

Enhancing the Insertion of NOP Instructions to Obfuscate Malware via Deep Reinforcement Learning

Daniel Gibert^{a,*}, Matt Fredrikson^b, Carles Mateu^c, Jordi Planes^c, Quan Le^a

^aUniversity College Dublin, CEADAR, Belfield Office Park, Dublin, Ireland

^bCarnegie Mellon University, Forbes Avenue Pittsburgh, PA 15213, United States of America

^cUniversity of Lleida, Jaume II, 69, Lleida, Spain

Abstract

Current state-of-the-art research for tackling the problem of malware detection and classification is centered on the design, implementation and deployment of systems powered by machine learning because of its ability to generalize to never-before-seen malware families and polymorphic mutations. However, it has been shown that machine learning models, in particular deep neural networks, lack robustness against crafted inputs (adversarial examples). In this work, we have investigated the vulnerability of a state-of-the-art shallow convolutional neural network malware classifier against the dead code insertion technique. We propose a general framework powered by a Double Q-network to induce misclassification over malware families. The framework trains an agent through a convolutional neural network to select the optimal positions in a code sequence to insert dead code instructions so that the machine learning classifier mislabels the resulting executable. The experiments show that the proposed method significantly drops the classification accuracy of the classifier to 56.53% while having an evasion rate of 100% for the samples belonging to the Kelihos_ver3, Simda, and Kelihos_ver1 families. In addition, the average number of instructions needed to mislabel malware in comparison to a random agent decreased by 33%.

Keywords: Malware Classification, Assembly Language Source Code, Obfuscation, Reinforcement Learning, Deep Q-Network

1. Introduction

Malware is on the rise. Global detections of newly-developed malware keep increasing year after year.¹ According to AV-TEST Institute,² there has been a dramatic spike of new malicious programs and potentially unwanted applications (PUA) year after year, doubling the number of total malware detected from 2015 to 2020. This recent surge of malicious programs is connected to the increasing dependency of people, things and organizations on the Internet, which provides cybercriminals with a vast range of targets to exploit, from traditional personal computers and laptops to industrial systems, mobile phones and Internet of Things (IoT) devices. In addition, the adoption of remote working due to global events such as the current global pandemic has propelled a new rise in cyberattacks, particularly the increase of ransomware attacks³.

To keep up with malware evolution and be able to mitigate the damage and impact of cyberattacks, it is necessary to constantly improve the computer systems defences. One

essential security element is endpoint protection, i.e. the use of security solutions to protect endpoints or end-user devices from being exploited against zero-day exploits, attacks, data leakages. Within the wide range of tools to secure an endpoint, anti-malware engines are the last layer of defence. More specifically, anti-malware engines are responsible for preventing, detecting and removing malicious software. Traditionally, anti-malware solutions relied on signature-based and heuristic-based methods. However, due to the huge volumes of new malware variants being deployed every day, traditional anti-malware solutions that rely solely on signatures and heuristics manually defined by domain experts cannot keep pace with the rapidly evolving malware.

During the last decade, research on machine learning (ML) solutions to tackle the problem of malware detection and classification increased because of the ability of machine learning systems to generalize to unseen malware and polymorphic mutations (Souri and Hosseini, 2018; Gibert et al., 2020b; Qiu et al., 2020). Although machine learning models (Ahmadi et al., 2015; Zhang et al., 2016) and in particular deep neural networks (Gibert et al., 2017; McLaughlin et al., 2017; Raff et al., 2018; Krčál et al., 2018; Gibert et al., 2019; Vinayakumar et al., 2019) have achieved significant success in the cybersecurity domain, they have been shown to be vulnerable to adversarial examples (Grosse et al., 2016; Demetrio et al., 2019; Anderson et al., 2018), i.e. modified examples with imperceptible perturbations that cause misclassification at test time.

The attacks presented in the literature mainly focus on mi-

*I am corresponding author

Email address: daniel.gibert@ucd.ie (Daniel Gibert)

URL:

[https://scholar.google.com/citations?user=1AAwRpMAAAAJ&hl=Daniel Gibert](https://scholar.google.com/citations?user=1AAwRpMAAAAJ&hl=Daniel%20Gibert)

¹<https://www.statista.com/statistics/680953/global-malware-volume/>

²<https://www.av-test.org/en/?r=1>

³<https://pages.checkpoint.com/cyber-attack-2021-trends.html>

nor modifications of the PE header, appending some bytes at the end of sections, inside code caves, at the end of the PE file, modifying the PE Headers, etcetera. However, these modifications are not usually generated by common obfuscation techniques employed by malware authors in a real world scenario but specifically designed to evade specific ML-based anti-malware engines. On the contrary, malware authors employ a variety of techniques to obfuscate the code and make it more difficult for humans to understand. Common obfuscation techniques are the dead code insertion technique, instruction replacement, code transposition, packing, encryption, among others.

The purpose of this work is to fill this gap. Accordingly, we have analyzed the robustness of a state-of-the-art malware classifier trained on the assembly language instructions of Portable Executables (Gibert et al., 2017; McLaughlin et al., 2017) against the simplest of the obfuscation techniques, the dead code insertion technique. Furthermore, the aforementioned technique has been enhanced with deep reinforcement learning to increase its evasion success rates. To this end, we present a framework that uses Double Q-learning to induce misclassification over malware families. Within this framework, an AI agent is set up to confront the malware classifier and select a sequence of functionality-preserving actions (dead code insertions) to modify the samples. For any given malware sample, the framework can eventually determine the optimal sequence of actions to make the ML-based classifier output an incorrect label. In this study, we have used the Portable Executable files provided by Microsoft for the Big Data Innovators Gathering challenge of 2015 for reproducibility purposes (Ronen et al., 2018).

The rest of the paper is organized as follows: Section 2 introduces the state-of-the-art approaches for malware detection and classification, and the adversarial attacks devised to bypass detection. Section 3 describes the proposed framework in detail. Section 4 presents the experimental setup and the results obtained. Section 5 summarizes the concluding remarks extracted from this work and presents some future lines of research.

2. Related Work

2.1. Machine Learning for Malware Detection and Classification

Recently, machine learning (ML) has become an appealing signature-less approach for malware detection and classification because of its ability to handle huge volumes of data and to generalize to never-before-seen malware (Gibert et al., 2020b; Souri and Hosseini, 2018; Qiu et al., 2020). Research has shifted from traditional approaches based on feature engineering (Ahmadi et al., 2015; Zhang et al., 2016) to deep learning approaches (Krčál et al., 2018; Gibert et al., 2017; Raff et al., 2018; Gibert et al., 2018, 2019; Vinayakumar et al., 2019; Venkatraman et al., 2019; Gibert et al., 2020a) because it allows to obviate and replace the time-consuming feature extraction process by an end-to-end system, which typically consists of a neural network with multiple layers, that performs both

feature learning and classification altogether. With deep learning, one can start with raw data as features will be automatically learned by the network through training on the labelled data. For instance, Gibert et al. (2017) and McLaughlin et al. (2017) presented a shallow convolutional neural network architecture to classify malware based on the assembly language instructions extracted from the assembly language source code of malware from PE executables and Android APKs, respectively. Similarly, Raff et al. (2018) and Krčál et al. (2018) designed a shallow and deep convolutional neural networks, respectively, to detect malware based on the bytes content. Instead of taking as input the byte sequence, which could consist of a few million time steps, Gibert et al. (2018) presented a method for classifying malware by compressing its binary content as a stream of entropy values (or structural entropy) using convolutional neural networks. In other works (Gibert et al., 2019; Vinayakumar et al., 2019), the executables are represented as a grayscale image by interpreting every byte as one pixel in an image, with values ranging from 0 to 255 (0:black, 255:white). Recently, researchers (Venkatraman et al., 2019; Gibert et al., 2020a) have started complementing traditional features with features extracted through deep learning. The reason behind this multimodal approach is that each executable includes multiple modalities of information. By only taking as input the raw bytes or opcodes, a lot of information for classification is overlooked. As a result, multiple types of features provide a better abstract representation of the executable characteristics, leading to more accurate ML models.

