IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 14, PP 14-18

Machine Learning Under Attack: Literature Survey

Aruna Pavate [1], Divya Kumawat [2], Suvrna Pansambal[3], Pranav Nerurkar [4], Dr. Rajesh Bansode[5]

[1, 2, 3](Computer Engineering, Atharva College of Engineering, Mumbai University, India)
[4](Computer Engineering, Veermata Jijabai Technological institute, Mumbai University, India)
[5](Computer Engineering, Thakur College of Engineering/ Mumbai University, India)

Abstract : *Machine Learning is an area that carries many advantages with it for businesses that assume it. Machine Learning is a data exploration process which influences machine learning algorithms to repeatedly study from the present data and assist the machine to find hidden patterns without being programmed for. Most of the present machine learning classifiers are extremely exposed to adversarial examples. The study clearly indicates that Machine learning is of huge importance given in computer vision applications but it may possible to attack the machine learning algorithm so that it is necessary to design and develop different defense systems. This study directed the detailed study of the previous approaches and the comparative analysis among these approaches.*

Keywords: *Machine Learning, DEEP Learning, Security, adversarial attack, Neural Network*

I. Introduction

Deep learning is functionally present in many safety-critical applications. Safety critical applications domain includes aircraft flight control, medical devices, nuclear systems and weapons. Safety critical systems are the systems whose failure would outcome in loss of life, significant property damage, or damage to the environment problem. [1] Most of the machine learning (ML) algorithms are used to solve the security problems such as Authentication, Email spam detection, Network Intrusion /Malware detection, Email spam detection.[2] Most of the present machines learning classifiers are extremely exposed to adversarial examples. The researcher provides a complete categorization of various attacks aimed at exploiting machine learning systems: causative, attacks on integrity and availability, exploratory attacks, targeted attacks; indiscriminate attacks, adversarial attack. [2] Deep neural network is one of the most popular algorithm has been used in a many areas like object detection [3], [4], in image classification for object recognition [5], [6], speech recognition [7], voice synthesis [9], language translation [8]. In deep neural network, data can be digging out with sophisticated and more abstract level depiction from raw data. Most of the machines learning classifiers are highly revealed to adversarial examples. However, recently deep neural network (DNN) has been found vulnerable to well-prepared input data .Large amounts of application based on deep neural network have been used in physical world especially in the safety-critical situations. A recent study shows that most of the adversarial examples applied in physical world .Any machine learning classifier could be fooled to give inaccurate predictions and with a little bit of knowledge, you can get the classifiers to give good result as you want revealed by Google Brain in his studies.[10] By using adversarial examples to any machine learning models make ML models produce wrong output with high confidence. Adversarial examples designed by employing few but deliberately worst-case perturbations to normal samples.

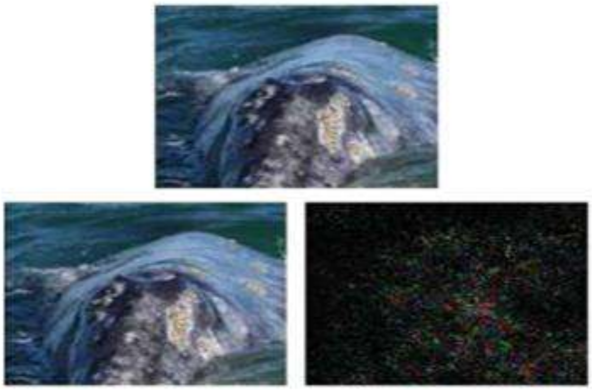

Fig 1: Source: DeepFool: a simple and accurate method to fool deep neural networks [11]

Figure 1 shows an adversarial example. The first row shows the original image classified as “whale” by the classifier. The second row shows the image classified as “turtle “after adding some perturbation to the original image by using DeepFool method. Moosavi-Dezfooli showed that by adding smaller perturbation leads to the classifier to give an incorrect result. Thus, an adversarial example necessitates being overlooked carefully. Adversarial examples used by many researchers to attack against the different system in different domains. In this paper, we outline the approaches for generating adversarial attacks. The respite of the paper is planned as follows: In section II we analysis different methods to generate adversarial examples. Section III describes comparative analysis of adversarial training algorithm. At last, section IV describes conclusion of the topic.

II. Methods For Rendering Adversarial Attacks

In this section, we discuss different approaches for rendering adversarial examples.

