JOURNAL OF TRANSLATIONAL ENGINEERING, VOL 5, OCTOBER 2019

A Critical review on Adversarial Attacks on Intrusion Detection Systems

Frank. N, *Research Associate, Department of Engineering and Technology, PhD Research Lab*. UK
Dr. Nancy, *Head, Technical Operations, PhD Research Lab*, UK.

***Abstract*—Deep Learning is a major leap forward in terms of technology, however they are easily susceptible to Adversarial attacks. Adversarial attacks are a major problem to Intrusion Detection Systems since they are difficult to implement. It is necessary to identify feasible and effective solutions for providing security against the attack.**

Index Terms— **Computer Science Research, Artificial Intelligence, Machine Learning, Training Error, Adversarial Attack**

I. Research Papers

Adversarial attacks are inputs to a model introduced deliberately and designed by attackers to malfunction the detection systems by modifying the trained models (Wang, Li, Kuang, Tan, & Li, 2019). In image data, the attacker modifies the images slightly in such a way that it seems good in a visual sense, but the model recognizes it as some other object. This is similar to creating an optical illusion for the models(S. Chen et al., 2018).

Let us review some of the literature dealing with Adversarial attacks.

A. Perturbantion optimized black-box adversarial attack via genetic algorithm (POBA-GA)

J. Chen, Su, Shen, Xiong, & Zheng,(2019) has studied adversarial attacks and proposed a novel technique for dealing with it. According to the paper, deep learning models are more vulnerable to the attacks. It is difficult to provide security against such attacks and the existing models are not robust enough. The recent attacks use their own target models with their own evaluation metrics. Since this attack takes place suddenly and its characteristics are difficult to analyze, it comes under the category of black box. White box attacks are easier to detect and prevent when compared to black box attacks. Hence, a novel Genetic algorithm based approach known as Perturbation Optimized Black-Box Adversarial Attack Genetic Algorithm (POBA-GA) has been proposed for getting the results comparable to white box attacks. The fitness function has been designed specifically for more efficiency. Also, the diversity of the population is modified accordingly for better perturbations. From the analysis, it has been seen that the algorithm works efficiently better than the existing techniques. However, this study uses CIFAR-10 and MINST datasets which are old and outdated. More recent datasets can be used for further analysis.

B. Adversarial attacks on deep-learning based radio signal classification

Sadeghi & Larsson,(2019) has also focussed on deep learning algorithms with respect to adversarial attacks. It has been seen that the deep learning is very vulnerable to these types of attacks in spite of its success in other applications. This paper uses the deep learning for radio signal applications and hence considers both black box and white box attacks. The classification efficiency is greatly reduced by these attacks with lots of perturbations in the input side. The work has been analysed and it has been seen that not much power is required for inducing the attacks into the signals. On the other hand, conventional jamming requires lots of power to attack. Hence, it has been conveyed that there is a loop hole in using deep learning techniques. However, a suitable solution is not provided to address the issue in this work.

C. Fuzzy classification boundaries against adversarial network attack

Iglesias, Milosevic, & Zseby,(2019) has proposed fuzzy based approach for classifying the boundaries against adversarial attacks. The methods will learn on their own for preventing the attackers from taking advantage. These attacks are either modified old attacks or new network attacks. This paper has proposed to distort the boundaries of classification approach to improve the efficiency of detection of the attacks. This is done to fix the problem where the machine learning has learning deficiency due to the fixed boundaries. The distorted boundaries fix this issue and hence the training takes place well. Decision Tree approach has been used as a classification technique and it has been tested with both linear decision tree and fuzzy decision tree. It has been seen that the performance of the classification has improved when the membership score is considered as additional feature. Since fuzzy is used, it is easier for identifying the adversarial attacks. The accuracy of detection is however not as high as expected since it has an accuracy of 76% for KDD dataset and 84% for NB15 dataset.

D. Adversarial examples for CNN-Based malware detectors

Convolutional Neural Network (CNN) has been useful in various applications like identifying images, classifying texts and speeches with great efficiency. However, when their performance is tested with adversarial attacks, it is not as expected. The existing techniques with neural networks have low efficiency, hence two novel approaches with white box techniques were proposed in B. Chen, Ren, Yu, Hussain, & Liu, (2019)

to evaluate other malware detection techniques. Gradient based algorithms has been incorporated to one of the white box algorithm and an accuracy of 99% has been obtained. However, if the parameters and structures are not known, then the obtained accuracy is 70%. Additionally, adversarial training is performed for resisting evasion attacks. Also, the security risks of existing adversarial approaches are used as an example for the training process. The process has been proved to work effectively. In this work, a novel classification technique is not proposed, but it is tested on existing techniques.

E. Defending non-Bayesian learning against adversarial attack

Su & Vaidya,(2019) have proposed a novel technique to provide security against adversarial attacks. The models collect data continuously to understand the characteristics of the attacks and hence train continuously. However, there is a problem of non-Bayesian learning in the network. This paper solves this problem by considering the adversarial agents for the analysis. Lots of Byzantine faults are present which must be addressed. Two rule based approaches are proposed and the approach works even when there is no data transfer. It has been seen that there is an adjustment between the ability to achieve consensus and the problem of network detectability. This problem will be addressed as a future scope.

