**Automated PE32 Treat classification using Import Table and Deep Neural Networks**

Visweswaran N

visweswaran.nagasivam98@gmail.com

**Abstract:**

A malware is a computer program which harms the computer in which it gets executed. Malware analysis play a major role in analysing the functionalities and behaviour of the malware. Malware analysis is a slow and tedious process which involves a lot of manual work. Finding the type of the malware will often boost up the analysis process and helps to the researcher to know what the binary is capable of. Usually researchers perform various static analysis techniques to find the category of the malware using various tools like strings, dependency walker etc., But each day there are millions [1] of new malwares gets released classifying them manually is a non-feasible solution. So, in our approach we are going to automate this process using deep neural networks.

**Introduction:**

Malware analysis helps the researchers to find out the functionality of the malware. Malware analysis comprises of two major types,

1. Static Malware analysis.
2. Dynamic Malware analysis,

In static malware analysis the malicious binaries are examined without executing and may be further subjected for reverse engineering using disassembler. Where as in dynamic malware analysis the malicious binary is actually executed in an isolated environment like sandbox to detect its behaviour.

**Keywords:** Malware, Malware Analysis, Static Malware Analysis, Malware Classification.

**Types of malware:**

**1. Backdoor:**

Backdoor is a malicious program which allows the remote attacker to gain access to the victim’s computer.

**2. Downloader:**

The sole purpose of the downloader is to download another malicious program and sometimes execute it.

**3. Keylogger:**

A keylogger is a program which continuously monitors the keystroke of the user. This helps the attacker to steal potential information like email address, password., etc.,

**4. Miners:**

This malicious program will use the resources of the victim’s computer to mine crypto-currency which is used to monetize the attacker’s wallet.

**5. Rouge software:**

Rouge software seems to behave like an original software say, antivirus and will trick the user to buy services which will end up paying to an attacker.

**6. Trojan:**

A trojan is a software which seems to behave like a legitimate program but does malicious activities in the background.

**7. Ransomware:**

A malicious program which encrypts the user’s files (pictures, documents, etc.,) and would demand a ransom for decryption. The ransom is generally collected through cryptocurrency like Bitcoins, Dash, etc.,

**Import Address Table:**

Import Address Table has the information about the functions which are used by an executable. These functions could help you to identify certain functionalities of the malware.

Here is an example of some functions used by a binary:



As you could see some functions like CryptEncrypt will encrypt the data which exhibits the properties of ransomware. So, we could use these functions to make our predictions.

**Dataset Preparation:**

**a. Sample collection:**

In order to prepare our dataset, we need actual malware samples. Various malware samples have been collected from open source GitHub repositories [2], [3] & from a famous website [4]. These repositories do already have most of the malware categorized which will be used for supervised learning.

All the collected samples are stored in a separate directory depending on the category of the malware which helps in labelling of the malware.