Christian Szegedy, presented the first adversarial attack for deep neural network. The network can execute to misclassify an image by maximizing the network’s prediction error. This method called as Linear based Fast Gradient Sign Method. The author proved that the generated adversarial attack can be applied to different models, the different network that was trained on different dataset misclassify the same input. They produced adversarial attack using an L-BFGS method to solve the general targeted the validation measure used was perturbation magnitude i.e. ℓ 2 & attack frequency was iterative. Error rate= 2.1% and distortion= 0.058.L-BFGS method was time consuming and not easy to implement practically as it was based on an expensive linear search method [12] Pedro Tabacof and Eduardo Valle pictured that shallow classifiers be there more robust adversarial examples as compared to deep convolutional network. [13]

Good fellow proposed a rapid method for rendering adversarial examples motivated by nonlinearity and over fitting problem. This method is referred as FGSM (Fast Gradient Sign Method). The author claimed that the main cause of neural networks vulnerability to adversarial perturbation is the linear nature of the neural network system but linear behavior raises the speed of training the network. Good fellow demonstrated that adversarial training can result in regularization. The validation measure used was perturbation magnitude i.e. element-wise & attack frequency was One -time. The author showed an accuracy of the system was 68% and misclassification rate=93%, with adversarial training error rate fell to 17.9%. The behavior of the system is more locally stable and interpretable. [14] Andras Rozsa developed a fast gradient value method (FGVM) by changing the sign of gradient with the raw gradient in the fast gradient sign method. This study showed the method for generating the much more varied set of adversarial examples than the above existing methods.[12][14]The proposed algorithm improved both the accuracy and robustness of the system. The author introduced the new measure to quantify adversarial images PASS(Perceptual Adversarial similarity Score) which is higher than 0.99 for this method. The validation measure used was perturbation magnitude i.e. element-wise & attack frequency was One –time. Fast gradient value method can generate images with the larger local difference. Gradient-based attacks are concluded ineffective. [15] A. Nguyen, J. Yosinski, and J. Clune exposed a novel kind of attack where deep neural network was used to classify adversarial examples. This attack referred as compositional pattern-producing network-encoded EA (CPPN EA). A deep neural network classifies the adversarial examples with high confidence (99%) but in this however, objects are not identifiable by the human eye. For generating adversarial examples Evolutionary algorithms (EAs) algorithm has been used. Performance of the attack generated measured by using perturbation magnitude i.e. N.A. & attack frequency was iterative [16]

Papernot et al. designed new attack called Jacobian-based Saliency Map Attack (JSMA). Jacobian-based Saliency Map Attack changes a small fraction of feature/pixel to cunning in each iteration which successfully introduces large variation of output that could fool the neural network. To pick the feature/ pixel to be cunning in each iteration authors introduced two adversarial saliency maps. Authors achieved 97% adversarial success rate by modifying only 4.02% input features per sample (Iterative and ℓ 2). However, this method runs very slow due to its considerable computational cost. [17]

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, invented new adversarial attack generation method called Deepfool to explore the nearest space from the concrete input to the decision boundary of adversarial examples . Authors performed an iterative attack by linear approximation to defeat the nonlinearity problem in high dimension space. The adversarial attack could be possible to generate for both the binary classifier & multiclass classifier. This method provides less perturbation as compared to Fast Gradient Sign Method (FGSM) and Jacobian-based Saliency Map Attack (JSMA). DeepFool method also concentrated the strength of perturbation in its place of the number of selected features.[11] A. Kurakin, I. Goodfellow, and S. Bengio proposed a new method mentioned as Basic Iterative Method (BIM). Most of the data generated in the physical world directly by using different devices like CCTV cameras, camera cell phone, sensors etc. When data directly comes from the physical world the generated result shows that adversarial examples are classified wrongly. This method is rarely used to attack on specific class and also this method tries to maximize the cross entropy loss. This approach known as Iterative Least-Likely Class method (ILLC).The result generated by this method showed that neural network can be straightforwardly fooled by image taken from physical world. Previous methods believed that adversarial attack can be generated by clearly given as input to the deep neural networks. Author tried this kind of attack to the real world problems. Fast Gradient Sign Method has been unmitigated by adding a smaller change for multiple iterations. All the previous methods considered adversarial data can be openly feed into the deep neural network. As most of the data in the real world comes from the devices directly the problem of openly feed has been solved to some extent. [18] Alexey Kurakin has developed new method referred as one -step Target Class Method. This method was an unmitigated version of Fast Gradient Sign Method (FGSN). This method does not guarantee that an adversarial image will be classified incorrectly. Author has proposed new method which was an unmitigated version of fast gradient sign method. This method raises the probability of attack to the target class. By using this method attacker either obtains the expected outcome and sometimes the machine learning model wins. This method is One –step Target Class method cannot be iterative methods and cannot stick with photo transformation [19] F. Tramer proposed an extension to fast gradient sign method. Author found that the fast gradient sign method with adversarial training is more robust to white box attack because of gradient masking than the black box attack and also author projected an addition of adversarial training that furthermore augments' training data with perturbed inputs. To fail adversarial training they added a random while updating the adversarial examples and the algorithm named as RAND-FGSM. The author claimed that their attack is "single-step" attacks on models trained with or without adversarial training. This attack is much more computationally expensive. [20]

