Poster: Defending malware detection models against evasion based adversarial attacks

Hemant Rathore*, Animesh Sasan†, Sanjay K. Sahay‡
Dept. of CS & IS, Goa Campus, BITS Pilani, India
{*hemantr, †f20170036, ‡ssahay}@goa.bits-pilani.ac.in

Mohit Sewak§
Security & Compliance Research, Microsoft, India
mohit.sewak@microsoft.com

Abstract

The last decade has witnessed a massive malware boom in the Android ecosystem. Literature suggests that artificial intelligence/machine learning based malware detection models can potentially solve this problem. But, these detection models are often vulnerable to adversarial samples developed by malware designers. Therefore, we validate the adversarial robustness and evasion resistance of different malware detection models developed using machine learning in this work. We first designed a neural network agent (*MalDQN*) based on deep reinforcement learning that adds noise via perturbations to the malware applications and converts them into adversarial malware applications. Malware designers can also generate these samples and use them to perform evasion attacks and fool the malware detection models. The proposed MalDQN agent achieved an average 98% fooling rate against twenty distinct malware detection models based on a variety of classification algorithms (standard, ensemble, and deep neural network) and two different features (android permission and intent). The MalDQN evasion attack reduced the average accuracy from 86.18% to 55.85% in the twenty malware detection models mentioned above. Later, we also developed defensive measures to counter such evasion attacks. Our experimental results show that the proposed defensive strategies considerably improve the capability of different malware detection models to detect adversarial applications and build resistance against them.

BIBLIOGRAPHIC REFERENCE

Rathore, Hemant, et al. *"Defending Malware Detection Models against Evasion based Adversarial Attacks."* Pattern Recognition Letters (2022) DOI: https://doi.org/10.1016/j.patrec.2022.10.010

Poster: Defending malware detection models against evasion based adversarial attacks

Hemant Rathore[1], Animesh Sasan[1], Sanjay K. Sahay[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

- Last decade witnessed a massive boom in smartphone malware in the android ecosystem.
- Existing literature suggests that AI (ML & DL) based malware detection systems can potentially solve malware detection problems.
- These next-gen malware detection models have shown superior performance but might be susceptible to adversarial attacks.
- We validate the adversarial robustness and evasion resistance of these ML & DL based malware detection models.
- We designed a neural network agent (*MalDQN*) based on deep reinforcement learning that adds noise via perturbations to the malware applications and converts them into adversarial malware applications.
- We also propose defensive strategies that considerably improve the capability of ML & DL based malware detection models to detect malicious adversarial applications and build resistance against them.