2.2. Adversarial Attacks on ML-based Malware Detectors

Although machine learning has demonstrated impressive performance in several application domains, ranging from computer vision and natural language processing to cybersecurity, ML models have been shown to be vulnerable to adversarial examples (Ren et al., 2020; Papernot et al., 2016), i.e. inputs to the models that an attacker has carefully designed to cause the model to make a mistake, and the domain of cyber security is no exception (Grosse et al., 2016; Demetrio et al., 2019; Kolosnjaji et al., 2018; Anderson et al., 2018). In addition, in the cyber security domain there really exist actual adversaries, the malware developers, who are strongly motivated to craft their malicious softwares to bypass detection systems in order to steal user information, spread across the network, or perform other hostile activities. Next, some relevant adversarial attack approaches in the literature are presented.

Grosse et al. (2016) proposed a gradient-based approach to generate adversarial Android malware examples to evade their own malware detector based on features extracted from (1) permissions and hardware components access requested, (2) API calls made by the application, (3) intents used to communicate with other applications and (4) application components, service, content provider and broadcast receivers used by the applications. They proposed a crafting process that iteratively modifies the feature whose gradient is the largest until the adversarial sample bypasses the detection model.

Suciu et al. (2019) explored the vulnerabilities of byte-based malware detectors to adversarial malware binaries and presented

various strategies to bypass binary detectors by appending a few bytes at the end of the PE headers and sections. Similarly, Kolosnjaji et al. (2018) presented an approach to evade malware detectors by appending a set of carefully-handpicked bytes at the end of the file. In addition, Demetrio et al. (2019) refined the aforementioned technique by using feature attribution to identify the most influential input features contributing to each decision and thus, reducing the number of bytes needed to be manipulated in order to evade the malware detector.

Alternatively, Anderson et al. (2018) presented a pioneering attack using reinforcement learning to craft adversarial examples using a series of actions to manipulate a malware binary. In their work, they trained an agent to evade a malware detector based on features extracted from the PE header and sections metadata, the import and export tables, counts of human readable strings, the byte histogram and a 2D byte-entropy histogram of the executable. To do so, they allowed the agent to perform various mutations on the executable that do not break its file format or alter the code execution, such as appending bytes at the end of sections, adding functions to the import address table. However, as their results show, the evasion rate between their agent and a random agent is practically the same, mainly because the nature of the mutations implemented, being the use of packing the most successful mutation as it is the technique that affects the most features used for training the ML detector.

The aforementioned attacks (Suciu et al., 2019; Kolosnjaji et al., 2018; Demetrio et al., 2019; Anderson et al., 2018) mainly rely on appending some bytes at the end of the sections or at the end of the PE file. However, these attacks are quite different from those employed by malware authors in a real world scenario to obfuscate the executables. Common obfuscation techniques are the dead code insertion technique, the instruction replacement technique and the subroutine reordering technique, just to name a few. Subsequently, in this work, we address this problem by evaluating the robustness of a state-of-the-art malware classifier (Gibert et al., 2017, 2021) against the simplest obfuscation technique, the dead code insertion technique, and we present a framework to enhance its performance. This classifier has been selected because of its superior performance with respect to deep learning byte-based approaches in the literature (Krčál et al., 2018; Raff et al., 2018).

3. Reinforcement Learning Framework

In this paper, we present a Proof of Concept (PoC) of a deep reinforcement learning framework to induce misclassification of malicious Portable Executable (PE) files over malware families.

Reinforcement learning (Kaelbling et al., 1996) is a branch of machine learning where an agent seeks to learn optimal decision-making by trying to maximize cumulative rewards to achieve a specific goal. The two main components are the environment and the agent. On the one hand, the environment represents the problem to be solved. On the other hand, the agent represents the learning algorithm used to perform actions in the environment in order to maximize the rewards. Thus, reinforcement

learning (RL) algorithms study the interaction of agents in such environments and learn to optimize the behavior of the agents in order to maximize the rewards. The learning process consists of an agent and an environment that continuously interact with each other, as shown in Figure 1. At each time step, the agent takes action a_t on the current observation of the environment s_t , based on its policy $\pi(a_t, s_t)$, and receives a reward r_{t+1} and the next observation of the environment s_{t+1} .

The problem of inducing misclassification of malware samples can be framed as a Markov Decision Process (Bellman, 1957), or MDP, which is a formalism that allows to define the interaction between the agent and environment as a tuple of four elements (S, A, T, R):

- **S**: Set of states. At each time step, the state of the environment is an element $s \in S$, where s_t denotes the state of the environment at time step t .
- **A**: Set of actions. At each time step, the agent chooses an action $a \in A$, where a_t denotes the action chosen at time step t . The set of actions that can be performed on a particular state $s \in S$, is denoted $A(s)$, where $A(s) \subseteq A$. Note that in some systems not all actions can be applied in every state. This occurs in our case, as there are positions in the assembly language source code that we will not be able to insert dead instructions because it will break the executable.
- **$T(s_t, a_t, s_{t+1})$** : State transition function that describes how the environment’s state changes when the agent performs an action in a given state. By applying action $a_t \in A$ in a state $s_t \in S$, the system makes a transition from state s_t to a new state $s_{t+1} \in S$, based on a probability distribution over the set of possible transitions. The transition function T is defined as the probability of ending up in state s_{t+1} after performing action a_t in state s_t .
Assuming that the system is Markovian, i.e. the result of an action does not depend on previous actions and previous states, but only depends on the current state, the transition function can be mathematically described as follows:
$$T(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \dots) \quad (1)$$
- **$R(s_t, a_t, r_{t+1})$** : Reward model used to describe the reward value the agent receives from the environment after transitioning from state $s \in S$ to the new state $s' \in S$ with action $a \in A$.

To sum up, at each time step, the agent will receive some state of the environment $s \in S$. Given this representation, the agent selects an action to take. Then, the environment transitions to a new state and the agent is given a reward as a consequence of its previous actions. This continuous interaction of the agent with the environment creates a trajectory of states, actions and rewards, which can be interpreted as:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \dots$$

Thus, the goal of the agent is to derive a policy π , the strategy that the agent will pursue, so as to maximize the total amount of rewards it receives over the course of action. In consequence, the agent does not want to maximize the immediate rewards but the cumulative rewards that it will receive over time. At a given time step t , the cumulative reward is defined as:

$$G_t = \sum_{i=0}^{T-t-1} \gamma^i R_{t+i+1} \quad (2)$$

where T is the final time step and γ is the discount factor used to control the importance of future rewards.

The policy π is a function that maps a given state s to probabilities of selecting each possible action from that state. In other words, for each state $s \in \mathcal{S}$, π is a probability distribution over $a \in A(s)$. Most MDPs derive optimal policies by learning value functions. There are two types of value functions: (1) functions of states, or (2) state-action pairs, that estimate how good it is for the agent to be in a given state, or how good it is for the agent to perform a given action in a given state, respectively, where the quality of a state or a state-action pair is given in terms of expected return. Considering that the rewards an agent expects to receive are dependent on the actions it takes in given states and that these are influenced by the policy the agent is following, we can see that the value functions are defined with respect to policies.

