Jaypee Institute of Information Technology, Noida
Dept. of Computer Science and Information Technology

Semester 6
Term Paper Report

CYBER ATTACKS ON MACHINE LEARNING MODELS: A STUDY OF ADVESARIAL VULNERABILITIES

Submitted By:

Piyush Gupta	17103067
Chitrank Mishra	17103103
Dharmesh Pratap Singh	17103279

Supervised By:
Dr. Sangeeta Mittal
Dept. of Computer Science and Information Technology

Introduction

Advances in the field of machine learning has led to revolutionizing technology in various cultures. It has also introduced capabilities that were not known before. With the advent of artificial learning expanding to support the physical world, rises the vulnerabilities that can be potential hazards to safety and security. Adversarial attacks on machine learning models are a way to exploit the learning structure of a system and create vulnerabilities which are beyond physical detection and recovery. These vulnerabilities houses capabilities from causing a classifier to misclassify, to causing trained and tested models to malfunction at run. Several algorithms have been introduced in the past few years which have happened to generate adversarial samples for detection of these anomalies. We intend to propose an ideology which check for the possible vulnerabilities which can be caused by such adversarial inputs and help detect them at the stage of training and testing.

This study is targeted at collecting various types of attacks possible on neural networks. Studies in the field have shown great advancements in the designing algorithms that hampers the raw input resulting into a misclassified objects. Researches have shown how these algorithms plays with arcade games like Atari, etc. With every neural network, there are some policies associated that parameterise the neural network. For example, for a CNN model designed to classify images, perturbations added on the training input side can cause complete fail of the trained model. There are multiple scenarios available for the study of the effect of these adversaries. Supervised learning and unsupervised learning have their own course of vulnerabilities. An adversarial model effective on one training model, is applicable on various other models as well due to property of transfer-ability in adversaries. Such vulnerabilities can target any machine model either during learning by tampering with the training data or during inference by manipulating inputs on which model is making predictions.

Neural Networks have achieved the desired state-of-the-art performance on recognizing images. It is found that these networks often suffer defeat from samples involving perturbation on samples from the datasets. Finding defense mechanisms that are effective enough and capable to protect the model from such adversarial attacks is still a vast field for research. People in the area have made a few advancements and the techiniques are growing with implementation.

In recent times, it's been determined that neural networks are fooled by adversarial examples simply. Several approaches are projected to form neural networks additional strong against white-box adversarial attacks, however they couldn't realize an efficient technique thus far. In this short paper, authors target the lustiness of the options learned by neural networks. they have a tendency to show that the options learned by neural networks aren't strong, and realize that the lustiness of the learned options is closely associated with the resistance against adversarial samples of neural networks.

Due to the complex nature of machine learning models, it is hard to identify the ways in which these models can be exploited when deployed. Recent findings on adversarial examples. which are inputs with some changes that result in different model predictions, is helpful in observing the robustness of these models by checking the adversarial situations where they fail. Although, such malicious examples are not natural as well as not applicable to complicated domains. In these paper, authors proposed a framework to make natural and reliable adversarial examples by observing in semantic space of dense and continuous data representation which is utilizing the recent findings in generative adversarial networks.

It is shown in various researches that machine learning algorithms are prone to adversarial perturbations. There are cases where physical adversaries are possible by printing malicious images and taking a picture of the same. But a major factor that hasn’t been given weightage in calculations is the physical aspects of the object. The camera can view objects from different angles and different distances. This paper shows that the current physical adversaries are not enough to create perturbations for object detection from a moving platform. It is believed that perturbed images can exhibit malicious behavior within a range of distances. Thus, the practical impact of these perturbations can be reduced when it comes to observation from a moving platform.

Identifying and collecting various types of adversarial attacks and methods to create and reciprocate their activity shall help us to understand the concepts better and help devise a state-of-the-art level defender some day.

Literature Review

Section 1: Introduction to Adversarial Attacks

[1]. Machine Learning in Adversarial Settings

Authors: Patrick McDaniel, Nicolas Papernot, and Z. Berkay Celik

Published in: IEEE Security & Privacy, May/June 2016, Cited 90 times

Summary: The paper conceptualizes the idea of how a model stores the encoded semantic information about how certain features or sets of features relate to the output class. An amount of modifications and perturbations is introduced in the data-set to yield a specific adversary-selected misclassification as output. The autonomous system can be misled into misclassifying stop signs as yield signs. To humans, these samples stay indistinguishable from the original input. Humans would classify both of these images as stop signs but the complexity for a machine to understand the image can be exploited to result in faulty classifier systems.

