Poster: Examining Adversarial Learning against Graph-based IoT Malware Detection Systems

Ahmed Abusnaina, Aminollah Khormali, Hisham Alasmary, Jeman Park,
Afsah Anwar, Ulku Meteriz, and Aziz Mohaisen
Departemet of Computer Science, University of Central Florida
{ahmed.abusnaina, aminkhormali, hisham, parkjeman, afsahanwar, meteriz}@knights.ucf.edu, mohaisen@cs.ucf.edu

***Abstract*—The main goal of this study is to investigate the robustness of graph-based Deep Learning (DL) models used for Internet of Things (IoT) malware classification against Adversarial Learning (AL). We designed two approaches to craft adversarial IoT software, including Off-the-Shelf Adversarial Attack (OSAA) methods, using six different AL attack approaches, and Graph Embedding and Augmentation (GEA). The GEA approach aims to preserve the functionality and practicality of the generated adversarial sample through a careful embedding of a benign sample to a malicious one. Our evaluations demonstrate that OSAAs are able to achieve a misclassification rate (MR) of 100%. Moreover, we observed that the GEA approach is able to misclassify all IoT malware samples as benign.**

***Index Terms*—Adversarial Learning, Deep Learning, Graph Analysis, Internet of Things, Malware Detection**

I. Introduction

Internet of Things (IoT) devices, including sensors, voice assistants, automation tools, *etc.* [1], are widely used, increasing the attack surface of the Internet due to their evolving and often insecure software. Thus, it is essential to understand IoT software to address security issues through analysis and detection [1]. However, the research work on IoT software analysis has been very limited not only in the size of the analyzed samples, but also the utilized approaches [2]. A promising direction leverages a graph-theoretic approach to analyze IoT malware. Representative static characteristics of IoT applications can be extracted from the Control Flow Graph (CFG), which can be utilized to build an automatic IoT malware detection system [3].

Machine Learning (ML) algorithms, specifically DL networks, are actively used in a wide range of applications, such as health-care, industry, cyber-security, and *etc.* [4], [5]. However, it has been shown that ML/DL networks are vulnerable to AL, where an adversary can force the model to his desired output, *e.g.,* misclassification. Although it is an active research area, there is very little research work done on understanding the impact of AL on DL-based IoT malware detection system and practical implications [6], particularly those that utilize CFG features for detection.

Goal of this study. Motivated by the aforementioned issues, our main goal is generating *adversarial IoT software samples that (1) fool the classifier and (2) function as intended.*

Approach. To tackle the above objectives, we designed two approaches to craft adversarial examples, including OSAA and GEA approaches. The OSAA approach incorporates six well-known adversarial learning methods to force the model to misclassification. Whereas, the GEA approach aims to preserve the functionality and practicality of the generated adversarial samples through a careful connection of benign graph to a malicious one.

Contributions. Our contributions are as follows: 1) We examined the robustness of CFG-based deep learning IoT malware detection system using two different approaches, including off-the-shelf adversarial learning algorithms and graph embedding and augmentation, while maintaining the practicality and functionality of the crafted AEs. 2) We found that the first approach can generate AEs with MR of 100%. However, they do not guarantee the practicality and functionality of the crafted AEs, unlike the GEA approach.

II. Generating Adversarial Examples

In order to generate realistic AEs that preserve the functionality and practicality of the original samples we design two approaches: generic adversarial machine learning attacks and GEA. More information regarding the proposed approaches are presented in §II-A and §II-B.

A. Off-the-Shelf Adversarial Attacks (OSAA)

This approach incorporates well-established adversarial machine learning attack methods into IoT malware detection. These methods apply small perturbation into the feature space to generate AEs that lead to misclassification.

B. Graph Embedding and Augmentation (GEA)

Assume an original sample x_{org} and a selected target sample x_{sel}, our main goal is to combine the two samples while preserving the functionality and practicality of x_{org} and achieving misclassification. Prior to generating the CFG for these algorithms, we compile the code using GNU Compiler Collection (GCC) command. Afterwards, Radare2 is used to extract the CFG from the binaries. 1(a) and 1(b) show the generated graphs for x_{org} and x_{sel}, respectively.