So, the state-value function for a given policy π , denoted as v_π , indicates how good any given state is for an agent following policy π . In other words, the value of state s under policy π is the expected return from starting from state s at time t and following π thereafter. Mathematically, $v_\pi(s)$ is defined as:

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} | S_t = s\right] \quad (3)$$

Similarly, the action-value function for a given policy π , denoted as q_π , indicates how good it is for the agent to take any given action from a given state while following policy π . In other words, the value of action a in state s under policy π is the expected return from starting from state s at time t , taking action a , and following policy π thereafter. Mathematically, $q_\pi(s, a)$ is defined as:

$$q_\pi(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} | S_t = s, A_t = a\right] \quad (4)$$

It is also common to refer to the action-value function q_π as the Q-function while the output from the Q-function for any given state-action pair (s, a) is referred to its Q-value. In other words, the Q-value associated with a state-action pair (s, a) represents the quality of taking action a in state s .

As already defined, the goal of the agent is to derive a policy π that maximizes the total amount of rewards received. That is, the policy that yields more return to the agent than all the other policies.

In terms of return, given two policies π and π' , policy π is considered to be better than or the same as policy π' if the expected return of π is greater than or equal to the expected return of policy π' for all states.

$$\pi \geq \pi' \text{ if and only if } v_\pi(s) \geq v_{\pi'}(s) \text{ for all } s \in \mathcal{S} \quad (5)$$

A policy that is better than or at least the same as all other policies is called the optimal policy, and will be denoted π_* from now on.

The optimal policy has an associated optimal state-value function, denoted as v_* and defined as

$$v_*(s) = \max_{\pi} v_\pi(s) \quad (6)$$

for all $s \in \mathcal{S}$. In other words, v_* gives the largest expected return achievable by any policy π for each state.

Similarly, the optimal policy has an optimal action-value function, or optimal Q-function, which we denote as q_* and define as

$$q_*(s, a) = \max_{\pi} q_\pi(s, a) \quad (7)$$

for all $s \in \mathcal{S}$ and $a \in A(s)$. In other words, q_* gives the largest expected return achievable by any policy π for each possible state-action pair.

One fundamental property of q_* is that it must satisfy the following equation.

$$q_*(s, a) = \mathbb{E}[R_{t+1} + \gamma \max_{a'} q_*(s', a')] \quad (8)$$

This is called the Bellman optimality equation. It states that, for any state-action pair (s, a) at time t , the expected return from starting in state s , selecting action a and following the optimal policy π_* thereafter is going to be the expected reward we get from taking action a in state s , which is R_{t+1} , plus the maximum expected discounted return that can be achieved from any possible next state-action pair (s', a') . Since the agent is following the optimal policy, the following state s' will be the state from which the best possible next action a' can be taken at time $t+1$.

As a result, the optimal policy can be determined from the optimal state-value function q_* , because once q_* has been solved, we can determine the optimal policy by finding the action a that maximizes $q_*(s, a)$. Mathematically, this can be written as follows:

$$\pi_*(s) = \arg \max_a q_*(s, a) \quad (9)$$

In other words, the best action from any given state is the action that has the highest expected return based on the possible next states resulting from taking that action.

One algorithm to find the optimal Q-values for each state-action pair is the Q-learning algorithm (Watkins and Dayan, 1992). This algorithm iteratively updates the Q-values for each state-action pair using the Bellman equation until the Q-function converges to the optimal Q-function q_* . In the Q-learning algorithm, the Q-function is stored as a table with each cell corresponding to the Q-value of an action a in a given state s . Then,

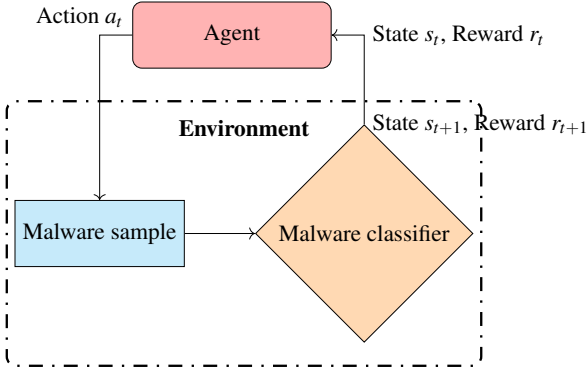


Figure 1: Reinforcement learning schema for evading a malware detector. At time step t , the agent takes as input the state s_t and the reward r_t . It selects the best action a_t to perform, i.e. the location to insert the NOP instruction that has a higher Q-value, and modifies the malware sample. Afterwards, it checks whether or not the mutated sample produces a misclassification. If not, the procedure is repeated until the sample is misclassified or it reaches the maximum number of modifications allowed to be performed by the agent before declaring failure.

the Q-values are updated iteratively according to the following update rule:

$$Q(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_a Q(s_{t+1}, a)) \quad (10)$$

Given infinite exploration time and a partly-random policy, it has been proved that the Q-learning algorithm can derive the optimal policy for any given finite MDP. However, although the Q-learning algorithm works well in very simple environments with a very limited number of actions and states such as the Frozen Lake⁴, CartPole-v1⁵, among others, this approach is totally impractical in complex environments as it requires having a finite state and action spaces and storing the full state-action table in memory (which is often unfeasible). As a result, the Double Q-learning algorithm (Mnih et al., 2013, 2015; Hasselt et al., 2016) with experience replay (Lin, 1992) has been implemented to estimate the optimal Q-function. For a more detailed description of the algorithm, we refer the readers to Section 3.3.

Next, Section 3.1 describes the environment, and Section 3.2 describes the action space. Afterwards, Section 3.3 describes the reinforcement learning algorithm used to approximate the optimal Q-function and the convolutional neural network architecture used to determine the best position to which insert the NOP instructions in order to bypass the malware classifier. Lastly, Section 3.5 briefly resumes the machine learning classifier that we aim to bypass by modifying the assembly language source code of the executable with the insertion of NOP instructions.

3.1. Environment

The environment consists of an initial malware sample (one malware sample per episode) and the malware classifier (the attack target). Each time step or turn within an episode provides the following feedback to the agent:

- A reward value $r_t \in \mathbb{R}$ given for mislabeling the malware family. The reward value r_t at time step t is equal to the difference between the loss of the previous state and the loss of the current state of the classifier that we want to evade.

$$r_t = -1 * (loss_{t-1} - loss_t) \quad (11)$$

where $loss_t$ refers to the multi-class logarithmic loss or cross entropy loss returned by the classification model for state s_t . Thus, if the reward at time step t , r_t , is positive, it indicates that the action performed at time step t , a_t , has increased the error loss of the target classifier. On the other hand, if the reward at time step t , r_t , is negative then, the action at time step t , a_t has negatively contributed to the misclassification of the current executable.

- The state s_t of the environment (malware sample) is represented as a sequence of assembly language instructions extracted from the assembly language source code of the Portable Executable file, whereas the assembly language source code can be obtained by disassembling the Portable Executable file using any disassembler of your choice, i.e. IDA Pro⁶, Radare2⁷, Ghidra⁸. One particularity of using the sequence of instructions to represent an executable is that the resulting length of the sequence will differ from one executable to the other. In addition, for each instruction added to the executable, the size of the resulting sequence of instructions will be incremented by one. Instead of taking the assembly language source code as a whole, including its arguments, we decided to only take as input the mnemonics of the instruction. The mnemonic of a given instruction refers to the portion of the machine language instruction that specifies the operation to be performed, i.e. on encountering the instruction `lea eax, [esp+8]`, we simply take as input the opcode `lea`.

Based on the feedback provided by the reward, the agent learns which action to choose from a set of mutations (See Section 3.2) given the environment’s state (sequence of assembly language instructions), while preserving the format and function of the PE file.

