Poster: Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning

Hemant Rathore*, Sanjay K. Sahay†, Piyush Nikam‡
Dept. of CS & IS, Goa Campus, BITS Pilani, India
{*hemantr, †ssahay, ‡h20180057}@goa.bits-pilani.ac.in
Mohit Sewak§
Security & Compliance Research, Microsoft, India
mohit.sewak@microsoft.com

Abstract—Since the inception of Andoroid OS, smartphones sales have been growing exponentially, and today it enjoys the monopoly in the smartphone marketplace. The widespread adoption of Android smartphones has drawn the attention of malware designers, which threatens the Android ecosystem. The current state-of-the-art Android malware detection systems are based on machine learning and deep learning models. Despite having superior performance, these models are susceptible to adversarial attack. Therefore in this paper, we developed eight Android malware detection models based on machine learning and deep neural network and investigated their robustness against the adversarial attacks. For the purpose, we created new variants of malware using reinforcement learning, which will be misclassified as benign by the existing Android malware detection models. We propose two novel attack strategies, namely single policy attack and multiple policy attack using reinforcement learning for white-box and grey-box scenario respectively. Putting ourselves in adversary' shoes, we designed adversarial attacks on the detection models with the goal of maximising fooling rate, while making minimum modifications to the Android application and ensuring that the app's functionality and behaviour does not change. We achieved an average fooling rate of 44.21% and 53.20% across all the eight detection models with maximum five modifications using a single policy attack and multiple policy attack, respectively. The highest fooling rate of 86.09% with five changes was attained against the decision tree based model using the multiple policy approach. Finally, we propose an adversarial defence strategy which reduces the average fooling rate by threefold to 15.22% against a single policy attack, thereby increasing the robustness of the detection models i.e. the proposed model can effectively detect variants (metamorphic) of malware. The experimental analysis shows that our proposed Android malware detection system using reinforcement learning is more robust against adversarial attacks.

Bibliographic Reference

Rathore, Hemant, et al. "Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning." Information Systems Frontiers (2020) pp. 1-16
DOI:https://doi.org/10.1007/s10796-020-10083-8

Poster: Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning

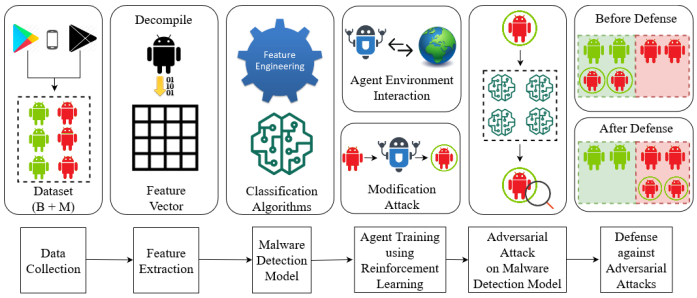
Hemant Rathore[1], Sanjay K. Sahay[1], Piyush Nikam[1], Mohit Sewak[2]

[1]Department of CS & IS, Goa Campus, BITS Pilani, India

[2]Security & Compliance Research, Microsoft, India

Problem Overview and Proposed Architecture

- Literature suggests malware detection systems based on ML/DL models are state-of-the-art and have shown promising results
- Despite having superior performance, these models are susceptible to adversarial attacks
- We investigated the robustness of malware detection models against the adversarial attacks
- We designed Single Policy Attack (SPA) and Multiple Policy Attack (MPA) against detection models using reinforcement learning for white-box and grey-box scenario, respectively
- We proposed adversarial retraining as the defense against adversarial attacks and thereby increased the robustness of malware detection models



Adversarial Attack Agent

- The proposed adversarial attack agent crafts perturbations governed by policy extracted from the Q-table(s)
- The policy is designed to modify malicious samples such that malware detection models are forced to misclassify them
- Goal of the optimal policy is to modify the maximum number of malicious samples with minimum modifications in each sample to generate new malicious variants that are misclassified by detection models
- Optimal policy ensures that each modification is syntactically possible and does not disrupt any functional or behavioral aspect of the sample/application
- SPA uses a single policy extracted from a Q-table for performing adversarial attacks in white-box scenario on malware detection models
- MPA uses a set of optimal policies extracted from many Q-tables and use them parallelly for adversarial attacks in black-box scenario on malware detection models