III. Evaluation and Discussion

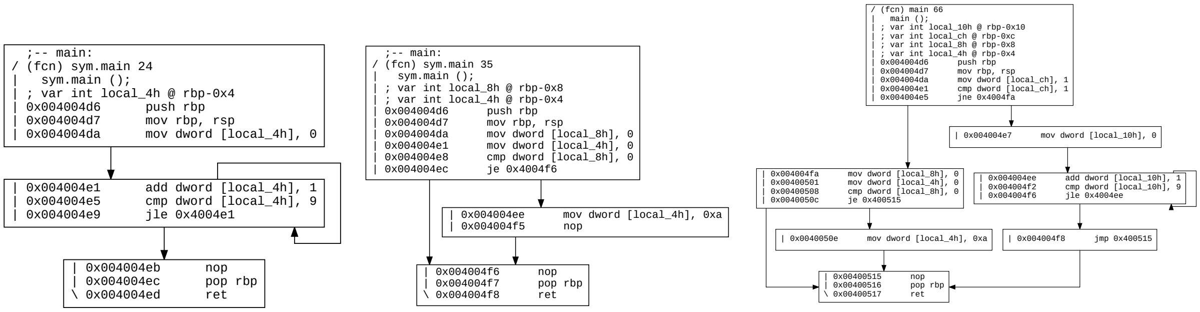
A. Dataset

We obtained the CFG dataset of the IoT malware from Alasmary *et al.* [3] to assess our proposed approach. The dataset consists of 2,281 malicious and 276 benign IoT samples. We extracted 23 different features in seven different groups, including betweenness centrality, closeness centrality, degree centrality, shortest path, density, # of edges, and # of nodes.

B. Results & Discussion

1) Deep Learning-based IoT Malware Detection System: We designed a CNN-based classifier, which distinguishes IoT malware samples from benign ones, trained over 23 CFG-based features categorized in seven groups, including betweenness centrality, closeness centrality, degree centrality, shortest path, density, # of edges, and # of nodes, extracted from CFGs of 2,281 malware and 276 benign samples. We achieved an accuracy rate of 97.13% with a False Negative Rate (FNR) of 11.26% and False Positive Rate (FPR) of 1.55%. It is worth mentioning that the high value of FNR is due to the imbalanced number of malware and benign samples.

2) OSAA: We implemented six generic adversarial learning attack methods to generate AE by perturbing the feature space. Overall, those approaches have shown, in general, a good performance (see Table I).



(a) Original sample's CFG (b) Taraget sample's CFG (c) Crafted adversarial CFG using GEA

Fig. 1. A practical implementation of the GEA approach. Fig. 1(a) shows the generated CFG for the original sample and used for extracting graph-based features (graph size, centralities, etc.) for graph/program classification and malware detection. 1(b) shows the graph for the selected target sample generated as in Fig. 1(a). Finally, The generated adversarial graph using GEA approach. Note that this graph is obtained logically by embedding the graph in Fig. 1(b) into the graph in Fig. 1(a).

TABLE I
EVALUATION OF THE GENERIC ADVERSARIAL LEARNING ATTACK METHODS. MR: MISCLASSIFICATION RATE, AVG.FG: AVERAGE NUMBER OF CHANGED FEATURES,AND CT: COMPUTATION TIME.

Attack Method	MR (%)	Avg.FG	CT (ms)
C&W [7]	100	12.60	25.30
DeepFool [8]	86.39	14.90	2.56
ElasticNet [9]	100	5.42	114.18
JSMA [10]	99.80	4.00	0.78
MIM [11]	100	20.60	0.90
PGD [12]	100	22.56	2.40

TABLE II
GEA: MALWARE TO BENIGN (MAL2BEN) AND BENIGN TO MALWARE (BEN2MAL) MISCLASSIFICATION RATE. MR: MISCLASSIFICATION RATE, CT: COMPUTATIONAL TIME.

	Size	# Nodes	MR (%)	CT (ms)
Mal2Ben	Minimum	2	7.67	33.69
	Median	24	95.48	37.79
	Maximum	455	100	1,123.12
Ben2Mal	Minimum	1	30.65	40.65
	Median	64	57.60	69.23
	Maximum	367	88.04	473.91

3) GEA: This approach is designed to generate a practical AE that fools the classifier, while preserving the functionality and practicality of the original sample. Here, we discuss the inherent overhead of the GEA approach. We investigate the impact of the size of the graph, determined by the number of the nodes in a graph, and graph density, determined by the number of edges in a graph while the number of nodes is fixed. Note that all generated samples maintain the practicality and the functionality of the original sample. The obtained results are discussed in more detail in the following.