F. Defending against whitebox adverasial attack via randomized descretization

While, other papers have used machine learning techniques, (Zhang & Liang, 2019) have used randomised discretisation for Whitebox adversarial attacks. Since the adversarial perturbations reduce the accuracy of the classification, a simple and effective defence strategy is proposed and analysed. First, random Gaussian noise is introduced, and then the pixels are discretised and then given to any classification approach. It has been seen that this randomised discretisation decreases the differences between actual and adversarial packets leading to better effectiveness in terms of classification accuracy. Datasets are selected and then implemented for measuring the performance. It has been seen that the proposed approach boosts the efficiency of detecting the adversarial attacks. In this work, a novel classification technique is not proposed, but it is tested on existing techniques.

TABLE I
SUMMARY OF THE LITERATURE ARTICLES

S.No	Work By	Method	Dataset	Results	Limitation / Future Scope
1	J. Chen, Su, Shen, Xiong, & Zheng (2019)	Perturbation Optimized Black-Box Adversarial Attack Genetic Algorithm (POBA-GA)	CIFAR-10 and MINST	It has been seen that the algorithm works efficiently better than the existing techniques.	The datasets used are old and outdated. More recent datasets can be used as a future scope
	Sadeghi & Larsson (2019)	Deep Learning Methods	RML2016	learning for radio signal applications and hence considers both black box and white box attacks. The classification efficiency is greatly reduced by these attacks with lots of perturbations in the input side. The work has been analysed and it has been seen that not much power is required for inducing the attacks into the signals	It has been conveyed that there is a loop hole in using deep learning techniques. However, a suitable solution is not provided to address the issue in this work.
	Iglesias, Milosevic, & Zseby,(2019)	Fuzzy based approach, Decision Tree	KDD CUP, NB-15	The boundaries are distorted in the classification approach to improve the efficiency of detection of the attacks.	The accuracy of detection is however not as high as expected since it has an accuracy of 76% for KDD dataset and 84% for NB15 dataset.
	B. Chen, Ren, Yu, Hussain, & Liu, (2019)	Gradient based algorithm	VirusShare, DAS, and Malware Benchmark	An accuracy of 99% has been obtained. However, if the parameters and structures are not known, then the obtained accuracy is 70%.	A novel classification technique is not proposed, but it is tested on existing techniques.
	Su & Vaidya,(2019)	Two rule based approaches	-	The models collect data continuously to understand the characteristics of the attacks and hence train continuously. The approach works even when there is no data transfer	There is an adjustment between the ability to achieve consensus and the problem of network detectability. This problem will be addressed as a future scope.
	(Zhang & Liang, 2019)	Randomised Discretisation	MNIST and ImageNet	Randomised discretisation decreases the differences between actual and adversarial packets leading to better effectiveness in terms of classification accuracy	A novel classification technique is not proposed, but it is tested on existing techniques.

II. Conclusion

- From the critical review of the literature, it is seen that there are difficulties in creating models theoretically for the adversarial process.
- The machine learning approaches must create very good outputs and must perform very well for efficient detection of adversarial attacks.
- Some of the literature have used old datasets (J. Chen et al., 2019), which others have identified the problems empirically without providing any particular solution to the problems (Sadeghi & Larsson, 2019).
- Solution is provided to existing techniques in some literature, but novel techniques are not suggested (B. Chen et al., 2019; Zhang & Liang, 2019). The solution has been provided and

implemented in Iglesias et al., (2019), however, the accuracy of detecting the attacks is low.

- Most of the available techniques are hence not very effective in predicting the adversarial attacks even though they are very efficient in detecting ordinary attacks.Hence, creating an effective defence approach is very much a necessity for the future. More adaptive techniques will be the major objective of future Intrusion Detection Systems.

References

1. Chen, B., Ren, Z., Yu, C., Hussain, I., & Liu, J. (2019). Adversarial Examples for CNN-Based Malware Detectors. IEEE Access, 7, 54360–54371. https://doi.org/10.1109/ACCESS.2019.2913439

2. Chen, J., Su, M., Shen, S., Xiong, H., & Zheng, H. (2019). POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm. Computers & Security, 85, 89–106. https://doi.org/10.1016/j.cose.2019.04.014

3. Chen, S., Xue, M., Fan, L., Hao, S., Xu, L., Zhu, H., & Li, B. (2018). Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach. Computers & Security, 73, 326–344. https://doi.org/10.1016/j.cose.2017.11.007

4. Iglesias, F., Milosevic, J., & Zseby, T. (2019). Fuzzy classification boundaries against adversarial network attacks. Fuzzy Sets and Systems, 368, 20–35. https://doi.org/10.1016/j.fss.2018.11.004

5. Sadeghi, M., & Larsson, E. G. (2019). Adversarial Attacks on Deep-Learning Based Radio Signal Classification. IEEE Wireless Communications Letters, 8(1), 213–216. https://doi.org/10.1109/LWC.2018.2867459

6. Su, L., & Vaidya, N. H. (2019). Defending non-Bayesian learning against adversarial attacks. Distributed Computing, 32(4), 277–289. https://doi.org/10.1007/s00446-018-0336-4

7. Wang, X., Li, J., Kuang, X., Tan, Y., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing, 130, 12–23. https://doi.org/10.1016/j.jpdc.2019.03.003

8. Zhang, Y., & Liang, P. (2019). Defending against Whitebox Adversarial Attacks via Randomized Discretization. Retrieved from http://arxiv.org/abs/1903.10586