**b. Extracting import functions:**

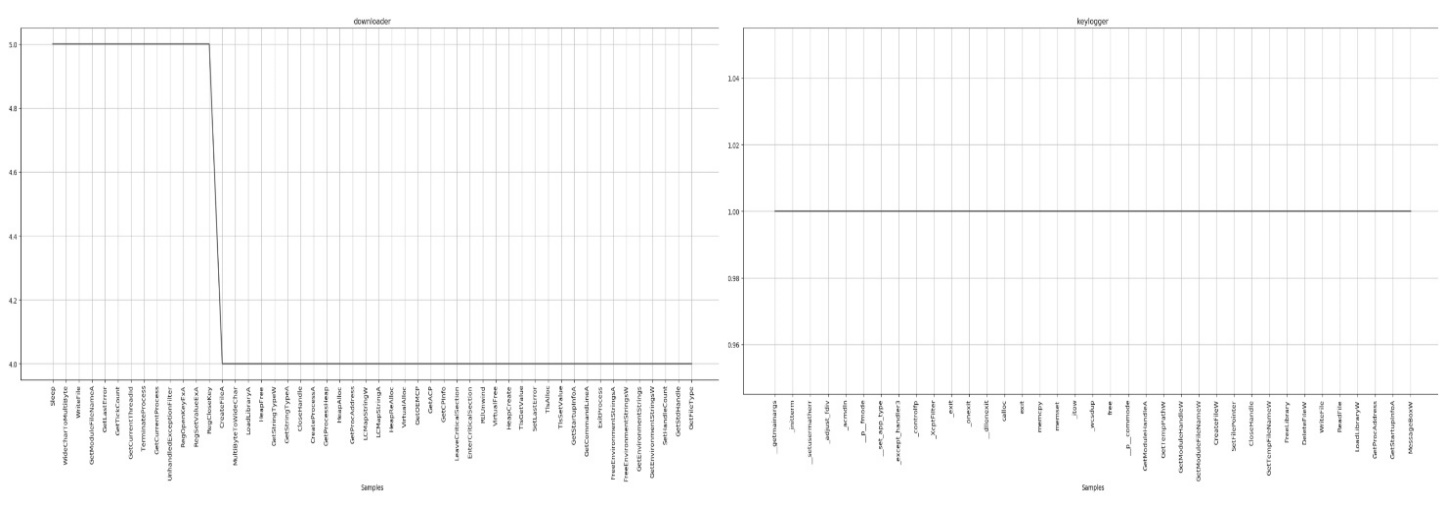
In order to prepare our dataset, we need to extract all the import functions used by the malware. A small C++ program is written which will extract all the imports from all the PE32 files present in the directory. MD5 hashing is used to prevent data duplication. Initially the program will create three separate files one to store the hashes of the scanned malware which is used to prevent data duplication, a separate file to store the imports used by all the executable of the same category and third file is used to notify when the PE32 has used a packer (UPX). Packed program will not be used for dataset preparation however it will be used stored in a separate text file for identification. The program also creates individual files for each PE32 executables containing it’s import with the name of its hash.

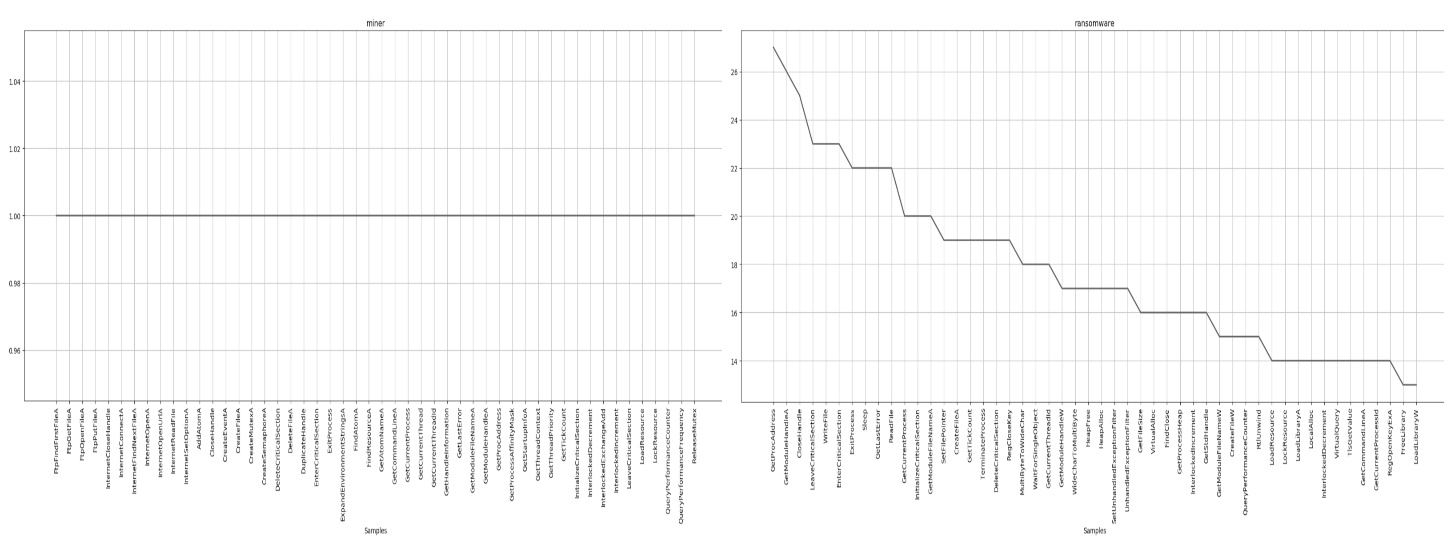
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

