2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) | 978-1-6654-0424-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/PERCOMWORKSHOPS51409.2021.9430980

Towards Robust Android Malware Detection Models using Adversarial Learning

Hemant Rathore
BITS Pilani, India
hemantr@goa.bits-pilani.ac.in
supervised by: Prof. Sanjay K. Sahay

***Abstract*—Malware analysis and detection is an endless competitive battle between malware designers and the anti-malware community. Recently researchers have proposed state-of-the-art malware detection models built using machine learning and deep learning which are necessary to detect advanced metamorphic malware. But these malware detection models are susceptible to adversarial attacks. Therefore we propose to design robust Android malware detection models against adversarial attacks using reinforcement learning. We propose the Single Policy and Multi Policy based Evasion Attacks for Perfect Knowledge and Limited Knowledge scenario respectively against many malware detection models built using four different sets of classifiers (bagging, boosting and deep neural network). The motivation is to identify the adversarial vulnerability in Android malware detection models and then propose defence against them.**

***Index Terms*—Android, Adversarial Learning, Machine Learning, Malware Detection, Reinforcement Learning, Smartphone**

I. Introduction

The Android smartphones have grown exponentially in the last decade, which has gained massive attention from malware designers to develop malicious applications which threaten the Android ecosystem. According to GDATA Mobile Malware Report, $11,500$ new malicious applications were detected each day in 2018 with a total number of malware for Android devices reaching $18,792,234$ in 2019. The primary defences against these malware attacks are designed and developed by the anti-malware research community and antivirus industry. Traditionally malware detection systems are based on Signature, Heuristic and Behaviour-based mechanism [1]. These mechanisms are highly human-driven, time-consuming, non-scalable and are unable to detect advanced metamorphic malware. Thus researchers have started investigating Android malware detection models based on Machine Learning (ML) and Deep Learning (DL) which has shown promising results [1] [2] [3]. Development of these detection models is a two-step process (1) Feature Extraction (2) Classification. Extraction of features can be done using static/dynamic analysis of Android applications followed by the use of classification algorithm(s) to construct malware detection models.

The recent state-of-the-art research shows that classification models constructed using ML/DL are prone to adversarial attacks. Goodfellow et al. demonstrated that small intentional worst-case permutation could be used to generate adversarial samples which can produce wrong classification results with high confidence [4]. Kurakin et al. showed that ML models are susceptible to adversarial attacks in physical world deployments as well [5]. Similar adversarial attacks can be devised on Android malware detection models which can ultimately jeopardize their real-world deployment. These attacks are developed based on *Attacker's Goal* to cause integrity, availability and privacy violation of detection models [6]. The attacker can design these attacks based on the different level of knowledge about the target system, namely training dataset, feature information and the classification algorithm. *Perfect Knowledge* scenario assumes that the attacker has complete information, whereas the *Zero-Knowledge* scenario assumes the attacker has no information about the detection system. Also, *Limited Knowledge* scenario assumes that the attacker has partial information about the system. Further, the attacker based on *Goal* and *Knowledge* can perform *Evasion Attack*, *Poisoning Attack* and *Privacy Attack*. The attacker in *Evasion Attack* tries to modify the training samples to force misclassification in detection models. *Poisoning Attack* is designed to intentionally injects wrongly labelled sample into the training set to decrease the capabilities of detection model. Finally, *Privacy Attack* aims to steal/gather information about the detection model.

II. Problem Overview and Proposed Architecture

In the proposed work, we will first act as an adversary and develop adversarial attack(s) on Android malware detection models. After a thorough investigation of vulnerabilities, we will propose the adversarial defence mechanism(s) to build the robust Android malware detection model(s).

A. Problem Definition

Consider a dataset (D) containing Android application (malware and benign) along with corresponding class labels. The dataset (D) contains a subset of malicious applications (M) and another subset of benign applications (B) where ($|B| \approx |M|$). Android applications can be decompiled to generate feature vector which is represented as $D = \{(x_i, y_i)\}$ where x_i is the vector representing the application and y_i is the corresponding class label. Various classification algorithms A can be used to construct different Android malware detection model(s), and they can be evaluated and optimized based on performance metric like accuracy, AUC, FPR etc.

