Poster: Robust Malware Detection Models: Learning from Adversarial Attacks and Defenses

Hemant Rathore*, Adithya Samavedhi†, Sanjay K. Sahay‡
Dept. of CS & IS, Goa Campus, BITS Pilani, India
{*hemantr, †f20170071, ‡ssahay}@goa.bits-pilani.ac.in
Mohit Sewak§
Security & Compliance Research, Microsoft, India
mohit.sewak@microsoft.com

Abstract—The last decade witnessed an exponential growth of smartphones and their users, which has drawn massive attention from malware designers. The current malware detection engines are unable to cope with the volume, velocity, and variety of incoming malware. Thus the anti-malware community is investigating the use of machine learning and deep learning to develop malware detection models. However, research in other domains suggests that the machine learning/deep learning models are vulnerable to adversarial attacks. Therefore in this work, we proposed a framework to construct robust malware detection models against adversarial attacks. We first constructed twelve different malware detection models using a variety of classification algorithms. Then we acted as an adversary and proposed Gradient-based Adversarial Attack Network to perform adversarial attacks on the above detection models. The attack is designed to convert the maximum number of malware samples into adversarial samples with minimal modifications in each sample. The proposed attack achieves an average fooling rate of 98.68% against twelve permission-based malware detection models and 90.71% against twelve intent-based malware detection models. We also identified the list of vulnerable permissions/intents which an adversary can use to force misclassifications in detection models. Later we proposed three adversarial defense strategies to counter the attacks performed on detection models. The proposed Hybrid Distillation based defense strategy improved the average accuracy by 54.21% for twelve permission-based detection models and 59.14% for intent-based detection models. We also concluded that the adversarial-based study improves the performance and robustness of malware detection models and is essential before any real-world deployment.

Bibliographic Reference

Rathore, H., Samavedhi, A., Sahay, S.K., Sewak, M., "Robust Malware Detection Models: Learning from Adversarial Attacks and Defenses." Forensic Science International: Digital Investigation (2021), vol 37, p. 301183 https://doi.org/10.1016/j.fsidi.2021.301183

Poster: Robust Malware Detection Models: Learning from Adversarial Attacks and Defenses

Hemant Rathore[1], Adithya Samavedhi[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

- Literature suggests that malware detection systems based on ML / DL models are currently state-of-the-art and are showing promising results
- Despite having superior performance, these detection models are susceptible to adversarial attacks
- We investigated the robustness of android malware detection models against the adversarial attacks
- We proposed Gradient Adversarial Attack Network (GAAN) which performs evasion attack(s) against permission / intent based malware detection models built using different machine / deep learning algorithms
- We also proposed three different defense strategies against adversarial attacks and thereby increased the robustness of malware detection models

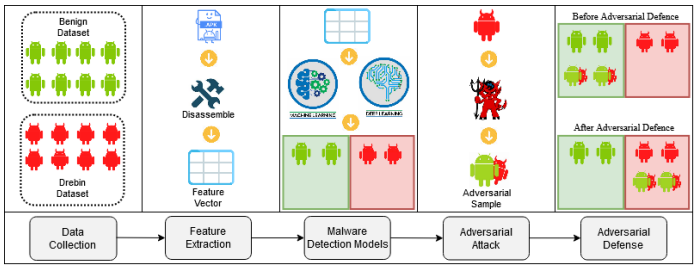

Fig: Proposed framework for constructing robust malware detection model(s)

Adversarial Attacks and Defenses

- The proposed adversarial attack agent crafts perturbations governed by policy extracted from the GAAN
- The policy is designed to perform evasion attacks by modifying malicious samples such that they are forcefully misclassified as benign by malware detection models
- 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 forcefully misclassified by detection models
- Optimal policy ensures that each modification is syntactically possible and does not disrupt any functional or behavioral aspect of the application(s)
- GAAN is designed for the grey-box scenario where an adversary is assumed to have knowledge about dataset and features but no information about malware detection models or their architecture