Methodology: The models they are focusing on are trained using supervised learning using labeled training data. This training data is a corpus sample taken from expected input distribution and labeled with their class. The training method iteratively processes the input data and creates an initial machine model. Further iterations use the input data to define the model better. This iterative refinement process strengthens the classification associations.

A problem arises when noise-injecting agents exploit the data to create adversaries. These inputs are indistinguishable to humans but the hidden structure of the images possesses the noise which hampers the quality of classification. The paper explains how the adversaries exploit the decision boundaries on a decision plane. Among the present sample set, the classification identifies a curve line separating the two classes. A small constricted data set causes the classifier to leave empty spaces in the decision plane, which homes the adversaries, which happens to change the curve and ruin the efficiency of the classifier. They explicitly drive the malicious adversaries into the input space that are ambiguous with respect to the model.

Result: With these propels, adversaries will try to bypass their controls and drive frameworks for their vindictive closures. In acknowledgment of this reality, the AI and security communities must undertake to inoculate frameworks against such abuse. Along these lines, we should return to our measures of value for AI procedures and weigh not just the results they produce yet in addition to their capacity to oppose tests cautiously produced by adversaries.

[2]. Adversarial Machine Learning at Scale.

Author: Alexey Kurakin, Ian J. Goodfellow and Samy Bengio

Published in: Conference at ICLR 2017, Cited 1077 times

Summary: Neural Networks and Machine learning models are highly vulnerable to attacks based on small modifications of the input to the model at the test time. This vulnerability possesses a transferability property. The infected input set for one machine model is also capable of infecting another machine model. Creating adversarial input requires injecting noise in the input set. The magnitude of the noise is variable according to the magnitude of the adversarial perturbation required. The robustness of such adversarially trained models increases with an increase in the model size.

Dataset: Imagenet large scale visual recognition challenge 2017. The data-set will contain 1,50,000 photographs, hand labeled into 1000 object categories, taken from Flickr and other sources.

Methodology: The detailed methodology includes adversarially training a model using synchronous distributed training on 50 machines, with a mini batch of 32 examples on each machine. They experimented with delaying adversarial training by 0, 10k, 20k, and 40k iterations. For the first N iterations in training, they used only clean examples and thereafter, included both clean and adversarial examples in the minibatch. They noticed that delaying adversarial training has almost no effect on accuracy on clean examples after a sufficient number of training iterations.

a. Results of adversarial training: Our results suggest that adversarial training should be employed in two scenarios:
    1. When a model is over-fitting, and a regularizer is required.
    2. When the concern is on security against adversarial examples. Here, any adversarial training which provides the required security is traded for losing a small amount of accuracy.
b. The property of adversarial examples lend to transfer from one model to another is a matter of security concern, allowing attacker in the black-box situation to create adversarial examples for their substitute

model, then deploy those adversarial examples to fool a target model. They studied the transferability of adversarial examples.

During each experiment, they fixed the source and target networks, created adversarial anomalies from 1000 randomly collected clean images from the test set and performed classification using the source and target networks.

Result: In this paper, they concentrated on how to expand robustness to adversarial examples of large models applied to large data-set. They showed that adversarial training gives robustness to adversarial examples generated using singular methods. While adversarial training didn't help much against iterative strategies they observed that adversarial examples generated by iterative methods are less likely to be transferred between networks, which provides indirect robustness against black-box adversarial attacks. Moreover, we saw that expansion of the model limit could likewise assist with expanding robustness to adversarial examples especially when used in conjunction with adversarial training.

[3]. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition

Author: Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter

Published in: CCS'16 October 24-28, 2016, Vienna, Austria, Cited 656 times

Summary: Machine learning has expanded its zone of action from detecting cancer-cells to operating self-driving cars. The limitless use of machine learning algorithms in various life activities where physical safety is at risk, explains well for the study of possible attacks on a machine learning model. The authors have focused on understanding the vulnerabilities of machine models working for facial biometric systems. These attacks are physically realizable and inconspicuous, and allow an attacker to use false identify or bypass the classifier by impersonating another individual. The research focuses on identifying vulnerabilities in white-box face-recognition systems, but they have also demonstrated the possible techniques for black-box scenarios to avoid face-detection.