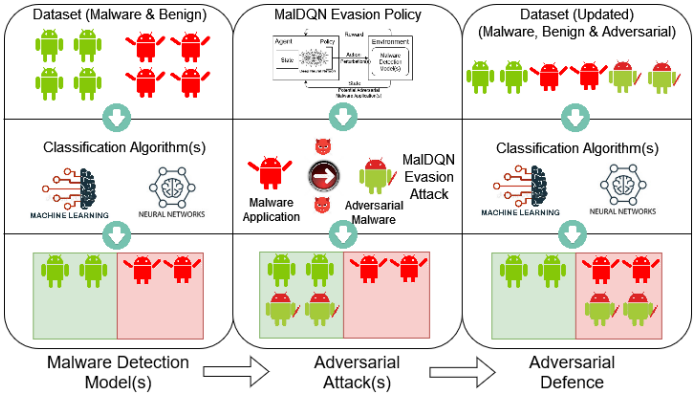

Fig: Framework to construct adversarially superior malware detection models

MalDQN Adversarial Attack

- We first constructed twenty distinct malware detection models using two features (permission/intent) and ten different classification algorithms from three categories (standard, ensemble, & neural network based algorithms).
- The twenty malware detection models achieved an average accuracy and AUC of 86.18% and 0.89, respectively. The highest accuracy (94.80%) and AUC (0.98) were achieved by the random forest model based on android permission.
- We then posed like an adversary and designed the MalDQN agent based on deep reinforcement learning to perform evasion attacks on the above malware detection models.
- The proposed MalDQN is a fully connected deep neural network with three hidden layers (128, 256, and 128 neurons). It uses *ELU* activation function in the hidden layers and *Linear* activation function in the output layer. It also used *Binary Cross Entropy Loss* function during training and *Adam* as an optimization algorithm.
- The MalDQN agent is trained to suggest minimum perturbation(s) while ensuring the structural, syntactical, and functional integrity of the modified malware applications.
- The MalDQN agent performs integrity violations in malware applications by adding perturbations and converting them into adversarial applications, forcing misclassifications in detection models and fooling the security ecosystem.
- Malware designers can also generate similar malicious samples and use them to actually fool real-world android malware detection models.
- The MalDQN attack achieved an average fooling rate of 98% against twenty malware detection models, along with 100% fooling against many models.
- We thoroughly investigated the adversarial vulnerabilities exposed by the evasion attack and developed a defense strategy to counter it.

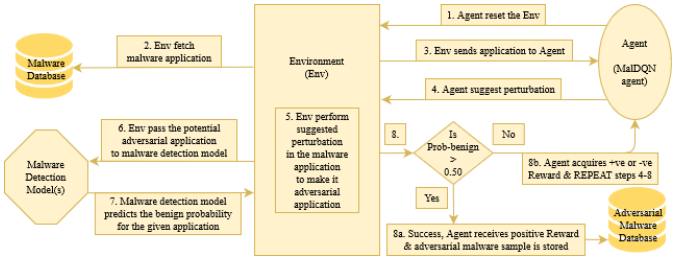

Fig: Flow diagram of the MalDQN attack against malware detection models

Experimental Results and Conclusion

- The MalDQN evasion attack reduced the average accuracy from 86.18% to 55.85% and average AUC from 0.89 to 0.48 in twenty malware detection models.
- The adversarial retraining defense mechanism improved the average accuracy and AUC to 96.36% and 0.88 of twenty detection models.
- The adversarial defense also improved the adversarial robustness and evasion resistance of the detection models by decreasing the forced misclassification of adversarial samples.

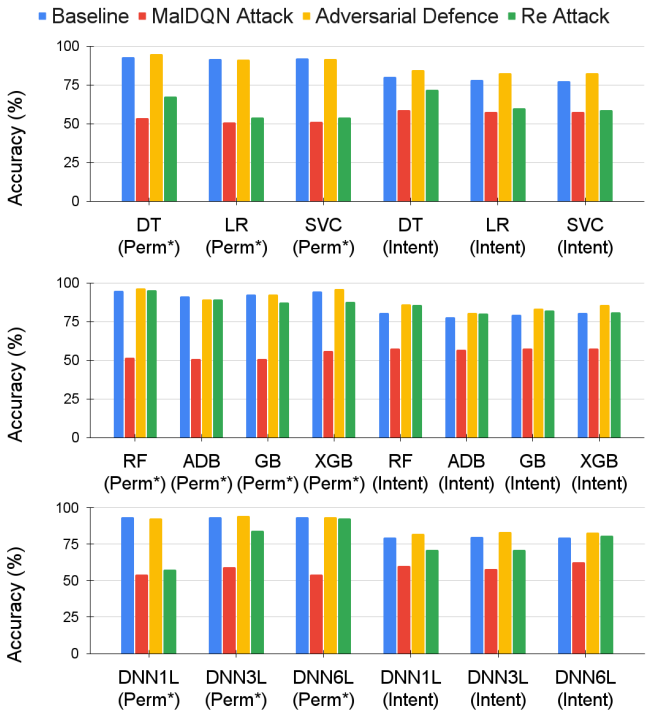

Fig: Performance (Baseline, MalDQN Attack, Adversarial Retraining, and MalDQN Reattack) of different android malware detection models

Bibliographic Reference

Rathore, Hemant, et al. *"Defending Malware Detection Models against Evasion based Adversarial Attacks."* Pattern Recognition Letters (2022) DOI: https://doi.org/10.1016/j.patrec.2022.10.010