Carlini and Wagner founded a targeted attack to overcome defensive distillation. Carlini & Wagners's attack is operative for most of existing adversarial detecting defenses. Author showed that if the magnitude of input is reduced artificially and other minor changes then the attack flourishes. The attack successfully misclassified 96.4 % of images by commuting average 4.7% of pixels. [21]

DongyuMeng , Hao Chen intended a framework known as MagNet. MagNet was designed to defend adversarial attack against Neural Network. MagNet does not require the information for producing the adversarial examples as well as this does not update the secured classifiers. MagNet is robust in the grey-box attack . In this attack attacker is known the defense is in place as well as the attributes of the main classifier but not the attributes of the defense. Magnet is not, knows the parameter of the base classifier, but not the parameters of the defense.MagNet is not rich to convenient adversarial examples. [22]

Chen et al. projected attack that can openly be laid in black box attack without transferability. This kind of attack does not employ gradients so directly placed in black box. The generators also altered stochastic coordinate descent (SCD) algorithmic rule by altering gradient function into an advanced loss function. The proposed system listed as ZOO-ADAM. The C&W'S white box attack results are evenly performed as compared with the ZOO. In the process of best fit solution, the system involves more iteration as compared to the white box attack. [23] Moosavi- Dezfooli et al. Formulated a universal adversarial attack grounded on the former research methods. From the experiments, authors analyzed vulnerability of deep neural networks to universal perturbations by observing the geometric correspondence between diverse parts of the decision boundary. The universal perturbation can be produced using a minor quantity of data models in its place of the whole dataset and highlighted several properties of such perturbations. [24] Jiawei Su rendered adversarial examples by simply altering single pixel to provide the solution for the measurement of perceptiveness. Their effects presented that 70.97% of images effectively tricked deep neural networks with at least one target class with confidence 97.47% on average. Experimental results show that this approach is effective on generating adversarial images in very limited conditions. [25]

III. Comparative Analysis

TABLE 1: Comparative Analysis of Adversarial image generation Methods

Ref.No.	**Attack Technique**	**Architecture**	**Dataset**	**Measurement Metric**	**Method**	**Attack Frequency**	**Applications**
[11]	Deepfool	LeNet,CaffeNet,GoogLeNet	MNIST, CIFAR10,ILSVRC 2012	ℓ p(p ϵ1,∞)	Deep neural networks	Iterative	Bioinformatics, speech, computer vision
[12]	L-BFGS	AlexNet,QuocNet	MNIST, ImageNet, Youtube	ℓ 2	Deep Neural Network	Iterative	Visual and speech recognition problem
[13]	L-BFGS-BLogistic	Torch7 to model the networks	MNIST,ImageNet	Fraction of images (in %)	logistic regression, convolution network, Over Feat	One-time	Computer Vision
[14]	FGSM	GoogLeNet, LeNet	ImageNet	Element-wise	logistic regression, softmax regression	One-time	Reinforcement Learning

[15]	FGVM	GoogLeNet ,LeNet, ResidualNet	MNIST,ImageNet, GoogLeNet	PASS	Deep Neural Network (Caffe Framework)	One-time	Computer Vision and Pattern Recognition
[16]	CPPN EA	LeNet,AlexNet	MNIST, ImageNet	N/A	convolution neural networks	Iterative	pattern-recognition
[17]	JSMA	LeNet	MNIST,	ℓ 2	Deep Neural Network	Iterative	Reading Comprehension, computer vision
[18]	BIM& ILLC	GoogLeNet	ImageNet , MNIST and CIFAR-10	N/A	neural network	Iterative	Computer vision
[19]	FGSN	GoogLeNet	MNIST, ImageNet	Element-wise	Deep Neural Network	One-time	Computer vision
[20]	RAND-FGSM	ResNet v2 Inception v3, IncRes v2	ImageNet	ℓ ∞	RBF networks	One-time	Computer Vision
[21]	C&W	GoogLeNet	MNIST, ImageNet, CIFAR-10	ℓ 0, ℓ 2,ℓ ∞	Neural Network	Iterative	Image Classification
[22]	MagNet	APE-GAN	MNIST and CIFAR-10	ℓ 2	Neural Network	*Defense for C&W	Image Classification
[23]	ZOO-ADAM	GoogLeNet	MNIST, CIFAR10 and ImageNet	ℓ 2,ℓ ∞	Deep Neural Network	Iterative	Image Classification
[24]	Universal Perturbation	GoogLeNet	ILSVRC 2012 validation set, Images captured from Camera	ℓ 2	Deep Neural Network	Iterative	Image Classification
[25]	One-pixel attacks	AlexNet	CIFAR-10, ImageNet (ILSVRC 2012)		All convolution network(AllConv), Network in network(NiN) and VGG	Iterative	Image Classification

