Adversarial Attacks and Defense Strategy in Deep Learning

Sarala D.V1[a, *] Dr.Thippeswamy Gangappa2[b, **]
1[a] Asst. Prof in Dept. of CS&E, Dayananda Sagar College of Engineering,
Bangalore-560078,Karnataka ,India,
email: sarala5.dv@gmail.com
2[b] Professor in Dept. of CS&E, BMSIT & M,
Bangalore-560064,Karnataka ,India,
email: swamy.gangappa@gmail.com

Abstract: With the rapid evolution of the Internet, the application of artificial intelligence fields is more and more extensive, and the era of AI has come. At the same time, adversarial attacks in the AI field are also frequent. Therefore, the research into adversarial attack security is extremely urgent. An increasing number of researchers are working in this field. We provide a comprehensive review of the theories and methods that enable researchers to enter the field of adversarial attack. This article is according to the "Why? → What? → How?" research line for elaboration. Firstly, we explain the significance of adversarial attack. Then, we introduce the concepts, types, and hazards of adversarial attack. Finally, we review the typical attack algorithms and defense techniques in each application area. Facing the increasingly complex neural network model, this paper focuses on the fields of image, text, and malicious code and focuses on the adversarial attack classifications and methods of these three data types, so that researchers can quickly find their own type of study. At the end of this review, we also raised some discussions and open issues and compared them with other similar reviews.

Keywords:
Adversarial Attack, Defenses, Deep Learning.

I. INTRODUCTION

As many other machine learning models, neural networks are known to be vulnerable to adversarial examples adversarial examples are maliciously designed inputs to attack a target model. They have small perturbations on original inputs but can mislead the target model. Adversarial examples can be transferred across different models. This transferability enables black-box adversarial attacks without knowing the weights and structures of the target model. Black-box attacks have been shown to be feasible in real-world scenarios, which poses a potential threat to security-sensitive deep learning applications, such as identity authentication and autonomous driving. It is thus important to find effective defenses against adversarial attacks. Since adversarial examples are constructed by adding noises to original images, a natural idea is to denoise adversarial examples before sending them to the target model we explored two models for denoising adversarial examples, and found that the noise level could indeed be reduced.

II. ADVERSARIAL NETWORK

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake; they're like optical illusions for machines. In this post we'll show how adversarial examples work across different mediums, and will discuss why securing systems against them can be difficult. At Open AI, we think adversarial examples are a good aspect of security to work on because they represent a concrete problem in AI safety that can be addressed in the short term, and because fixing them is difficult enough that it requires a serious research effort. To get an idea of what adversarial examples look like, consider this demonstration from Explaining and Harnessing Adversarial Examples: starting with an image of a panda, the attacker adds a small perturbation that has been calculated to make the image be recognized as a gibbon with high confidence. An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon. The approach is quite robust; recent research has shown adversarial examples can be printed out on standard paper then photographed with a standard smart phone, and still fool systems. Adversary model Attackers can be formalized depending on their degree of knowledge, the ways in which they can tamper with the system, as well as the expected reward. For the purpose of our contribution, modeling the reward is not required. In this paper, we consider attackers that only have access to test data and, optionally, the trained model. They are thus unable to tamper with the training sample, unlike in other contexts, such as learning with

Manuscript received January 5, 2024
Manuscript revised January 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.1.14

malicious error. We address different settings, depending on the degree of knowledge of the adversary. The attacker can gain access to information about the learning algorithm, which can include only the architecture of the system or values of the parameters as well, the feature space and the data which was used for training. Of course, from the perspective of the attacker, the white-box setup is the most advantageous, making the crafting easier. A good defense method should be able to sustain the strongest type of attack achievable in practice On the other hand, it has been shown that in some cases a black-box attack, when the attacker only has access to the input and output of the model, achieves better results than its white-box counterpart. We thus consider both black-box and white-box attacks when evaluating our method.

