints* (2017).

[6] Ronen, R., Radu, M., Feuerstein, C., Yom-Tov, E., and Ahmadi, M. Microsoft Malware Classification Challenge. *ArXiv e-prints* (2018).

[7] Saxe, J., and Berlin, K. Deep neural network based malware detection using two dimensional binary program features. *CoRR* (2015).

[8] Schultz, M. G., Eskin, E., Zadok, E., and Stolfo, S. J. Data mining methods for detection of new malicious executables. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy* (Washington, DC, USA, 2001), SP '01, IEEE Computer Society, pp. 38–.

[9] Shahi, G., Pang, E., and Fong, P. *Technology in a Changing World.* Lulu Enterprises Incorporated, 2009.

[10] Tian, R., Islam, R., Batten, L., and Versteeg, S. Differentiating malware from cleanware using behavioural analysis. In *2010 5th International Conference on Malicious and Unwanted Software* (2010).

[11] Yuan, X. Phd forum: Deep learning-based real-time malware detection with multi-stage analysis. In *2017 IEEE International Conference on Smart Computing (SMARTCOMP)* (2017).

[12] Yuxin, D., and Siyi, Z. Malware detection based on deep learning algorithm. *Neural Computing and Applications* (2017).