Methodology: To demonstrate the objective of the research, the authors performed the following implementations:

a. White-box DNNs For Face Recognition: They identified three different DNNs and labeled them as A, B, and C. DNN_A refers to the globally acclaimed facial recognition model. DNNB was trained in the lab using facial features of 5 women and 5 men. DNN_C recognized a fairly larger set of people. DNN_B and DNN_C were trained over the layer of DNN_A.

b. Attacking White-box Systems: They used softmax loss to define the impersonation and dodging targets of the attacker. For impersonation, they found a modification image such that it is at least variable to the target class.

c. Facilitating physical realizability: Physical realizability can be ensured by the use of facial accessories or by tweaking the attacker's mathematical formulation.

Result: With this research, the authors were able to demonstrate the techniques for generating accessories in the form of eyeglass frames that could fool the state-of-the-art facial recognition systems. These eyeglasses can both dodge and impersonate others. Their demonstration is aimed at informing the future deliberations for defining the extent up to which the machine models can be trusted in adversarial settings

Integrated Summary for Section 1

S. No	Methods Used	Dataset	Results	Remarks
1	Pre-processing the model with available input data-set and testing the model for correct classification. Testing it for the adversarial counter data-set. Calculating the deviation from correct classification.	-Not-Used-	With these propels, adversaries will try to bypass their controls and drive frameworks for their vindictive closures. In acknowledgment of this reality, the AI and security communities must undertake to inoculate frameworks against such abuse.	This paper provides an easy insight to the concept of adversarial learning. It has an array of examples defining various scenarios where adversaries can cause damage. Good to understand the concept and know how a machine model system works.

2	Adversarially training a model using synchronous distributed training on 50 machines, with a minibatch of 32 examples on each machine.	Imagenet large scale visual recognition challenge 2017. The data-set will contain 1,50,000 photographs, hand labeled into 1000 object categories, taken from Flickr and other sources.	They showed that adversarial training gives robustness to adversarial examples generated using singular methods.	This paper aimed at showing the vulnerabilities of a faulty machine model. It also made the reader understand how the adversaries can be transferred with the learning characteristics from one model to another.
3	a. White-box DNNs For Face Recognition. b. Attacking White-box Systems. c. Facilitating physical realizability.	-Not Used-	The authors were able to demonstrate the techniques for generating accessories in the form of eyeglass frames that could fool the state-of-the-art facial recognition systems.	This paper shows the various methods which are employed to create adversarial input set. It was well enough for one to understand the concept of how adversaries are created.

Section 2: Study on Adversarial Attacks: Types

[4]. Fundamental limits on adversarial robustness

Author: Alhussein Fawzi, Omar Fawzi, Pascal Frossard

Published in: ICML 2015 Workshop on Deep Learning, Lille, France, 2015, Cited 52 Times

Summary: Paper focuses on finding if there is any difference between noise and adversarial noise. Also this focuses on finding out if there is a way to reduce or eliminate adversarial noise in Deep Learning Networks or is it the inherent part of it. This paper studies adversarial attacks and their effects on linear and quadratic classifiers in binary settings. In both the cases, paper's results showed their existence of a fundamental limit on the robustness to adversarial perturbations. It is found out that quadratic models perform better in every case and have better results then linear models.

Methodology: A framework is introduced to analyze the robustness of classifiers. Both Linear and Quadratic classifier models have been tested on adversarial perturbations and noise and the results have been plotted out on the graph of their accuracy on training and testing data.

In the paper, an example of linear and quadratic classifiers shows that the first examples(a-e) are original images on which model was supposed to be train but these images were mixed with adversarial images shown from (f-j) and the linear model gave results as (b)f_{lin} and quadratic model gave results as (c) f_{quad}, which clearly shows that quadratic classifiers are very much more robust than linear classifiers.

Result: This paper shows how the increase in the level of classification increases the robustness of the model to adversarial perturbations also to noise. In the experiments conducted in the model clearly shows that linear models can handle noise to some extent by extensive training but are unable to handle adversarial perturbations where as quadratic did a better in almost every place to reduce the effect of perturbation.

[5]. Adversarial Examples are not Bugs, they are Features.