3.2. Action Space

The mutations represent the actions or moves available to the agent within the environment. Formally, the set of all possible actions is defined as $A = \{insert_0, insert_1, \dots, insert_n\}$, where $insert_i$ refers to inserting a *NOP* instruction at position i in the assembly language instruction sequence. Therefore, at time t , the agent chooses action a_t . In this paper, we only considered as valid actions the insertion *NOP* instructions as it is the simplest dead code instruction available (Balakrishnan and Schulze, 2005). The *NOP* or no-op instruction (short for no operation) is an assembly language instruction that does nothing. An example is provided in Figure 2. In addition, the number

⁴<https://gym.openai.com/envs/FrozenLake-v0/>

⁵<https://gym.openai.com/envs/CartPole-v1/>

⁶<https://www.hex-rays.com/ida-pro/>

⁷<https://rada.re/n/>

⁸<https://ghidra-sre.org/>

of actions available differs for each sample because it depends on the size and the content of their assembly language source code. Take into account that the locations to which the NOP instructions are inserted must not break the executable.

```
.text:00408AA0 ; ===== S U B R O U T I N E =====
.text:00408AA0
.text:00408AA0
.text:00408AA0 ; int __stdcall sub_408AA0(void *Src)
.text:00408AA0 sub_408AA0 proc near ; CODE XREF: sub_455FBC[]j
.text:00408AA0
.text:00408AA0 Src = dword ptr 4
.text:00408AA0 0B 54 24 04 mov edx, [esp+Src]
.text:00408AA4 56 push esi
.text:00408AA5 8B F1 mov esi, ecx
.text:00408AA7 8B C2 mov eax, edx
.text:00408AA9 57 push edi
.text:00408AAA C7 46 18 0F 00 00 00 mov dword ptr [esi+18h], 0Fh
.text:00408AB1 C7 46 14 00 00 00 00 00 mov dword ptr [esi+14h], 0
.text:00408AB8 C6 46 04 00 mov byte ptr [esi+4], 0
.text:00408ABC 8D 78 01 lea edi, [eax+1]
.text:00408ABF 90 nop
```

Figure 2: Snapshot of a piece of assembly language source code with a NOP instruction inserted at the location 0x00408ABF.

3.3. Agent

The agent takes the assembly language instructions of the sample and executes the action that will eventually yield the highest cumulative reward. This action is selected according to the Q-values provided by a deep Q-network (DQN). See Figure 3. For each given input state, the Q-network outputs the estimated Q-values for each action that can be taken from that state.

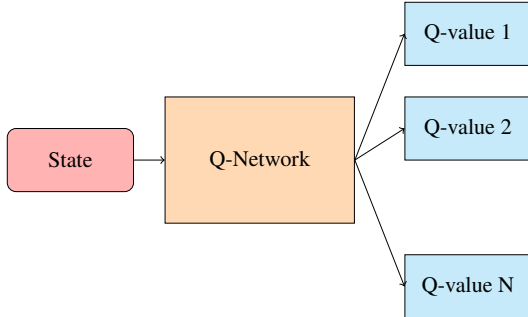


Figure 3: Black-box representation of the Q-Network. The Q-Network takes as input the state, i.e. the assembly language source code of malware, and outputs the estimated Q-values of inserting a NOP instruction to all locations of the source code.

To train this network, we used the Double Q-learning algorithm (Hasselt et al., 2016), which decouples the action selection from the target Q-value generation, with experience replay (Lin, 1992). The Double Q-learning algorithm uses two networks, the primary network, Q , and the target network, Q_{target} , to select what is the best action to take for the next state and to calculate the target Q-value of taking that action at the next state, respectively. The weights of the primary network at time step t are denoted as θ while the weights of the target network at time step t are denoted as θ' . Notice that the primary network and the target network architectures are the same. The

only difference between them is their weights. Mathematically, this can be formulated as follows:

$$Q_*(s_t, a_t) \approx r(s_t, a_t) + \gamma Q_{target}(s_{t+1}, \arg \max_a Q(s_{t+1}, a; \theta); \theta') \quad (12)$$

The update of the target network stays unchanged from the primary network, and it slowly copies the weights of the primary network Q using the Polyak averaging:

$$\theta' \leftarrow \tau \theta + (1 - \tau) \theta' \quad (13)$$

To train Q-networks it is used a technique called experience replay during training (Lin, 1992). The idea behind is to store the agent's experiences at each time step in a replay memory data set and use these experiences to update the weights of the Q-network through gradient descent.

At a time step t , the agent's experience denoted as e_t is defined as a tuple of four elements containing the state of the environment s_t , the action a_t taken from state s_t , the reward r_{t+1} the agent receives at time step $t + 1$ as the result of performing action a_t on the state s_t , and the next state of the environment s_{t+1} .

$$e_t = (s_t, a_t, r_{t+1}, s_{t+1})$$

The advantages of experience replay are various. First, each experience might be potentially used in multiple weight updates allowing for greater data efficiency. Second, previous online learning approaches learned directly from consecutive samples with strong correlations between them. By sampling a subset of samples from the experience replay data set, these correlations are broken and the variance of the updates is reduced (Mnih et al., 2013).

In addition, to balance between exploration and exploitation during training, we use the epsilon-greedy action selection algorithm. This algorithm tackles the exploration-exploitation trade-off by taking an exploratory action with probability ϵ and a greedy action with probability $1 - \epsilon$. Mathematically, the epsilon-greedy action selection algorithm selects an action a_t at time step t as follows:

$$a_t = \begin{cases} \max Q_t(a) & \text{with probability } 1 - \epsilon \\ \text{random action} & \text{with probability } \epsilon \end{cases} \quad (14)$$

In our case, the value of ϵ is set to 1.0 at the beginning of the training, and it is continuously decreased at each time step until reaching 0.5. This has been done to allow the agent to explore more often than not as the state and action spaces in our problem are very large.

For a complete description of the Double Q-learning algorithm, we refer the readers to Algorithm 1, the work of Hasselt et al. (2016), and the references therein.

3.4. Q-Network Architecture

An overview of the Q-network is described in Figure 4. It consists of a convolutional layer to extract features from the sequence of assembly language instructions followed by a time-distributed layer that applies the same fully-connected layer to every time step. The layers description is as follows:

