Malware Detection using Deep Learning based Images Analysis

Ipshita Roy Chowdhury
Advisor: Dr Deepayan Bhowmik
Email: ipshita.roy.chowdhury@stir.ac.uk, deepayan.bhowmik@stir.ac.uk
Division of Computing Science and Mathematics, University of Stirling

SICSA CONFERENCE 2021: Sustainability and Resilience

Abstract

- Malware can compromise a computer system in any form such as Internet worms, viruses, Trojan horses etc. Malware writers always adopt new sophisticated techniques to generate various types of malicious codes which can develop **sophisticated anti-detection techniques**.
- There are three types of malware depending upon the code structure: Basic, Polymorphic and Metamorphic. Advanced techniques like **Camouflage, Packing or code obfuscation** are adopted by the malware developers to make the code unreadable and avoid detection.
- Image and visualization based detection methods are useful to analyze malicious codes. Deep Learning based methods are being widely used to detect malware variants and code visualization.
- **Attention mechanism and Neural Network based architectures** can be effective to detect structural changes of malware and classification from a set of malware images.

Introduction

- Because of ubiquitous nature of Internet and cyber systems there are always new emerging cyber threats that are causing significant growth of new generation malware and attack variants. Web applications, mobile platforms and social networking cites are constantly making the end-users highly vulnerable to novel malware attacks.
- There exists a wide variety of malware types, including Trojan horses, ransomware, viruses, spyware, adware, worms, DDoS, zombies, backdoors, and so on. As a result, it's really a great challenge for the anti-malware companies to release new patches within a short period of time.

Dataset Collection

- **Malimg Dataset:** An image-based benchmark malware dataset was collected to classify the malware families.
- The dataset contains 9339 images of 25 Malware Families.

Motivation

According to AV test report there is a constant distribution of Windows malware and PUA (Potentially Unwanted Applications) per year (Figure 1) Here are few Statictics of Windows malware from AV test report.

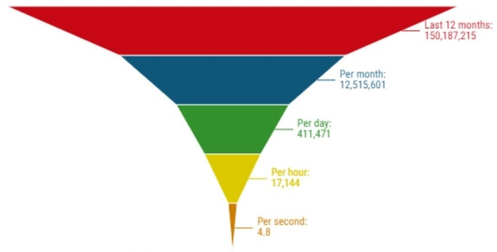

Figure 1:Windows malware and PUA

Figure 2 is showing the growth of malware and PUA in every year.

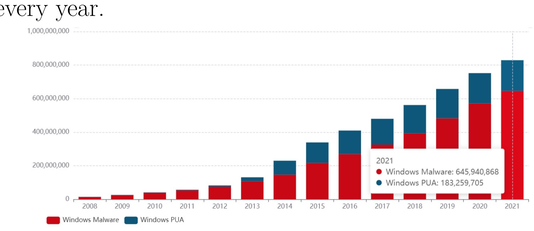

Figure 2:Growth of malware and PUA

Experiment

A sample of the dataset is shown in Figure 3.

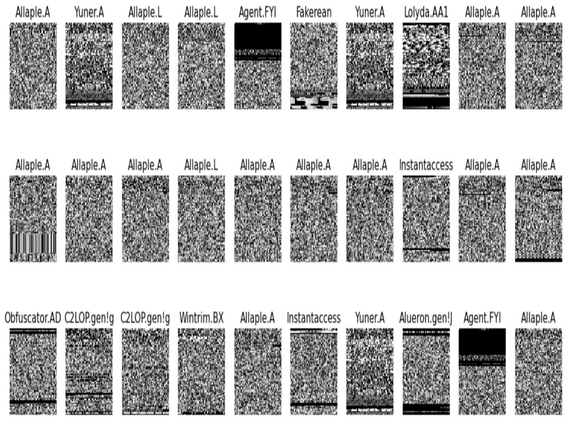

Figure 3:Sample of Malimg Dataset

Experiment & Results-Using CNN

Convolutional Neural Network was used to deal with the unbalanced data and to build a model using Keras.

```
{'Adialer.C': 0,
 'Agent.FYI': 1,
 'Allaple.A': 2,
 'Allaple.L': 3,
 'Alueron.gen!J': 4,
 'Autorun.K': 5,
 'C2LOP.P': 6,
 'C2LOP.gen!g': 7,
 'Dialplatform.B': 8,
 'Dontovo.A': 9,
 'Fakerean': 10,
 'Instantaccess': 11,
 'Lolyda.AA1': 12,
 'Lolyda.AA2': 13,
 'Lolyda.AA3': 14,
 'Lolyda.AT': 15,
 'Malex.gen!J': 16,
 'Obfuscator.AD': 17,
 'Rbot!gen': 18,
 'Skintrim.N': 19,
 'Swizzor.gen!E': 20,
 'Swizzor.gen!I': 21,
 'VB.AT': 22,
 'Wintrim.BX': 23,
 'Yuner.A': 24}
```

Figure 4:Classification of Malware Families

The 25 malware classes have been identified in Figure 4. After using CNN the Classification accuracy reached 95.11% (Figure 5).

```
Final CNN accuracy:  0.9511063694953918
```

Figure 5:Classification Accuracy

Conclusion & Future Work

Because of the widespread use of Deep Learning and Computer Vision image and visualization based malware detection and classification approaches are being popular in cyber security field. Convolutional Neural Network is the mostly used algorithm to train and test images in DL based architectures.

- The current work is CNN based classification of malware dataset from given images.
- Other neural network based methods also need to carry out to explore and visualize the dataset.
- Code Visualization can immensely help in malware classification and to identify new malware variants.
- Visualization techniques can help in malicious code detection to predict and detect zero-day attacks.
- Future Work will include to detect malware using Attention mechanism and neural network based techniques. Also alteration of malware codes needs to be identified for polymorphic and metamorphic malware detection.

Information Sources

- Dataset: Malimg
- AV test report: AV test