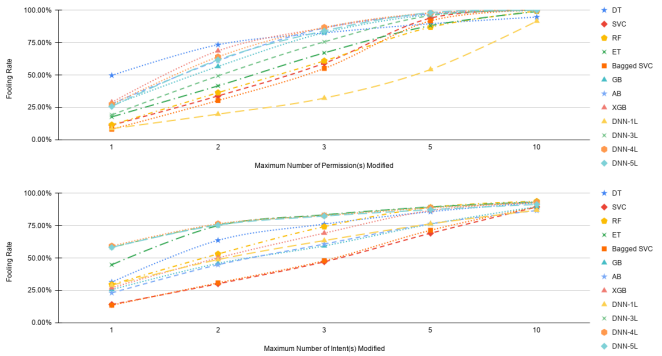

Fig: Performance of permission / intent based malware detection models against GAAN attack concerning fooling rate

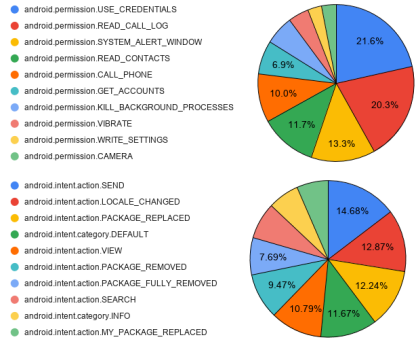

Fig: Distribution of 10 most frequently modified permissions / intent during 10-bit GAAN attack on twelve detection models

Experimental Results and Conclusion

- We proposed the GAAN strategy to perform evasion attacks against twelve distinct malware detection models built using a variety of classification algorithms (ML, bagging, boosting, DNN)
- The twelve different permission based malware detection models achieve an average accuracy of 93.35% whereas twelve intent-based detection models attain an average accuracy of 80.17%
- The evasion attack with a maximum of 10 modifications achieved an average fooling rate of 98.68% against twelve permission-based malware detection models and 90.71% against intent based detection models
- We also developed vulnerability lists of permissions / intents, which adversaries can use to force misclassifications against malware detection models
- We designed three adversarial defense strategies (Adversarial Retraining, GAN Retraining, and Hybrid Distillation) for malware detection models to counter evasion attacks. The proposed hybrid distillation defense strategy achieved an average accuracy improvement of 54.21% for twelve permission-based malware detection models and 59.14% for intent-based detection models
- The highest accuracy was achieved by the Random Forest model (95.80%) followed by the Extra Tree model (95.16%) with adversarial retraining based defense strategy. Overall the hybrid distillation performed best, followed by adversarial retraining and GAN based defense
- We conclude that the adversarial defense does improve the robustness of malware detection models and should be validated before any real-world deployment of any malware detection models

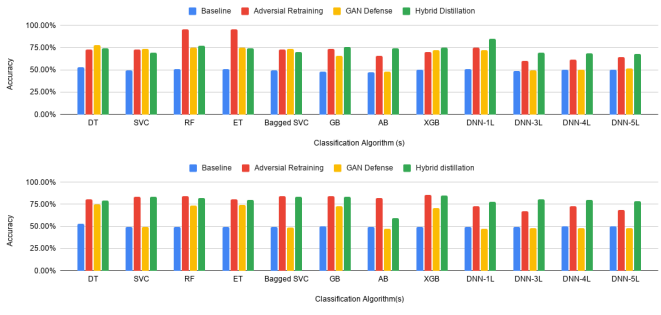

Fig: Performance of permission / intent based malware detection models (baseline and defense strategies) against GAAN attack

Bibliographic Reference

Rathore, H., Samavedhi, A., Sahay, S.K., Sewak, M., *"Robust Malware Detection Models: Learning from Adversarial Attacks and Defenses."* Forensic Science International: Digital Investigation (2021), vol 37, p. 301183 https://doi.org/10.1016/j.fsidi.2021.301183