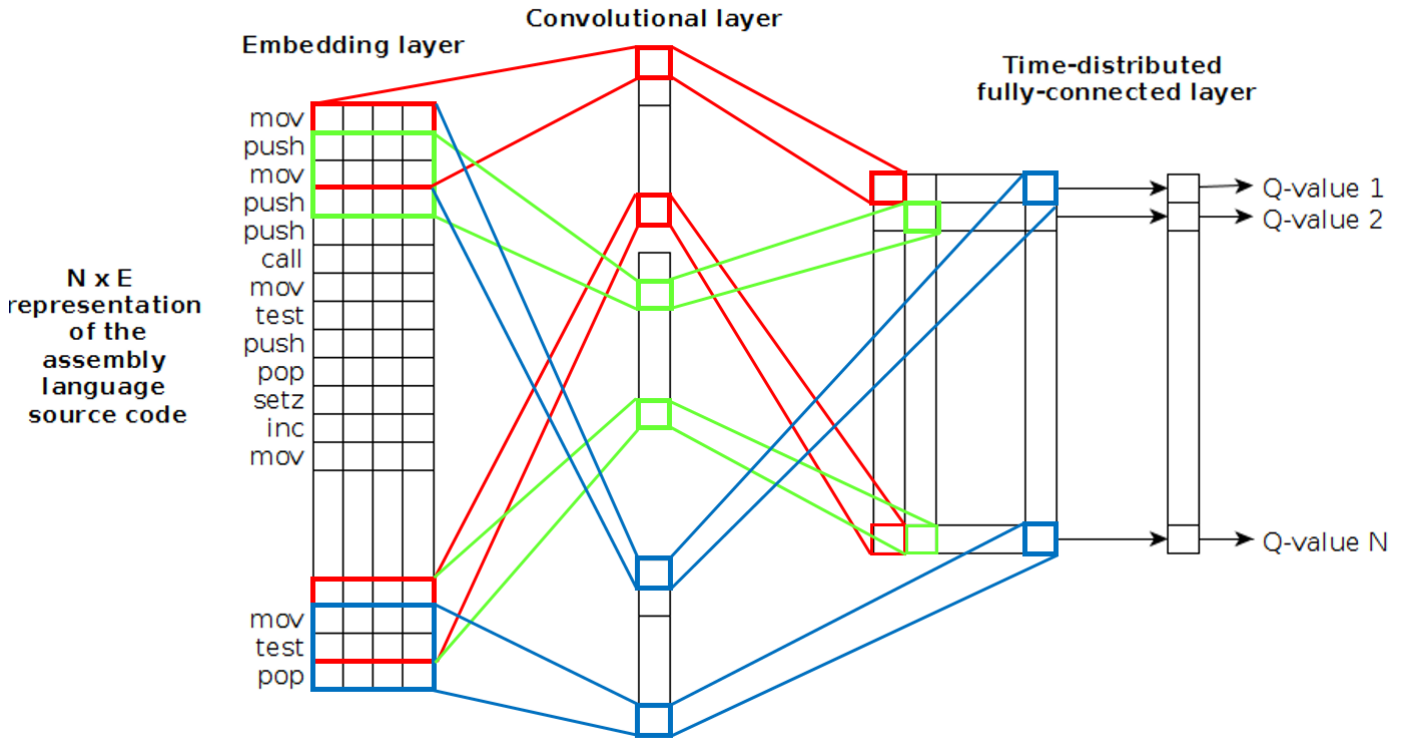


Figure 4: Graphical representation of the shallow convolutional Q-network architecture. The Q-network takes as input a Portable Executable (PE) file represented as a sequence of mnemonics (length N) and represents each mnemonic as a word embedding (size E). Afterwards, a convolutional layer with filters of size 3 is applied to extract features from the sequence of embedded mnemonics, followed by a time-distributed layer to retrieve the Q-value of inserting a NOP instruction at every location in the mnemonics sequence.

- **Input layer.** The network takes as input an executable represented as a sequence of mnemonics. A mnemonic is the name of the operation a machine can execute. For instance, the assembly language instruction `mov ebp, esp` is reduced to the mnemonic `mov`. The main argument behind this simple representation is that it will generalize better as it would not be affected by small permutations in the arguments and thus, the obfuscation technique known as register reassignment would not alter the output of the classifier.
- **Embedding layer.** Each mnemonic is represented as a low-dimensional vector of real values (word embedding) of size E , where each value captures a dimension of the mnemonics' meaning. E was set to 4 as in (Gibert et al., 2017). The rationale behind using distributed representations is to better capture the semantic relationships between comparable mnemonics.
- **Convolutional layer.** The convolutional layer extracts N -gram like features from the sequence of assembly language instructions. The size of each filter is $h \times E$ where $h = 3$. The number of different filters in the convolutional layer is 10.
- **Time-Distributed layer.** This layer applies a fully-connected layer to each of the N mnemonics (N is equal to the size

of the assembly language instructions sequence), independently. The input of the fully-connected layer is equal to the number of filters in the convolutional layer while the number of output neurons o is equal to 1. At the end, the softmax function is applied to output the estimated Q-values of inserting a NOP instruction in any location of the assembly language instructions given as input to the network.

For a complete description of the trainable parameters of the Q-network we refer the readers to Table A.4.

3.5. CNN classifier

The classifier model that our agent is seeking to evade is a shallow convolutional neural network specifically designed to learn N -gram based features from a sequence of text (Gibert et al., 2017, 2021). This classifier has been selected among the deep learning classifiers in the literature because of its state-of-the-art performance (Gibert et al., 2021). See Figure 5 and Table A.5 for a detailed description of the network architecture.

The network takes as input an executable represented as a sequence of mnemonics and convolves various filters of different sizes to extract n -gram like features. In particular, the filters have sizes 3, 5 and 7. Afterwards, a global max-pooling layer is applied to extract the maximum activation of each of the feature map activations from the previous layer. At the end,

Input: Initialize weights θ and θ' of primary and target networks Q and Q_{target} , replay buffer D , $\tau \ll 1$

for each episode do

for each environment step do

Select action a_t using Equation 14
Execute a_t and observe next state s_{t+1} and reward r_t . Store transition $(s_t, a_t, r_{t+1}, s_{t+1})$ in replay buffer D

end

for each update step do

sample $e_t = (s_t, a_t, r_{t+1}, s_{t+1}) \sim D$
Compute target Q-value using Equation 12
Perform gradient descent step on

$$(Q_*(s_t, a_t) - Q(s_t, a_t))^2 \quad (15)$$

Update target network parameters using Eq. 13

end

end

Algorithm 1: Double Q-learning algorithm (Hasselt et al., 2016)

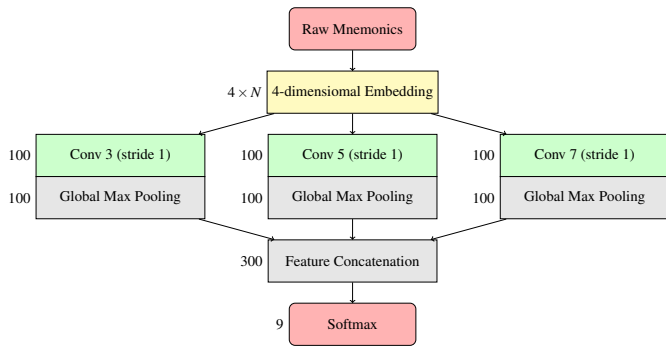


Figure 5: State-of-the-art convolutional neural network architecture for malware classification (Gibert et al., 2017).

the softmax function is applied to the linear combination of the learned features to output the probability distribution over malware families.

4. Evaluation

4.1. The Microsoft Malware Classification Dataset

To validate our method, the data provided by Microsoft for the Big Data Innovators Gathering (Ronen et al., 2018) has been used instead of creating our own in-house (non-reproducible) dataset of benign and malicious samples for reproducibility purposes. Notice that, unlike other domains, there are legal restrictions in place that forbid sharing benign binaries such as copyright laws. In addition, determining whether a file is malicious and its corresponding family is very time-consuming even for experienced security analysts. Furthermore, the Microsoft dataset has become the standard benchmark to evaluate the performance of machine learning algorithms for the task of Windows malware classification because it provides high quality

samples from various malware families. The dataset is publicly accessible and hosted on Kaggle ⁹, and it contains almost half of a terabyte of malware belonging to 9 malware families. See Table 1. For each file, it is provided the hexadecimal repre-

Table 1: Class distribution in the Microsoft dataset (Ronen et al., 2018).

Family Name	# Train Samples	Type
Ramnit	1541	Worm
Lollipop	2478	Adware
Kelihos_ver3	2942	Backdoor
Vundo	475	Trojan
Simda	42	Backdoor
Tracur	751	TrojanDownloader
Kelihos_ver1	398	Backdoor
Obfuscator.ACY	1228	Obfuscated malware
Gatak	1013	Backdoor

sentation of the file’s binary content and the assembly language source code generated with the IDA disassembler tool ¹⁰. The assembly language source code of a computer program is the low-level representation of the program’s statements and machine code instructions. As observed in Table 1 and Figure 6, the dataset is imbalanced and heterogeneous, with the average number of assembly language instructions of samples belonging to different families very distinct. This particularity will greatly affect the performance of the framework as shown in Section 4.4.