Author: Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry

Published in: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019, Cited 280 times

Summary: This paper states that the Adversarial Examples are not bugs but actually they are the feature of the machine learning model. Machine Learning models are built in such a way that they are going to learn any feature they find common in most of the data set and that is the thing which is exploited by Adversarial Perturbations. This paper provides an alternative approach to learning by differentiating features into robust and non robust features. Basically this paper argues that we need to make the machine learning model more human-like then model oriented on what is stored in pixels.

Methodology: This paper states how to classify features of the model into robust and non robust features while training a model. It states that while doing training we need to find features which remain unchanged during both adversarial perturbations and during normal training. And then give more importance only to the feature that remains unchanged. This paper basically says that adversarial perturbations are not because of the higher dimensionality of the model but due to the very nature of the model. The paper did the research by training on robust data set and non robust data set and found that adversarial perturbations has not effect on robust data set but have a high effect on non-robust data set which proves that adversarial perturbations are not an related to training but to the choice of features.

Result: This paper proves that the previous theory which plainly blames the higher dimensionality of the data set are not completely correct and the adversarial perturbations depends highly on the choice of features. This paper also probes that adversarial perturbations can be stopped to very high extent using robust features as features to be trained on, which is basically making machine learning more human brain like and less machine like.

[6]. Universal adversarial perturbations

Author: Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard.

Published in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, Cited 951 times.

Summary: This paper tries to find out that if there is an image which can be added to any image and then that image will be misclassified by most of the classifiers. This paper proposes an algorithm to find out these kinds of images and proves that these kinds of images are possible and can be found using an algorithm. This paper also proves that universal perturbations have a remarkable property of misclassification of any image by any model.

Methodology: This paper uses the following algorithms to generate universal perturbations. This algorithm solves most m instances of the optimization problem in every pass. Even if this problem is not a convex problem if ˆk is a standard classifier, there are many efficient approximations methodology have been devised to solve this problem. We use this method for improving the efficiency.

Result: This paper proves that there exist many universal perturbations which can be applied to any image and that image will be majorly misclassified by most of the classifiers. This is a big foundation because now the prevention of such perturbations becomes more necessary as otherwise lives of many people can come in danger with the advent of driverless cars and introduction of AI in day to day life. Hence research of prevention of adversarial becomes an important topic for future research.

[7]. Poisoning Attacks against Support Vector Machines

Author: Battista Biggio, Blaine Nelson, Pavel Laskov

Published In: 29th International Conference on Machine Learning, Edinburgh, 2013, Cited 650 times

Summary: This paper described the implementation of a family of poisoning attacks against Support Vector Machines (SVM). The attack proposed in the paper uses a gradient ascent method in which properties of the SVM's optimal solution are the basis of gradient computation. Attacks on learning algorithms can be classified into exploratory (exploitation of the classifier) and causative (manipulation of training data). Poisoning refers to a causative attack (manipulation of training data) in which crafted attack points are merged into training data.

Dataset: MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: In this algorithm, arbitrary points from the attacked class are cloned and their labels are flipped to initialize the attack vector. Any point within the attacking class margin can be used as an initial point. If this point is very close to the boundaries of the attacking class, the repetitive adjusted attack point may become a reserve point, which stops further process. After each update of the attack point, the effective solution is recomputed from the solution on the training dataset with the help of the incremental support vector machine.

Result: Here, the classification error is overestimated by the validation error due to a smaller sample size. Even so, a single attack data point caused the classification error to increase from the initial error rates of 2-5% to 15-20%. This concludes that this attack can gain higher error rates than labels flipped randomly, and detects the vulnerability of the support vector machine (SVM) to poisoning attacks.

[8]. Evasion attacks against machine learning at test time

Author: Battista Biggio, Davide Maiorca, Igino Corona, Nedim Srndic,Blaine Nelson, Pavel Laskov, Giorgio Giacinto, and Fabio Roli

Published In: ECML PKDD 2013, Part III, LNAI 8190, pp. 387–402, 2013, Cited 607 times

Summary: This paper's author proposed a gradient based approach that can be used to identify the vulnerability of mainly used classification algorithms with respect to evasion attacks. Some attacking scenarios are explained which make various risk levels for the classifier by increasing the attacker's knowledge about the system and increasing the ability of the attacker to manipulate attack samples.

Dataset: PDF corpus with 500 malicious samples from the Contagio dataset and 500 gentle samples.

MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: In this paper, two experiments were conducted. First one was a toy example from the MNIST handwritten digit classification task to visually demonstrate how the proposed algorithm modies digits to mislead classification. Another is detection of malware in PDF files which shows the effectiveness of the proposed attack on a practical scenario.

