TITLE

Names

[#]Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh

ABSTRACT

content...

1. DEFENSE METHODS

Nowadays many studies have been done to ensure AI's security.So, The attackers won't be able to affect the data in various cases. Three methods were suggested by Hendrycks and Gimpel [1] for finding adversarial images. A hardware-assisted randomization method against adversarial examples was proposed by Zhang et al. [2] to defend against many adversarial attacks. Now AI attacks have three divisions: evasive attack, poisoning attacks, and backdoor attack. Defense Techniques of the AI system during various phases like collection of data, training and use of model are listed in Table **??**. Moreover, more studies are needed to strengthen the understanding of ML and to build AI security platforms [3].

4.1 Evasive Attack's Defense Methods

4.1.1 Adversarial Training. An essential way to strengthen the robustness of neural network models is Adversarial Training. This technique's main rule is, it uses known attack techniques to generate adversarial samples in the model training stage. After that it adds samples to training set of the model, and the model is trained again and again till a new model is generated which can resist disturbance. So, this technique strengthens the robustness of the new model as well as it also strengthens accuracy and standardization of the model.

E-ABS (analysis-by-synthesis) proposed by Ju et al. [4], can broaden the ABS robust classification model to more difficult image domains. The basic elements include: (1) generation model; (2) discriminative loss; (3) variational inference; (4) lower bound for the robustness of E-ABS. Nonetheless, the generation model is responsive to image likeliness calculations. The operational effectiveness and the reasoning of ABS-like model are low on huge datasets.

4.1.2 Network Distillation. The method of compressing the knowledge of a large network into smaller networks is called Distillation. Specialist models signifies that, multiple specialized networks can be trained to upgrade the model performance for a large network. The use of distillation is normally to train a model at first which is large. Then the large model is heated. Soft target is the output of large model and the data label becomes hard target. Then the two are merged to train the small model.

The basic idea of the network distillation technique is to series many DNNs in the training stage. Here, Former DNN generates classification results and these results are used to train the later DNN. Papernot et al. [5] established that the transfer knowledge decreases sensitivity of model and enhances the robustness of the AI model. Hence, network distillation technology is suggested to provide defense against evasive attacks, and it has been tested on MNIST and CIFAR-10 datasets.

4.1.3 Adversarial Example Detection. The main purpose of adversarial example detection is to detect that is either an example is adversarial or not, and it is done by adding the detection element of the external detection model or the original model which is used is the usage state. The detection model identifies an example as adversarial example or not before the input example gets to original model. The detection model is also able to extract appropriate information from each part of the original model and synthesize different information. Several detection models can use various criteria to evaluate whether the input is adversarial or not [6].

4.1.4 Input Reconstruction. The concept of input reconstruction is that the input samples are changed to withstand evasive attack when the model is in use stage, and the changed data will not influence the normal classification function. Some reconstruction methods are noising, preprocessing, denoising, gradient-masking and auto encoder that changes the input examples [7].

4.1.5 DNN Verification. DNN verification technology verifies properties of DNN models by using solvers, like proving that no adversarial example is found within a particular disturbance range. Nonetheless, the efficiency of solver is low and the DNN model is also confirmed as NP-complete problem. But the operation efficiency can be improved by selection and optimization, like validation by region, sharing validation information [8].

4.1.6 Data Augmentation. The insufficiency of data is a

Table 1. The security defense technology of AI

Type	Phase		
	Data collection phase	Model train phase	Model usage phase

Evasive attack	Generating adversarial examples	Network distillation;
		adversarial training
			Adversarial examples detection;
			input reconstruction;
			DNN model validation

Poisoning attack	Filtering training data; regression analysis	Integration analysis	
Back door attack		Model pruning	Input preprocessing

real-life problem. Data augmentation solves the problem by enlarging the original training sets. And it is done by generating adversarial samples. When a large amount of data is lacking, it produces enough samples to make sure that the training of model is done effectively[9].

4.2 Poisoning Attack's Defense Methods

4.2.1 Training Data Filtering. The main purpose of this technique is to control the training dataset and it uses various methods to stop the poisoning attack. Particular directions include [10]: finding probable poisoning attack data points and filtering the points during next training; a method is used to decrease sampling data which could be used by poisoning attack, this method is called model contrast filtering method. The filtered data is used against the attack.

