Year 4 Project Proposal

Project Title: MalNet
Student Name: Conor Hanlon
Student ID: 15445378
Stream: CASE
Project Supervisor Name: Ray Walshe

General Outline

In today's modern technology environment, hackers have become significantly more sophisticated with their techniques and approaches to exploiting systems. A common trend that is emerging is a zero-day attack. A zero-day vulnerability is a possible entry route for hackers which is currently unknown by the vendor of the target software. These bugs can be exploited to adversely affect programs, data and systems. This is a severe threat because the application is at risk until the developers discover the zero-day vulnerability and release a patch to fix the issue.

Max Secure Software, one of the world's leading cybersecurity firms, carried out a study and found that machine learning techniques can be applied to detect zero-day vulnerabilities. Unlike traditional security systems, which have to be designed to search for different aspects of malware, machine learning techniques involve training a network to look for similar traits of a malicious file. For my project, I plan to train a neural network to be able to identify malware files from legitimate files using the dataset provided by Max Security Software.

Tech Stack

- **Python:** Programming language with many libraries useful for modelling a neural network as well as training the network.
- **PIP:** Package manager for Python.
- **Tensorflow:** Used to train deep neural network through Python.
- **Numpy:** Python library with functions helpful for training the network.
- **Flask:** Micro web framework for Python that I will use to deploy the neural network model.

- **Docker:** Open source software that provides containerization. I will create a container on an Amazon Web Services EC2 instance to set up and configure the training environment.

Learning Challenges

This project presents a lot of learning challenges for me. It is the first time I will be creating my own neural network from scratch and implementing the training process. The Tensorflow and Flask libraries are both new to me, so I will have to research best practices to ensure that I use them to the best of their ability. It will also be challenging to work with such a large dataset. I will have to ensure that the data is pruned properly to be able to make accurate predictions, which is something I am only beginning to learn about this year in our Data Mining module. Understanding the nature of the malware will be critical in this project. This will play a key part in training the network to identify the most common characteristics of malicious files.

Hardware / Software Requirements

My dataset consists of roughly 136,000 entries consisting of malicious and legitimate files, so it requires massive processing power to train the neural network. I plan to use AWS to obtain this processing power. They provide elastic GPUs which enhance the performance of your EC2 instance running the application. The pricing to rent these GPUs start at $0.05 per hour, which is quite cheap considering the quality of the service that AWS provide.

The development environment will be set up on my local machine.