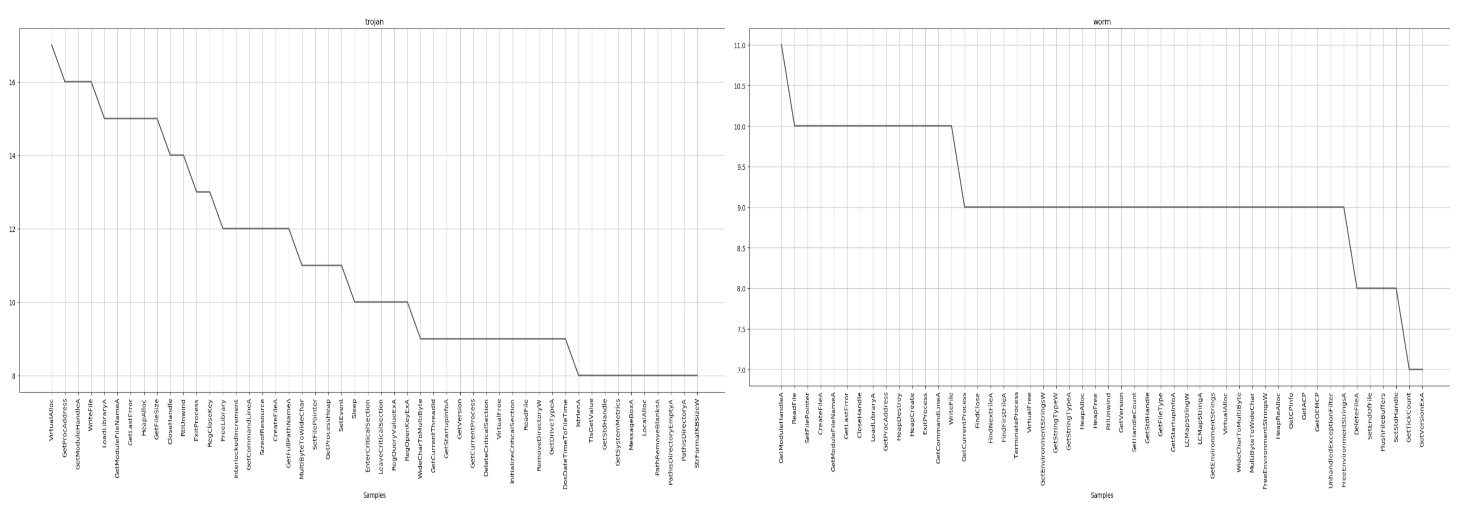
Algorithm:



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_







**c. Compiling the data to dataset:**

By now we only have individual data and yet we need to compile the data into dataset (collection of data) along the labels of the malware. The import functions which has 1728 features is used as the column names and one additional column to include the type of the malware ranging from 0-6. To generate rows for the dataset every malware is iterated and if the import function is present the column is marked with 1 and if not, it will be marked with the 0. The final column will be marked with the type of malware.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Algorithm:

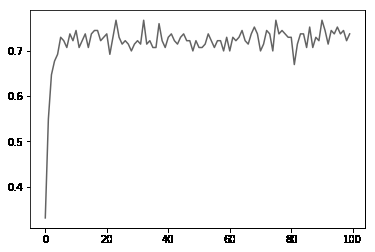


\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Training our model:**

Once the dataset has been prepared, we are ready to train our model. The feed forward neural network model consists of one input layer, two hidden layers and one final output layer each using rectified linear unit as their activation function. The model has been trained with 100 epochs and reached accuracy of more than 70%.

Here is our accuracy graph,



**Conclusion:**

In this research we have concluded that,

1. Import tables play a major role in categorizing the malware.
2. Categorizing the malware can be automated using Deep Neural Networks.

**References:**

1. Virus total statistics: <https://www.virustotal.com/en/statistics/>
2. Fabrizio Monaco, <https://github.com/fabrimagic72/malware-samples>
3. Yuval Nativ, <https://github.com/ytisf/theZoo>
4. J.-M. Roberts. Virus Share. <https://virusshare.com/>