Graph Size Impact. We selected three graphs, as targets, from each of the benign and malicious IoT software, consisting of a minimum, median and maximum graph size, and the goal was to understand the impact of size on MR with GEA. The results are shown in Table II. We found that the MR increases when the number of nodes increases, which is perhaps natural. In addition, the time needed to craft the AE is proportional to the size of the selected sample. We achieved a malware to benign MR of as high as 100%, and a benign to malware MR of 88.04%, while insuring that the original samples are executed as intended, a property not guaranteed with the off-the-shelf adversarial attack methods.

IV. CONCLUSION

In this work, we generated the CFGs of the IoT samples, we then extracted 23 representative features from the CFGs to train our DL model. The focus of this study is to investigate the robustness of the trained DL model. Thus, we designed two approaches, including OSAA methods and GEA. OSAA methods incorporates six different attacks to generate the AE. In our evaluation, we obtain a MR of up to 100% using these attacks. GEA approach focuses on preserving the functionality and practicality of the generated samples, which is not guaranteed in OSAA methods. Our evaluation showed that GEA is able to misclassify all malware samples as benign.

V. ACKNOWLEDGMENT

This work was supported by NRF-2016K1A1A2912757, NVIDIA GPU Grant Program, Office of Research and Commercialization Fellowship, and a collaborative seed grant from the Florida Cybersecurity Center (FC2).

REFERENCES

[1] A. Gerber. (Retrieved, 2017) Connecting all the things in the Internet of Things. [Online]. Available: https://ibm.co/2qMx97a
[2] A. Azmoodeh, A. Dehghantanha, and K.-K. R. Choo, "Robust malware detection for Internet Of (Battlefield) Things devices using deep eigenspace learning," *IEEE Transactions on Sustainable Computing*.
[3] H. Alasmary, A. Anwar, J. Park, J. Choi, D. Nyang, and A. Mohaisen, "Graph-based comparison of IoT and android malware," in *Proceedings of the 7th International Conference on Computational Data and Social Networks, CSoNet*, 2018, pp. 259–272.
[4] A. Mohaisen, O. Alrawi, and M. Mohaisen, "AMAL: high-fidelity, behavior-based automated malware analysis and classification," *Computers & Security*, vol. 52, pp. 251–266, 2015.
[5] A. Khormali and J. Addeh, "A novel approach for recognition of control chart patterns: Type-2 fuzzy clustering optimized support vector machine," *ISA transactions*, vol. 63, pp. 256–264, 2016.
[6] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. D. McDaniel, "Adversarial examples for malware detection," in *Proceedings of the 22nd European Symposium on Research Computer Security - ESORICS, Part II*, 2017, pp. 62–79.
[7] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.
[8] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.
[9] P. Chen, Y. Sharma, H. Zhang, J. Yi, and C. Hsieh, "EAD: elastic-net attacks to deep neural networks via adversarial examples," in *Proceedings of the Conference on Artificial Intelligence*, 2018.
[10] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy*, Saarbrücken, Germany, Mar. 2016, pp. 372–387.
[11] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the 2018 International Conference on Learning Representations.*, 2018.

Examining Adversarial Learning against Graph-based IoT Malware Detection Systems

Ahmed Abusnaina Aminollah Khormali Hisham Alasmary Jeman Park Afsah Anwar Ulku Meteriz Aziz Mohaisen

{ahmed.abusnaina, aminkhormali, hisham, parkjeman, afsahanwar, meteriz}@knights.ucf.edu, and mohaisen@cs.ucf.edu

INTRODUCTION

Background

- Internet of Things (IOT) devices are heavily used in a wide variety of applications, e.g. sensors, automation tools, etc.
- They can be exploited through software vulnerabilities, leading to security threats and impacts, e.g. DDoS attacks.
- Analyzing, understanding, and classification of IoT software vulnerabilities are critical, specifically in sensitive applications.
- Research on analysis of IoT is limited in both size of the analyzed samples and the utilized approaches.
- One promising research direction is incorporating graph-theoretic approach and machine/deep learning to analyze IoT malware.
- Deep learning networks are vulnerable to adversarial examples.