III. DEEP LEARNING:

Deep learning is a particular kind of machine learning that achieves great power and flexibility by learning to represent the world as a nested hierarchy of concepts, with each concept defined about simpler concepts, and more abstract representations computed in terms of less abstract ones. The concept of deep learning is not new. It has been around for a couple of years now. It's on hype nowadays because earlier we did not have that much processing power and a lot of data. As in the last 20 years, the processing power increases exponentially, deep learning and machine learning came into the picture. A formal definition of deep learning is- neurons Deep learning has proven its prowess across a wide range of computer vision applications, from visual recognition to image generation. Their rapid deployment in critical systems, like medical imaging, surveillance systems, or security-sensitive applications, mandates that reliability and security are established a priori for deep learning models. Similar to any computer-based system, deep learning models can potentially be attacked using all the standard methods (such as denial of service or shooing attacks), and their protection only depends on the security measures deployed around the system. 2.3 Deep Learning, Deep convolutional networks have been long studied in computer vision. Successful results on digit recognition using supervised back-propagation networks have been achieved in early research. More recently, similar networks are applied on large benchmark datasets consisting of more than one million images, such as ImageNet with competition-winning results. The learned deep representations can be transferred across tasks. It has been extensively studied in an unsupervised setting. However, such models in convolutional networks have been limited to relatively small datasets such as CIFAR and MNIST, and only achieved modest success. Sermanet et al propose to use unsupervised pre-training, followed by supervised fine-tuning to solve the problem of insufficient training data. A supervised pre-training approach using a concept-bank paradigm is also proven successful in computer vision and multimedia settings. It learns the features on large-scale data in a supervised setting, then transfers them to different tasks with different labels. that supervised pre-training on a large dataset, followed by domain-adaptive fine-tuning on the smaller dataset is an efficient paradigm for scarce data. Additionally, DNNs are sensitive to a threat specific to prediction models. adversarial examples. These are input samples that have deliberately been modified to produce the desired response by a model. A Convolutional neural network (CNN) is a neural network that has one or more convolutional layers and is used mainly for image processing, classification, segmentation, and also for other auto correlated data.

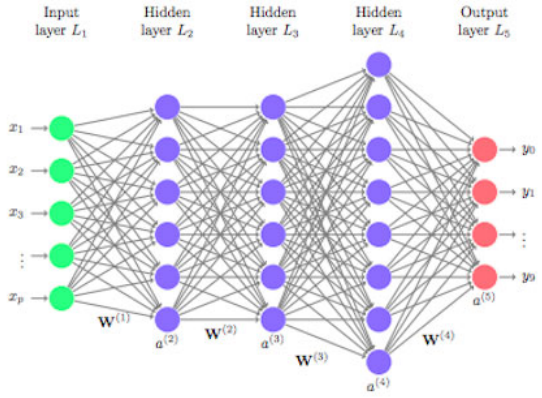

Fig.1: The **input layer, hidden layer** and **output layer** of Neural Network

A convolution is essentially sliding a filter over the input. One helpful way to think about convolutions is this quote from Dr. Prasad Samarakoon: "A convolution can be thought as "looking at a function's surroundings to make better/accurate predictions of its outcome."Rather than looking at an entire image at once to find certain features, it can be more effective to look at smaller portions of the image.

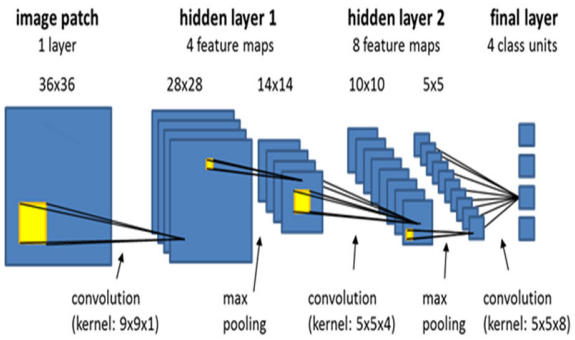

Fig 2: Convolutional Neural Network

IV. ATTACKS:

Big Data powered machine learning and deep learning has yielded impressive advances in many fields. One example is the release of ImageNet consisting of more than 15 million labeled high-resolution images of 22,000 categories which revolutionized the field of computer vision. State-of-the-art models have already achieved a 98% top-five accuracy on the ImageNet dataset, so it seems as though these models are foolproof and that nothing can go wrong. However, recent advances in adversarial training have found that this is an illusion. A good model misbehaves frequently when faced with adversarial examples. The image below illustrates the problem.