The goal of the adversary is to design an *Evasion Attack* which aims to modify malicious applications such that mal-

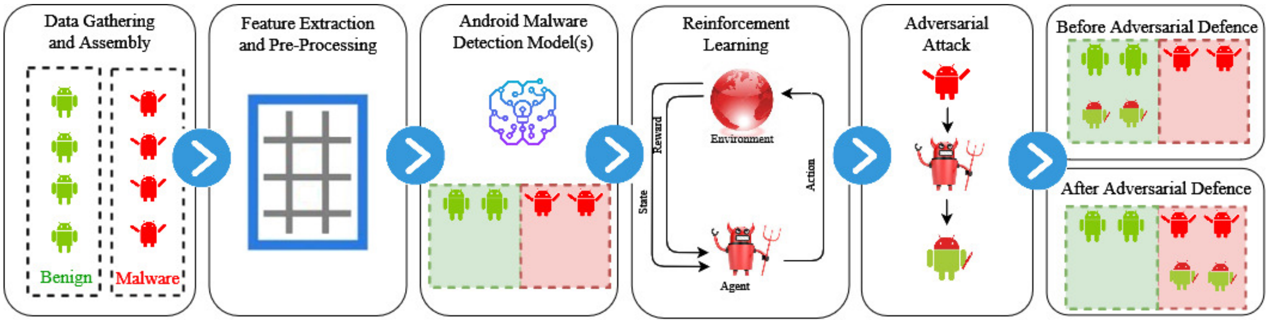
Data Gathering and Assembly
Benign
Malware
Feature Extraction and Pre-Processing
Android Malware Detection Model(s)
Reinforcement Learning
Environment
Reward
State
Action
Agent
Adversarial Attack
Before Adversarial Defence
After Adversarial Defence

Fig. 1. An overview of the proposed architecture of designing robust Android malware detection model(s)

.

ware detection model(s) misclassifies them. The adversary targets to convert maximum malicious applications from M with minimum modifications in each application to reduce the cost of attack. Also, all the modification(s) should be syntactically possible and shall not break any functional or behavioural property of the modified application. Various vulnerabilities in malware detection models could be identified after successful evasion attack(s) which are proposed to be countered using adversarial defence mechanisms like retraining, distillation, etc.

B. Architecture Overview and Research Progress

Figure 1 illustrates a six-step modular approach to develop robust Android malware detection model(s). The first step consists of data gathering of malicious and benign applications. We have used Drebin dataset and Android Malware Dataset (AMD) containing malicious applications while benign applications are downloaded from Google Play store and verified by VirusTotal [1]. The second step consists of feature extraction from Android applications and then pre-processing them using feature engineering techniques. We did the static analysis of Android applications and extracted permission, intent and API calls. The third step proposes to build malware detection models where we have constructed many malware detection models using four different set of classifiers like classical ML, bagging, boosting and deep neural network. The fourth module train's agent(s) for evasion attack using reinforcement learning techniques. We have explored q-learning, deep q-learning, actor-critic etc. methods to devise the adversarial attack policy against malware detection models. The fifth module performs actual evasion attack by modifying malicious samples such that they force misclassification in above detection model(s). The last step uses adversarial defence to counter the evasion attack and propose robust Android malware detection model(s). With the above research plan, we have achieved and further anticipate the following thesis contributions:

- We proposed a *Single Policy based Evasion Attack* for *Perfect Knowledge* scenario based on reinforcement learning. The adversarial attack was designed using deep q-learning and was applied against a group of twelve different malware detection models build using four different sets of classifiers, including classical ML, bagging, boosting and deep neural network. Currently, we have achieved an average fooling rate of 90% across twelve different detection models with a maximum of 10% modification. The attack policy is highly interpretable and can list the most vulnerable features, thus reducing the overall cost of the adversarial attack.
- We also proposed a *Multi Policy based Evasion Attack* for *Limited Knowledge* scenario where the adversary is assumed to have knowledge about the training dataset and features, but no information about classification algorithm used to build detection model. We tested this attack policy against the same set of twelve detection models and achieved an average fooling rate of 95% with a maximum 10% modification.
- We are in the process of developing another adversarial attack for *Zero-Knowledge* scenario where the adversary is assumed to have no information about the dataset, feature information or classification algorithm.
- The adversarial defence mechanism should be able to defend against evasion attacks on malware detection models. Thus we started with a basic strategy of model retraining and were able to reduce the average fooling rate threefold against Single Policy based evasion attack. We are also designing defensive distillation and GAN based mechanism to defend against adversarial attacks.

References

[1] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A survey on malware detection using data mining techniques," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–40, 2017.

[2] K. Tam, A. Feizollah, N. B. Anuar, R. Salleh, and L. Cavallaro, "The evolution of android malware and android analysis techniques," *ACM Computing Surveys (CSUR)*, vol. 49, no. 4, 2017.

[3] M. Sewak, S. K. Sahay, and H. Rathore, "Doom: a novel adversarial-drl-based op-code level metamorphic malware obfuscator for the enhancement of IDS," in *International Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2020.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations (ICLR)*, 2014.

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *International Conference on Learning Representations (ICLR)*, 2016.

[6] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning (ICML)*, 2017.