Goal

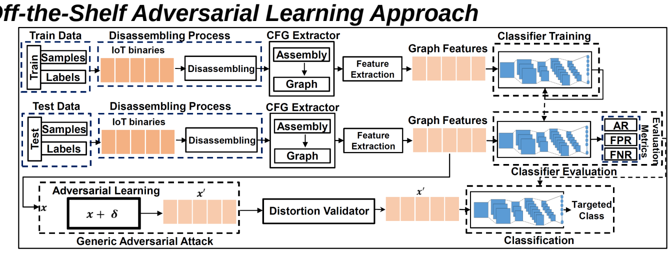
- Investigating the robustness of graph-based deep learning IoT malware detection systems against adversarial learning attacks.

Approach

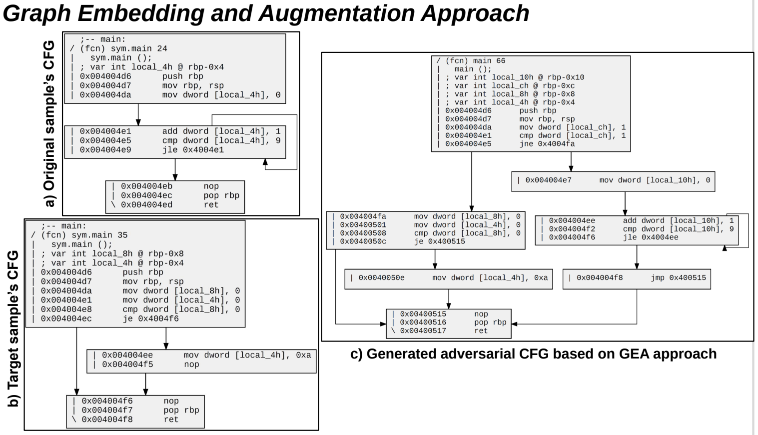
- *Off-the-Shelf Adversarial Learning:* exploring feasibility of six generic adversarial attack methods to generate practical adversarial graph samples that misleads the CNN-based model to misclassification.
- *Graph Embedding and Augmentation (GEA):* Carefully embedding and augmenting selected target graph to the original graph, while maintaining the practicality and functionality of the crafted AE.

Methodology

Off-the-Shelf Adversarial Learning Approach



Graph Embedding and Augmentation Approach



c) Generated adversarial CFG based on GEA approach

Evaluation

Dataset

- 2,281 malicious and 276 benign IoT samples, totaling to 2557.
- 23 different CFG-based features categorized in seven different groups, including betweenness centrality, closeness centrality, degree centrality, shortest path, density, # of edges, and # of nodes.

Deep learning-based IoT malware Detection System

- A Convolutional Neural Network (CNN) model consisting of four consecutive layers, trained over 23 CFG-based features
- Reached a classification rate of 97.13% with false positive rate and false negative rate of 1.55% and 11.26%, respectively.

Off-the-Shelf Adversarial Learning Approach*

Attack Method	MR (%)	Avg.FG	CT (ms)
C&W [8]	100	12.60	25.30
DeepFool [9]	86.39	14.90	2.56
ElasticNet [10]	100	5.42	114.18
JSMA [11]	99.80	4.00	0.78
MIM [12]	100	20.60	0.90
PGD [13]	100	22.56	2.40

Graph Embedding and Augmentation Approach*

	Size	# Nodes	MR (%)	CT (ms)
Mal2Ben	Minimum	2	7.67	33.69
	Median	24	95.48	37.79
	Maximum	455	100	1,123.12
Ben2Mal	Minimum	1	30.65	40.65
	Median	64	57.60	69.23
	Maximum	367	88.04	473.91

* Here, MR: misclassification rate, AVG.FG: average number of changed features, CT: computation time.

Discussion and Conclusion

Discussion

- Off-the-shelf adversarial attack methods are able to achieve a high misclassification rate of 100%.
- GEA approach is able to misclassify all malicious samples as benign.
- Also, more than 88% of the benign samples are classified as malware.
- While the first approach cannot guarantee the practicality of the generated AEs, the GEA approach crafts realistic CFGs.
- The performance of the GEA approach is directly correlated with the size (# nodes) of the selected target sample.
- The computation time increases once the size of the IoT software increases.

Conclusion

- Graph-based deep learning IoT malware detection systems are vulnerable to adversarial machine learning attacks.
- More robust IoT malware detection systems are required, such as sophisticated features, unlike CFG-based ones.

NDSS Symposium ,
San Diego, California
24-27 February, 2019

UCF

This work was supported by NRF-2016K1A1A2912757, NVIDIA GPU Grant Program, Office of Research and Commercialization Fellowship, and a collaborative seed grant from the Florida Cybersecurity Center (FC2).