Notice that the methodology described through this paper can be applied for the task of malware detection with minor modifications, i.e. reducing the number of output neurons from 9 (number of families in the Microsoft dataset) to 1, indicating the maliciousness of the executable.

4.2. Experimental Setup

The experiments were run on a computer with the following specifications: Intel i7-7700K, 32 GB RAM, 2 x Nvidia GTX 1080Ti. The Microsoft dataset comprises two sets, the training and the test set. But unfortunately, the labels of the test set are not provided. To evaluate the models on the test set, a file with the predicted class probabilities for each sample on the set must be submitted on Kaggle. In consequence, to evaluate our framework, we divided the training set into three sets: training, validation and test set, each containing 70%, 15% and 15% of the samples of the original training set, respectively.

4.3. Parameters Setting

In the experiments, the malware classifier to be evaded is trained on the training set for 15 epochs and evaluated on the validation set. The test set is then used to provide an unbiased evaluation of the fitness of the final model. Figure 7 shows the confusion matrix, also known as the error matrix, which summarizes the results of testing the model on the test set. The

⁹<https://www.kaggle.com/c/malware-classification/overview>

¹⁰<https://www.hex-rays.com/products/ida/>

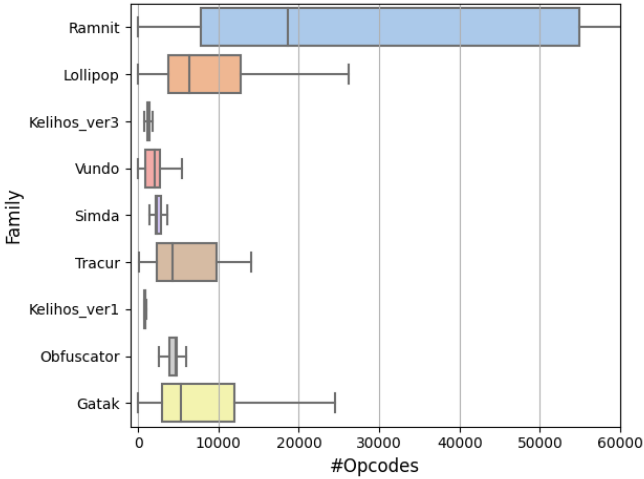


Figure 6: Average number of instructions per family. It can be observed that the size of the samples belonging to different families is not similar. For instance, the samples belonging to the Ramnit and Lollipop families contain 46601.44 and 19157.06 instructions per sample, respectively, while the samples belonging to the Kelihos_ver1 family have only an average of 1326.65 instructions per sample.

accuracy achieved is 98.65%. Regarding the hyperparameters of the classifier, the convolutional layer has 300 filters of size $h \times M$, where $h \in \{3, 5, 7\}$ and M is the embedding size ($M = 4$).

For the reinforcement learning framework parameters setup, the maximum number of modifications per episode is 50 to limit the number of modifications to the minimum. As observed in Figure 9, the average number of dead code insertions needed to degrade the performance of the classifier using a random agent on average is less than 50 for the Gatak, Kelihos_ver1, Tracur, Simda, Vundo and Kelihos_ver3 families. Thus, we considered that 50 insertions were more than enough to test whether or not the AI agent outperforms the random agent. For each family, a different model was trained for a total number of 1000 episodes. In consequence, each model learns which N-grams need to break or mimic in order to mislabel the samples of a given malware family. The discount factor γ is set to 0.99997. The major parameter settings in the training algorithm are shown in Table 2.

4.4. Results

The reinforcement learning models are trained on the samples from the validation set. To evaluate the performance of the learned models in mislabeling the malicious samples of any given family, we recorded the average total reward value, the average number of NOP insertions and the average accuracy per family on the test set. The average total reward plot in Figure 10 shows a 87.62%, 260.95%, 103.09%, 288.05%, 22.88%, 68.75% increase with respect to the cumulative rewards achieved by a random agent on the Kelihos_ver3, Vundo, Simda, Tracur, Kelihos_ver1 and Gatak malware families, respectively. In addition, as it can be observed in Figures 8 and 9, the AI agent was able to misclassify all the samples on the test set belonging

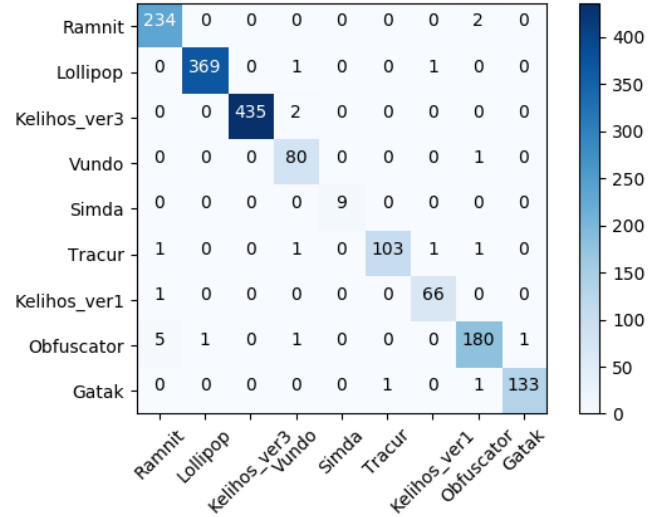


Figure 7: Confusion or error matrix of the CNN classifier on the test set. Each row represents a predicted class and each column represents the instances in an actual class. By definition, a confusion matrix C is such that $C_{i,j}$ is equal to the number of observations known to be in family i and predicted to be in family j . The diagonal represents the case where the prediction of the CNN model is family i and the actual class is i too. Any off-diagonal entry indicates some mistake.

to the Kelihos_ver3, Simda, Kelihos_ver1 and Gatak families while also reducing by 81.81%, 92.62%, 76.47%, and 88.40% the number of NOP instructions inserted, respectively. As evidenced by the experiments (Table 3), our framework is able to generate adversarial examples that bypass a state-of-the-art CNN malware classifier by only inserting NOP instructions. However, the agent has problems mislabeling the samples belonging to the Ramnit, Lollipop, Tracur and Obfuscator.ACY families because of the following two reasons: First, the samples of malware belonging to the Kelihos_ver3, Simda, Kelihos_ver1 and Gatak families in the training set do not contain any NOP instruction. As a result, their respective Q-networks ended up learning the optimal locations to where to insert a NOP instruction in such a way that it makes the malware classifier detect a pattern characteristic of samples in a different malware family. On the contrary, the malware samples belonging to the Ramnit and Lollipop families hardly break the patterns the classifier has learned to detect as their assembly language instructions are at least twice as bigger than those of the other families, and some patterns learned for those families consist of at least one NOP instruction. You can observe in Table 3 that inserting 50 NOP instructions is not enough to misclassify the executables belonging to these families. This could be solved by increasing the maximum number of mutations to be performed or by allowing the agent to insert other dead instructions such as (1) MOV Reg, Reg, (2) PUSH Reg; POP Reg, (3) ADD Reg, 0, to see if with the insertion of those instructions we could make the agent mimic patterns learned for other families.

Nevertheless, in the present paper, we have demonstrated that the use of simple obfuscation techniques is more than enough to decrease the overall accuracy of deep learning models, and we have provided a general framework that uses reinforcement

Table 2: List of major parameters and their values in the training algorithm.

Parameter settings	Value	Description
MAXTURN	50	The maximum number of modifications allowed to be performed by the agent before declaring failure.
EPISODES	1000	The maximum number of episodes played for each family.
Memory buffer max size	2000	The capacity of the replay buffer memory.
Discount factor	0.99997	The discount factor γ used in the Q-learning update.
Initial exploration	1	Initial value of the ϵ in ϵ -greedy exploration
Final exploration	0.5	Final value of the ϵ in ϵ -greedy exploration
Number of filters (Q-Network)	10	Number of filters in the convolutional layer of the Q-Network.