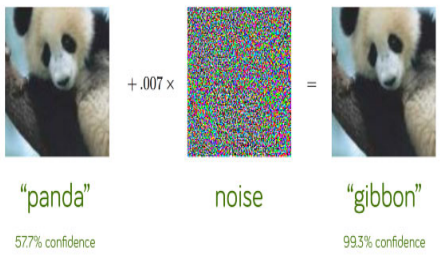

Fig 3: Noise injecting into the picture

The model initially classifies the panda picture correctly, but when some noise, imperceptible to human beings, is injected into the picture, the resulting prediction of the model is changed to another animal, gibbon, even with such a high confidence. To us, it appears as if the initial and altered images are the same, although it is radically

different to the model. This illustrates the threat these adversarial attacks we may not perceive the difference so we cannot tell an adversarial attack as happened. Hence, although the output of the model may be altered, we cannot tell if the output is correct or incorrect. This formed the motivation behind the talk for Professor Ling Liu's keynote speech at the 2019 IEEE Big Data Conference, where she touched on types of adversarial attacks, how adversarial examples are generated, and how to combat against these attacks. Without further ado, I will get into the contents of her speech. Adversarial attacks are classified into two categories — targeted attacks and untargeted attacks. The targeted attack has a target class, Y, that it wants the target model, M, to classify the image I of class X as. Hence, the goal of the targeted attack is to make M misclassify by predicting the adversarial example, I, as the intended target class Y instead of the true class X. On the other hand, the untargeted attack does not have a target class which it wants the model to classify the image as. Instead, the goal is simply to make the target model misclassify by predicting the adversarial example, I, as a class, other than the original class, X. Researchers have found that in general, although untargeted attacks are not as good as targeted attacks, they take much less time. Targeted attacks, although more successful in altering the predictions of the model, come at a cost (time).

Definition: Given data D = $\{x_1, x_2, \ldots, x_n\}$, target labels L = $\{y_1, y_2, \ldots, y_n\}$ find a hypothesis *H* such that

$$\underset{H}{\operatorname{argmin}} \sum_{x_i \in D} \ell(H(x_i), y_i)$$

where ℓ is some loss function.

V. SURVEY DETAILS

Paper no.	Short notes	Advantages	Disadvantages
[1]	Deep Convolutional Neural network is used to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes.	Minimum error rates achieved is 17.0%	Maximum error rates achieved is 37.5%
[2]	Introduces a visual sentiment concept classification method based on deep convolutional neural networks (CNNs). Nearly one million Flickr images tagged with these ANPs are downloaded to train the classifiers of the concepts.	Performance evaluation shows the newly trained deep CNNs model SentiBank 2.0 is significantly improved in both annotation accuracy and retrieval performance, compared to its predecessors which mainly use binary SVM classification models.	The adversary additionally has no knowledge of the training algorithm or hyper parameters.
[3]	Generating adversarial examples is a critical step for evaluating and improving the robustness of learning machines. So far, most existing methods only work for classification and are not designed to alter the true performance measure of the problem at hand.	The realistic deployments show an apparent boom within the obligation rankings and protection focus.	Consumes more time for massive computations.
[4]	Robustness of neural networks has recently been highlighted by the adversarial examples, i.e., inputs added with well-designed perturbations which are imperceptible to humans but can cause the network to give incorrect outputs. In this paper, we design a new CNN architecture that by itself has good robustness. Introducing a simple but powerful technique, Random Mask, to modify existing CNN structures.	Developing a very simple but effective method, Random Mask. We show that combining with Random Mask, existing CNNs can be significantly more robust while maintaining high generalization Performance.	However, it might not be appropriate to apply Random Mask to deep layers.
[5]	Deep neural networks have emerged as a widely used and means for tackling complex, real-world problems. However, a major obstacle in applying them to safety-critical systems is the great difficulty in providing formal guarantees about their behavior. presenting a novel, scalable, and efficient technique for verifying properties of deep neural networks	Speculating that the mechanism we applied to ReLUs can be applied to other piecewise linear layers, such as max-pooling layers.	While the technique is general in the sense that it is not tailored for a specific activation function