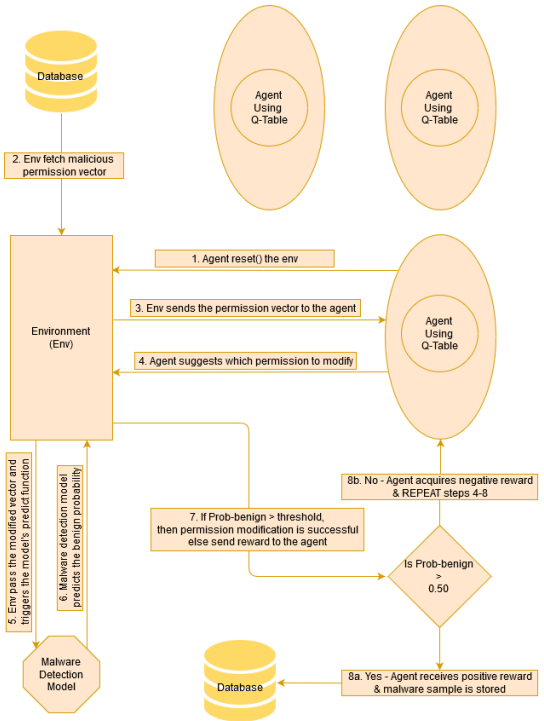
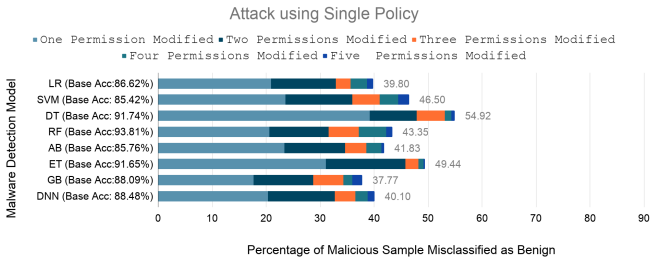


Fig: SPA/MPA based Adversarial Attack on Malware Detection Models

Experimental Results and Conclusion

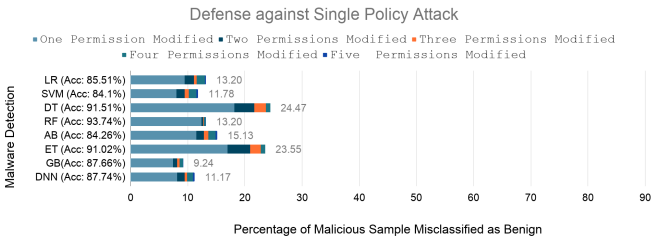
*Fooling rate after the **Adversarial Attack** on Malware Detection Models*

- Eight different malware detection models were constructed using a variety of classification techniques like traditional algorithms (LR, SVM, and DT), bagging algorithms (RF, ET), boosting algorithms (AB, GB) and DNN
- Highest malware detection accuracy was achieved using RF (93.81%) and all the other detection models have attained more than 85% accuracy. However, these models are susceptible to adversarial attacks
- SPA for white-box scenario achieved an average fooling rate of 44.21% across eight detection models with maximum five modifications
- MPA for black-box scenario achieved an average fooling rate of 53.20% across eight detection models with maximum five modifications



*Fooling rate after the **Adversarial Defense** on Malware Detection Models*

- Adversarial defense strategy reduced the average fooling rate against the single policy attack by threefold to 15.22% and twofold for the multi-policy attack to 29.44%, i.e., it can now effectively detect variants (metamorphic) of malware



Bibliographic Reference

Rathore, Hemant, et al. *"Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning."* Information Systems Frontiers (2020) pp. 1-16 DOI:https://doi.org/10.1007/s10796-020-10083-8