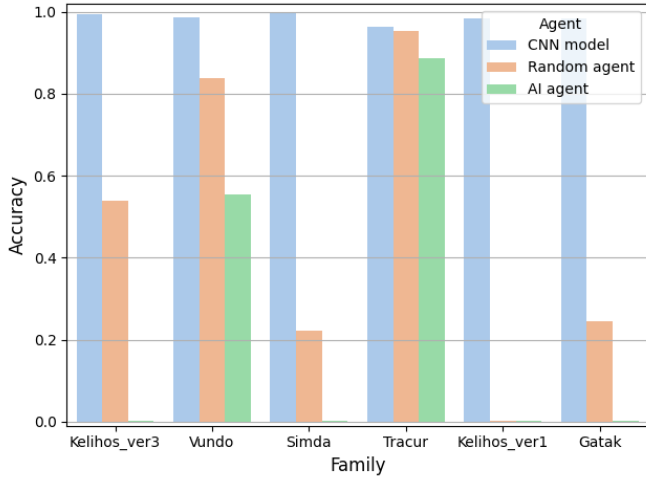


Figure 8: Performance comparison of the CNN model on the test set. It can be seen that the DQN agent outperforms the random agent in almost all families (The lower is the accuracy of the model on the resulting obfuscated test set, the better). Notice that those families in which the agents haven't succeeded in misclassifying any of the samples are not displayed, i.e. Ramnit, Lollipop and Obfuscator.ACY families.

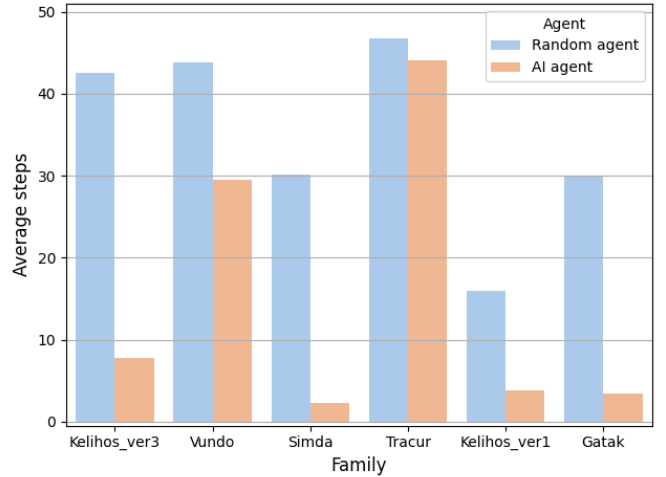


Figure 9: Comparison of the average number of NOP instructions required by the DQN and random agents to misclassify the state-of-the-art CNN classifier (Gibert et al., 2017, 2021). Notice that those samples in which the agents haven't succeeded in misclassifying any of the samples are not displayed, i.e. Ramnit, Lollipop and Obfuscator.ACY. In those cases, the average number of NOP insertions is 50, which is equal to *MAXTURN*, the maximum number of modifications allowed to be performed by the agent before declaring failure.

learning to boost the efficiency of the dead code insertion technique.

5. Conclusions

This paper proposes a general framework using reinforcement learning to make a state-of-the-art malware classifier to incorrectly label the samples of a given family. The core component is an intelligent agent, which constantly interacts with malware samples to learn to choose the optimal locations to where to insert NOP instructions. This has been achieved by training a shallow convolutional Q-network using the double Q-learning algorithm. Experiments show that the proposed framework dropped the classification accuracy of the classifier from 98.65% to 56.53% on the test set while having an evasion rate of 100% for the samples belonging to the Kelihos_ver3, Simda, and Kelihos_ver1 families. In addition, the Q-network enhanced the evasion rate of the dead code insertion technique by 11.29% while also reducing the average number of NOP insertions needed to mislabel the malware sample by 33%.

5.1. Future Work

One future line of research is the use of additional dead code instructions (such as the *MOV Reg, Reg* instruction, the *ADD*

Reg, 0 instruction, the *SUB Reg, 0* instruction, the *SHL Reg, 0* instruction, among others), and the investigation of other common obfuscation techniques such as the instruction substitution technique, the subroutine reordering technique, and the code reordering through jumps technique. Moreover, instead of developing adversarial attacks, one line of research could be the study of potential methods for hardening anti-malware engines against adversarial crafting.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 847402. This research has been partially funded by the Spanish MICINN Projects TIN2015-71799-C2-2-P, ENE2015-64117-C5-1-R, PID2019-111544GB-C22, and supported by the University of Lleida.

References

- Ahmadi, M., Giacinto, G., Ulyanov, D., Semenov, S., Trofimov, M., 2015. Novel feature extraction, selection and fusion for effective malware family classification. CoRR abs/1511.04317. URL <http://arxiv.org/abs/1511.04317>

Table 3: Comparison of the accuracy of the CNN model on the samples of the test set and the obfuscated versions generated by the random agent and the DQN agent, as well as the average cumulative reward and the average number of NOP insertions of both agents.

Malware family	CNN accuracy			Average number of NOP insertions		Average cumulative reward	
	No agent	Random agent	DQN agent	Random agent	DQN agent	Random agent	DQN agent
Ramnit	99.15	99.15	99.15	48.58	48.58	0.01	0.01
Lollipop	99.46	99.46	99.19	48.74	48.612	0.01	0.01
Kelihos_ver3	99.54	53.78	0.0	42.60	7.75	0.85	1.60
Vundo	98.77	83.95	55.56	43.79	29.47	0.22	0.81
Simda	100.0	22.22	0.0	30.11	2.22	0.81	1.64
Tracur	96.26	95.33	88.79	46.84	44.14	0.05	0.21
Kelihos_ver1	98.51	0.0	0.0	15.99	3.76	1.39	1.71
Obfuscator.ACY	95.74	95.21	95.74	46.85	46.95	-0.01	0.01
Gatak	98.52	24.44	0.0	29.83	3.46	0.88	1.48

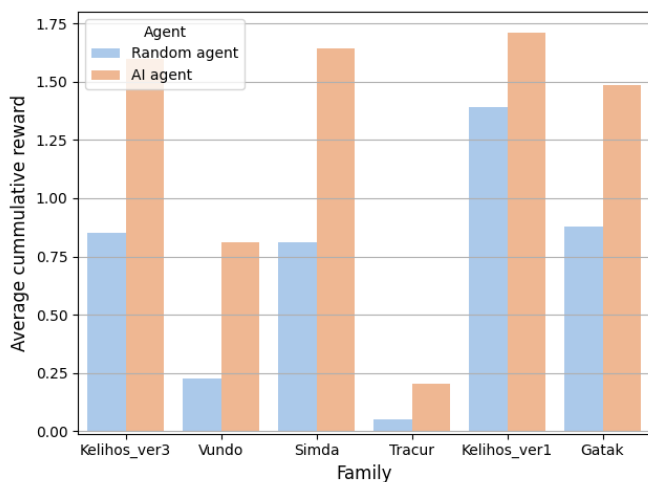


Figure 10: Comparison of the average cumulative reward achieved by the DQN and random agents. Notice that those families in which the agents haven't succeeded in misclassifying any of the samples are not displayed as their average cumulative reward is almost 0, i.e. Ramnit, Lollipop and Obfuscator.ACY families.

Anderson, H. S., Kharkar, A., Filar, B., Evans, D., Roth, P., 2018. Learning to evade static PE machine learning malware models via reinforcement learning. CoRR abs/1801.08917.

URL <http://arxiv.org/abs/1801.08917>

Balakrishnan, A., Schulze, C., 2005. Code obfuscation literature survey.