4.2.2 Regression Analysis. This method is used to detect outliers and noise in datasets, and it is based on statistics. There are particular method including various loss functions to check the outliers. This uses distribution properties of data for identification [11], etc.

4.2.3 Ensemble Learning. It is used to build and merge multiple machine learning classifiers to enhance the ability of the ML system to withstand poisoning attacks. Multiple models which are independent joins and forms the AI system. And the systems possibility to get affected by poisoning attack is reduced to the numerous training datasets affected by multiple models [12].

4.2.4 Iterative Retraining. Iterative retraining means repeated training of neural networks. This method generates adversarial examples following an attack model and adds these examples to the training data. Then it attacks the neural model and repeats the process [13].

4.3 Backdoor Attack's Defense System

4.3.1 Input Processing. The principle of this method is to purify the input. The filtering process decreases the risk of input triggering the back door. It also reduces the risk of changing the model judgement [14]. Generally, data can be divided into two types: continuous and discrete. The data in image is continuous. The Preprocessing operation is differentiable and linear such as mean standardization. The data in a text is discrete and the pretreatment operation is nondifferentiable and nonlinear.

4.3.2 Model Pruning. The term pruning in neural network came from synaptic pruning which occurs in human brain. The axons and dendrites of human brain becomes dead in this synaptic elimination process. It occurs between childhood and puberty in mammals. This model focuses on cutting of the neurons of the original model. Thus, it reduces the possibility of backdoor neurons, where the condition for backdoor neuron is that normal function is consistent. The backdoor neurons can be removed by using a fine-grained pruning method [15].

2. REFERENCES

[1] Dan Hendrycks and Kevin Gimpel, "Early methods for detecting adversarial images," *arXiv preprint arXiv:1608.00530*, 2016.

[2] Jiliang Zhang, Shuang Peng, Yupeng Hu, Fei Peng, Wei Hu, Jinmei Lai, Jing Ye, and Xiangqi Wang, "Hrae: hardware-assisted randomization against adversarial example attacks," in *2020 IEEE 29th Asian Test Symposium (ATS)*. IEEE, 2020, pp. 1–6.

[3] Zixiao Kong, Jingfeng Xue, Yong Wang, Lu Huang, Zequn Niu, and Feng Li, "A survey on adversarial attack in the age of artificial intelligence," *Wireless Communications and Mobile Computing*, vol. 2021, 2021.

[4] An Ju and David Wagner, "E-abs: extending the analysis-by-synthesis robust classification model to more complex image domains," in *Proceedings of the 13th ACM Workshop on Artificial Intelligence and Security*, 2020, pp. 25–36.

[5] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[6] Fabio Carrara, Rudy Becarelli, Roberto Caldelli, Fabrizio Falchi, and Giuseppe Amato, "Adversarial examples detection in features distance spaces," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[7] Shixiang Gu and Luca Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

[8] YG Qian, Xi-Ming Zhang, Bin Wang, Wei Li, Jian-Hai Chen, Wujie Zhou, and Jing-Sheng Lei, "Towards robust dnns: a taylor expansion-based method for generating powerful adversarial examples," *CoRR*, 2020.

[9] Yucheng Shi and Yahong Han, "Schmidt: Image augmentation for black-box adversarial attack," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[10] Ricky Laishram and Vir Virander Phoha, "Curie: A method for protecting svm classifier from poisoning attack," *arXiv preprint arXiv:1606.01584*, 2016.

[11] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 19–35.

[12] Deqiang Li and Qianmu Li, "Adversarial deep ensemble: Evasion attacks and defenses for malware detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3886–3900, 2020.

[13] Sungrae Kim and Hyun Kim, "Zero-centered fixed-point quantization with iterative retraining for deep convolutional neural network-based object detectors," *IEEE Access*, vol. 9, pp. 20828–20839, 2021.

[14] Yuntao Liu, Yang Xie, and Ankur Srivastava, "Neural trojans," in *2017 IEEE International Conference on Computer Design (ICCD)*. IEEE, 2017, pp. 45–48.

[15] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 273–294.