Result: Researchers observed the attack in the case of perfect and limited knowledge of the attacked system, and described that widely used classification algorithms (majorly SVMs and neural networks) can escape with high probability even if the adversary can only detect a copy of the classifier from a small substitute dataset. Hence, this observation raises some questions on whether such algorithms can be reliably used in security-sensitive applications.

[9]. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods

Author: Nicholas Carlini, David Wagner

Published In: AISec'17, November 3, 2017, Cited 772 times

Summary: Carlini and Wagner proposed ten defensive techniques which detect several adversarial examples which were considered from seven papers. It is previously stated that classification of adversarial examples attempts have failed mostly, that is why the research was back on detecting only adversarial inputs. Carlini and Wagner stated that even it is quite difficult that such approaches can be defeated by a zero-knowledge attack (in which detector is not visible to the attacker) mostly. A zero-knowledge attack works against the two scenarios, that's why this attack is tried first. Perfect-knowledge attacks (white-box attack) can sometimes be adapted to the limited-knowledge situation by designing a substitute neural network and making a white-box attack against that network. Carlini and Wagner also stated that limited-knowledge attacks (black-box attack) only came into consideration if zero-knowledge attacks fail and perfect-knowledge attacks are successful.

Dataset: CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.

MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: For ten defences approached by C&W , each defense is first evaluated on the FGS (Fast-Gradient-Sign) and JSMA (Jacobian Saliency Map Attack) baselines to confirm that it was implemented correctly.

These approaches are categorised into 4 categories.

a. Secondary classification: Here authors are using a neural network to detect adversarial examples directly. Grosse 2017 added a new class to their network mainly for adversarial examples. Gong 2017 trains a separate binary classifier.
b. PCA and dimensionality reduction: Hendrycks and Gimpel 2017 used PCA(Principal Component Analysis) on the input images and identified that natural and adversarial examples have a visibly different distribution over the later components. That's why they build a classifier that focuses on these lower-order PCA components.
c. Classical statistical approaches: These approaches look at the distribution of adversarial examples. Grosse 2017 tried a Maximum Mean Discrepancy test, but this fails for zero-knowledge attacks. Feinman 2017's kernel density estimation (KDE) approach is more promising , which focuses on activations from the final hidden layer of the classifier.
d. Randomization and Blur: Feinman 2017 proposed the strongest defensive approach with Bayesian neural network uncertainty method. The basis is simply adding randomness, expecting that classification of natural images will be impacted less than adversarial inputs, which certainly seems reasonable.

Result: Zero-Knowledge Attack Evaluation: Grosse 2017 observed that 98.5% of attacks

were adversarial. It also classified half of the remaining 1.5% correctly. Gong 2017 observed that 98% of the adversarial examples are detected accurately.

Perfect-Knowledge Attack Evaluation: This paper concludes that none of these approaches are effective on MNIST. On the other hand, adversarial examples are identified with neural networks while they are not attempting to be evasive. For CIFAR-10, these attacks work in the same way and results are the same.

Limited-Knowledge Attack Evaluation: Authors only describe the attack on Grosse's scheme. They proposed two secured models implementing Grosse's defense, and applied the same attack. They detected 98% success rate, with 5.3 mean L2 distortion (4% more than the baseline). Hence, Grosse's defense is not effective and can be easily attacked even by an attacker who does not have the knowledge of the model parameters.

[10]. Adversarial vulnerability for any classifier.

Author: Alhussein Fawzi ,Hamza Fawzi, Omar Fawzi

Published In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, Cited 121 times.

Summary: Despite achieving impressive performance, state-of-the-art classifiers remain highly vulnerable to small, imperceptible, adversarial perturbations. This vulnerability has proven empirically to be very intricate to address. In this paper, we study the phenomenon of adversarial perturbations under the assumption that the data is generated with a smooth generative model. We derive fundamental upper bounds on the robustness to perturbations of any classification function, and prove the existence of adversarial perturbations that transfer well across different classifiers with small risk.

Dataset: SVHN dataset: It is a real-world image dataset for developing machine learning and object recognition algorithms. It contains 10 classes, 1 for each digit. Digit '1' has label 1, '9' has label 9 and '0' has label 10.

73257 digits for training, 26032 digits for testing, and 531131 additional, somewhat less difficult samples, to use as extra training data
CIFAR-10 dataset. This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.