Bellman, R., 1957. A markovian decision process. Journal of Mathematics and Mechanics 6 (5), 679–684.

URL <http://www.jstor.org/stable/24900506>

Demetrio, L., Biggio, B., Lagorio, G., Roli, F., Armando, A., 2019. Explaining vulnerabilities of deep learning to adversarial malware binaries. CoRR abs/1901.03583.

URL <http://arxiv.org/abs/1901.03583>

Gibert, D., Béjar, J., Mateu, C., Planes, J., Solis, D., Vicens, R., 2017. Convolutional neural networks for classification of malware assembly code. In: Recent Advances in Artificial Intelligence Research and Development - Proceedings of the 20th International Conference of the Catalan Association for Artificial Intelligence, Deltebre, Terres de l'Ebre, Spain, October 25-27, 2017. pp. 221–226.

URL <https://doi.org/10.3233/978-1-61499-806-8-221>

Gibert, D., Mateu, C., Planes, J., 2020a. Hydra: A multimodal deep learning framework for malware classification. Computers & Security 95, 101873.

URL <https://www.sciencedirect.com/science/article/pii/S0167404820301462>

Gibert, D., Mateu, C., Planes, J., 2020b. The rise of machine learning for

detection and classification of malware: Research developments, trends and challenges. Journal of Network and Computer Applications, 102526.

URL <http://www.sciencedirect.com/science/article/pii/S1084804519303868>

Gibert, D., Mateu, C., Planes, J., Marques-Silva, J., 2021. Auditing static machine learning anti-malware tools against metamorphic attacks. Computers & Security 102, 102159.

URL <http://www.sciencedirect.com/science/article/pii/S0167404820304326>

Gibert, D., Mateu, C., Planes, J., Vicens, R., 2018. Classification of malware by using structural entropy on convolutional neural networks.

URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16133>

Gibert, D., Mateu, C., Planes, J., Vicens, R., 2019. Using convolutional neural networks for classification of malware represented as images. J. Comput. Virol. Hacking Tech. 15 (1), 15–28.

URL <https://doi.org/10.1007/s11416-018-0323-0>

Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P. D., 2016. Adversarial perturbations against deep neural networks for malware classification. CoRR abs/1606.04435.

URL <http://arxiv.org/abs/1606.04435>

Hasselt, H. v., Guez, A., Silver, D., 2016. Deep reinforcement learning with double q-learning. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI'16. AAAI Press, p. 2094–2100.

Kaelbling, L. P., Littman, M. L., Moore, A. W., May 1996. Reinforcement learning: A survey. J. Artif. Int. Res. 4 (1), 237–285.

Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C., Roli, F., Sep. 2018. Adversarial malware binaries: Evading deep learning for malware detection in executables. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 533–537.

Krčál, M., Švec, O., Bálek, M., Jašek, O., 2018. Deep convolutional malware classifiers can learn from raw executables and labels only.

URL <https://openreview.net/forum?id=HkHrmM1PM>

Lin, L.-J., May 1992. Self-improving reactive agents based on reinforcement learning, planning and teaching. Machine Learning 8 (3), 293–321.

URL <https://doi.org/10.1007/BF00992699>

McLaughlin, N., Martinez del Rincon, J., Kang, B., Yerima, S., Miller, P., Sezer, S., Safaei, Y., Trichel, E., Zhao, Z., Doupé, A., et al., 2017. Deep android malware detection. In: Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy. CODASPY '17. Association for Computing Machinery, New York, NY, USA, p. 301–308.

URL <https://doi.org/10.1145/3029806.3029823>

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. A., 2013. Playing atari with deep reinforcement learning. CoRR abs/1312.5602.

URL <http://arxiv.org/abs/1312.5602>

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nat. 518 (7540), 529–533.

URL <https://doi.org/10.1038/nature14236>

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., Swami, A., 2016. The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS P). pp. 372–387.

Qiu, J., Zhang, J., Luo, W., Pan, L., Nepal, S., Xiang, Y., Dec. 2020. A survey of android malware detection with deep neural models. *ACM Comput. Surv.* 53 (6).
URL <https://doi.org/10.1145/3417978>

Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., Nicholas, C. K., 2018. Malware detection by eating a whole EXE. In: The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018. pp. 268–276.
URL <https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16422>

Ren, K., Zheng, T., Qin, Z., Liu, X., 2020. Adversarial attacks and defenses in deep learning. *Engineering* 6 (3), 346 – 360.
URL <http://www.sciencedirect.com/science/article/pii/S209580991930503X>

Ronen, R., Radu, M., Feuerstein, C., Yom-Tov, E., Ahmadi, M., 2018. Microsoft malware classification challenge. *CoRR* abs/1802.10135.
URL <http://arxiv.org/abs/1802.10135>

Souri, A., Hosseini, R., Jan 2018. A state-of-the-art survey of malware detection approaches using data mining techniques. *Human-centric Computing and Information Sciences* 8 (1), 3.
URL <https://doi.org/10.1186/s13673-018-0125-x>

Suciu, O., Coull, S. E., Johns, J., 2019. Exploring adversarial examples in malware detection. In: 2019 IEEE Security and Privacy Workshops (SPW). pp. 8–14.

Venkatraman, S., Alazab, M., Vinayakumar, R., 2019. A hybrid deep learning image-based analysis for effective malware detection. *Journal of Information Security and Applications* 47, 377–389.
URL <https://www.sciencedirect.com/science/article/pii/S2214212618304563>

Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Venkatraman, S., 2019. Robust intelligent malware detection using deep learning. *IEEE Access* 7, 46717–46738.

Watkins, C. J. C. H., Dayan, P., May 1992. Q-learning. *Machine Learning* 8 (3), 279–292.
URL <https://doi.org/10.1007/BF00992698>

Zhang, Y., Huang, Q., Ma, X., Yang, Z., Jiang, J., 2016. Using multi-features

and ensemble learning method for imbalanced malware classification. In: 2016 IEEE Trustcom/BigDataSE/ISPA. pp. 965–973.

Appendix A. Configuration Details of the Neural Networks

Layer (type)	Output shape	Parameters #
input_1 (Input layer)	(None, N, 1)	0
embedding_1 (Embedding layer)	(None, N, E)	4*557 (E*V)
conv1d_3(Conv_1D)	(None, N, 10)	10*3*4
global_maxpool1d_3 (GlobalMaxPooling1D)	(None, 10)	0
time_distributed (dense_1)	(None, N)	10*1
softmax_1 (Softmax)	(None, N)	0
Total trainable parameters		2358

Table A.4: Configuration details of the Q-network architecture used to select the locations to which insert NOP instructions in a given malware sample in order to bypass detection. N is the size of the input mnemonics sequence. E is the embedding size. V is the vocabulary size, i.e. the number of different mnemonics found in the Microsoft dataset.

Layer (type)	Output shape	Parameters #
input_1 (Input layer)	(None, N, 1)	0
embedding_1 (Embedding layer)	(None, N, E)	4*557 (E*V)
conv1d_3(Conv_1D)	(None, N, 100)	100*3*4
global_maxpool1d_3 (GlobalMaxPooling1D)	(None, 100)	0
conv1d_5(Conv_1D)	(None, N, 100)	100*5*4
global_maxpool1d_5 (GlobalMaxPooling1D)	(None, 100)	0
conv1d_7(Conv_1D)	(None, N, 100)	100*7*4
global_maxpool1d_7 (GlobalMaxPooling1D)	(None, 100)	0
features concatenation	(None, 300)	0
dense_1 (Dense)	(None, 9)	300*9
softmax_1 (Softmax)	(None, 9)	0
Total trainable parameters		10928

Table A.5: Configuration details of the shallow CNN architecture for malware classification.