IV. Conclusion

Many methods and theorems have been proposed and developed in recent years based on machine learning; a lot of fundamental questions need to be well explained, and a lot of challenges need to be addressed. The reason for the existence of adversarial examples is an interesting and one of the most fundamental problems for both adversaries and researchers, which exploits the vulnerability of neural networks and help defenders to resist adversarial examples. This study depicted performance of different methods like L-BFGS Attack, FGSM, BIM, JSMA, CPPN EA Fool, DeepFool, C&W's Attack, ZOO was done for generating adversarial examples. Performance of the different methods, measures was evaluated and their results were analyzed. It has been observed that for generating adversarial images require in some cases knowledge of internal model or training data and in some cases it does not. Many defenses are unable to detect adversarial examples. Most of the defenses are not effective at classifying adversarial examples correctly. Generating defenses does not mean to provide robustness we need the model which provides transferability as well as robustness.

References

Journal Papers:

[1] John C. Knight, Safety Critical Systems: Challenges and Directions, Proc. 24th Int'l Conf. Software Eng. pp. 547-550 2002.

[2] Vitaly Ford and Ambareen Siraj, Applications of Machine Learning in Cyber Security,7th International Conference on Computer Applications in Industry and Engineering, October 2014,

[3] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, arXiv preprint arXiv:1612.08242, 2016.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, In Advances in neural information processing systems, pages 91–99, 2015.

[5] Alex Krizhevsky, IlyaSutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[6] Karen Simonyan and Andrew Zisserman., Very deep convolutional networks for large-scale image recognition,arXiv preprintarXiv:1409.1556, 2014.

[7] George Saon, Hong-Kwang J Kuo, Steven Rennie, and Michael Picheny, The ibm 2015 english conversational telephone speech recognition system, arXiv preprint arXiv:1505.05899, 2015

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V Le., Sequence to sequence learning with neural networks., In Advances in neural information processing systems, pages 3104–3112, 2014.

[9] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan,OriolVinyals, Alex Graves, NalKalchbrenner, Andrew Senior, and Koray Kavukcuoglu. ,Wavenet: A generative model for raw audio. arXivpreprint, arXiv:1609.03499, 2016

[10] https://www.datascienceweekly.org/newsletters/data-science-weekly-newsletter-issue-205

[11] Seyed-Mohsen Moosavi-Dezfooli, AlhusseinFawzi, and PascalFrossard, Deepfool: a simple and accurate method to fool deep neural networks, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015, pages 2574–2582

[12] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199, 2013.

[13] Pedro Tabacof and Eduardo Valle, Exploring the space of adversarial images, In Neural Networks (IJCNN), 2016 International Joint Conference on, pages 426–433. IEEE, 2016.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572, 2014.

[15] AndrasRozsa, Ethan M Rudd, and Terrance E Boult., Adversarial diversity and hard positive generation, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 25–32, 2016.

[16] A. Nguyen, J. Yosinski, and J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images , in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, , pp. 427–436, 2015

[17] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, The limitations of deep learning in adversarial settings, in Security and Privacy (EuroS&P), IEEE European Symposium on. IEEE, , pp. 372–387 ,2016

[18] Alexey Kurakin, Ian Goodfellow, and SamyBengio, Adversarial examples in the physical world, arXiv preprint arXiv:1607.02533, 2016.

[19] Alexey Kurakin, Ian Goodfellow, and SamyBengio, Adversarial machine learning at scale, Proceedings of the International Conferenceon Learning Representations (ICLR), 2017.

[20] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel, Ensemble adversarial training: Attacks and defenses , Published as a conference paper at ICLR 2018 arXiv preprint arXiv:1705.07204, 2017.

[21] N. Carlini and D. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, AISEC, 2017

[22] D. Meng and H. Chen, MagNet: a two-pronged defense against adversarial examples., In ACM Conference on Computer and Communications Security (CCS), 2017. arXiv preprintarXiv:1705.09064

[23] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models , arXiv preprint arXiv:1708.03999, 2017.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, Universal adversarial perturbations , in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[25] J. Su, D. V. Vargas, and S. Kouichi, One pixel attack for fooling deep neural networks, arXiv preprint arXiv:1710.08864, 2018