[6]	Analyzing an attack in an extremely limited scenario where only one pixel can be modified. For that proposing a novel method for generating one-pixel adversarial perturbations based on differential evolution (DE).	Creating tools that can effectively generate low cost adversarial attacks against neural networks for evaluating robustness.	showing that current DNNs are also vulnerable to such low dimension attacks. Besides
[7]	Machine learning classifiers are known to be vulnerable to inputs maliciously constructed by adversaries to force misclassification.Such adversarial examples have been extensively studied in the context of computer vision applications. In this work, we show adversarial attacks are also effective when targeting neural network policies in reinforcement learning.	Unlike supervised learning applications, where a fixed dataset of training examples is processed during learning.	The adversary additionally has no knowledge of the training algorithm or hyperparameters.
[8]	Formalizing the space of adversaries against deep neural networks (DNNs) and introduce a novel class of algorithms to craft adversarial samples based on a precise understanding of the mapping between inputs and outputs of DNNs.	Specializes identification of dangers, threats in advanced computing environments. Aimed at system high quality assurance and renovation	This transferability enables black-box adversarialattacks without knowing the weights and structures of the target model
[9]	Adversarial samples are crafted with a deliberate intention of undermining a system. In the case of DNNs, the lack of better understanding of their working has prevented the development of efficient defenses. In this paper, we propose a new defense method based on practical observations which is easy to integrate into models and performs better than state-of-the-art defenses.	The implementation of our method brings almost no overhead to the training procedure, while maintaining the prediction performance of the original model on clean samples.	Making its prediction more stable and less likely to be fooled by adversarial samples.
[10]	Different attack strategies have been proposed to generate adversarial examples, but how to produce them with high perceptual quality and more efficiently requires more research efforts. In this paper, we propose AdvGAN to generate.	Described attack has placed the first with 92.76% accuracy on a public MNIST black-box attack challenge.	In black-box attacks, dynamically train a distilled model for the black-box model and optimize the generator accordingly.
[11]	AmI (Attacks meet Interpretability) is an “attribute-steered” defense to detect adversarial examples on face recognition models. By applying interpretability techniques to a pre-trained neural network, AmI identifies “important” neurons.	AmI is no more robust to untargeted attacks than the undefended original network.	AmI rejects inputs where the original and augmented neural network disagree.

[12]	Neural networks are vulnerable to adversarial examples, which poses a threat to their application in security sensitive systems. We propose high-level representation guided denoiser (HGD) as a defense for image classification.	With HGD as a defense, the target model is more robust to either white-box or black-box adversarial attacks.	Standard denoiser suffers from the error amplification effect, in which small residual adversarial noise is progressively amplified and leads to wrong classifications.

VI. CONCLUSION

Broadly speaking, this survey paper has explored deep learning convolutional neural network CNN and adversarial behavior in deep learning systems. Most of the algorithms in survey paper can reliably produce samples correctly classified by human subjects but misclassified in specific targets by a DNN with a 97% adversarial success rate while only modifying on average 4.02% of the input features per sample. Solutions to defend DNNs against adversaries can be divided into two classes: detecting adversarial samples and improving the training phase. The detection of adversarial samples remains an open problem. Interestingly, the universal approximation theorem formulated by Hornik et al. states one hidden layer is sufficient to represent arbitrarily accurately a function [13]. Thus, one can conceive that improving training is key to resisting adversarial samples.

REFERENCES

[1] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th Conference on Neural Information Processing Systems; 2012 Dec 3–6; Lake Tahoe, NV, USA;2012. p. 1097–105.

[2] Tao Chen, Damian Borth, Trevor Darrell and Shih-Fu Chang: DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks .arXiv:1410.8586v1

[3] Cisse M, Adi Y, Neverova N, Keshet J. Houdini: fooling deep structured prediction models. 2017. arXiv:1707.05373.

[4] Luo T, Cai T, Zhang M, Chen S, Wang L. Random mask: towards robust convolutional

neural networks. In: ICLR 2019 Conference; 2019 Apr 30; New Orleans, LA, USA; 2019.

[5] Guy Katz, Clark Barrett, David Dill, Kyle Julian and Mykel Kochenderfer: Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks?. arXiv:1702.01135v2

[6] Jiawei Su*, Danilo Vasconcellos Vargas* and Kouichi Sakurai: One Pixel Attack for Fooling Deep Neural Networks. arXiv:1710.08864v7

[7] Huang S, Papernot N, Goodfellow I, Duan Y, Abbeel P. Adversarial attacks on neural network policies. 2017. arXiv:1702.02284.

[8] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy; 2016 Mar 21–24; Saarbrucken, Germany; 2016. p. 372–87.

[9] Valentina Zantedeschi, Maria-Irina Nicolae, Ambrish Rawat: E_cient Defenses Against Adversarial Attacks. arXiv:1707.06728v2

[10] Xiao C, Li B, Zhu JY, He W, Liu M, Song D. Generating adversarial examples with adversarial networks. 2018. arXiv:1801.02610.

[11] Carlini N. Is AmI (attacks meet interpretability) robust to adversarial examples? 2019. arXiv:1902.02322v1.

[12] Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 1778–87.

[13] K. Hornik, M. Stinchcombe, et al. Multilayer feedforward networks are universal approximators. Neural networks, 2(5):359–366, 1989.