Methodology: In this paper, authors measure their bounds on the SVHN dataset that contains color pictures of house numbers, and also the task is to classify the digit at the middle of the image. The dataset contains seventy three, 257 coaching pictures, and 26, 032 take a look at pictures. They train a DCGAN generative model on this dataset, with a latent vector dimension d = 100, and think about many neural networks architectures for classification. For every classifier, the empirical lustiness is compared to our boundary. additionally to news the in-distribution and at liberty lustiness, additionally report the lustiness within the latent space:

$$r_Z = \min_r \|r\|_2 \text{ s.t. } f(g(z+r)) \neq f(g(z)).$$

Result: Experiments on SVHN dataset. Authors report 25 % of the normalized lustiness at every cell, wherever chances are squared, measured either on paper.

	Upper bound on robustness	2-Layer LeNet	ResNet-18	ResNet-101
Error rate	-	11%	4.8%	4.2 %
Robustness in the $\mathcal{Z}$-space	16×10^{-3}	6.1×10^{-3}	6.1×10^{-3}	6.6×10^{-3}
In-distribution robustness	36×10^{-2}	3.3×10^{-2}	3.1×10^{-2}	3.1×10^{-2}
Unconstrained robustness	36×10^{-2}	0.39×10^{-2}	1.1×10^{-2}	1.4×10^{-2}

[11]. Generating Natural Adversarial Examples.

Author: Zhengli Zhao,Dheeru Dua ,Sameer Singh.

Published In: ICLR 2018, Cited 230 times.

Summary: Due to the complex nature of machine learning models, it is hard to identify the ways in which these models can be exploited when deployed. Recent findings on adversarial examples. which are inputs with some changes that result in different model predictions, is helpful in observing the robustness of these models by checking the adversarial situations where they fail. Although, such malicious examples are not natural as well as not applicable to complicated domains. In this paper, authors proposed a framework to make natural and reliable adversarial examples by observing in semantic space of dense and continuous data representation which is utilizing the recent findings in generative adversarial networks.

Dataset: MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.
LSUN dataset in which they randomly sample the amount of 126,227 images from the "Tower" category.

Methodology: Authors apply their approach to two standard datasets, MNIST and LSUN, and generate natural adversaries. They use $\Delta r = 0.01$ and $N = 5000$ with model details.Two different applications are described in paper are as follows:

Handwritten Digits: Authors train a W-GAN on 60,000 MNIST images.They included the inverter and with fully connected layers on top of the last hidden layer. They train two classifiers to generate adversaries against Random Forests which have 5 trees with test accuracy of 90.45% and LeNet with 98.71% test accuracy.

Church vs Tower : Training procedure is same as above except that the generator and critic in WGAN are deep residual networks. Authors train an MLP classifier on these two classes with test accuracy of 71.3%.

Result: For MNIST's hand-written digits , author picked up 20 images, 2 for each digit and generated adversaries against RF and LeNet then observed 13 responses for each of the questions. They also checked adversaries for the LeNet model generated by FGSM and found that 78% of the time the program agrees that adversaries changed to the original images and are more natural.

[12]. Learning More Robust Features with Adversarial

Author: Shuangtao Li , Yuanke Chen , Yanlin Peng , Lin Bai

Published In: AISec'18, September 12, 2018, Cited by 7.

Summary: In recent times, it's been determined that neural networks are fooled by adversarial examples simply. Several approaches are projected to form neural networks additional strong against white-box adversarial attacks, however they couldn't realize an efficient technique thus far. In this short paper, authors target the lustiness of the options learned by neural networks. they have a tendency to show that the options learned by neural networks aren't strong, and realize that the lustiness of the learned options is closely associated with the resistance against adversarial samples of neural networks. They have a tendency to conjointly realize that adversarial coaching against quick gradients sign technique (FGSM) doesn't build the learned options terribly strong, notwithstanding it will build the trained networks terribly proof against FGSM attack

Dataset: CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.

MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: In order to create the options learned by neural networks that are additional sturdy, authors tend to add a distortion term to the initial adversarial objective performance to encourage the distortions to be smaller throughout coaching. Formally, they tend to train neural networks with this objective function:

$$\tilde{J}_{adv}(\theta, x, y) = \alpha \cdot J(\theta, x, y) + (1-\alpha) \cdot J(\theta, x^*, y) + \sum_{i=1}^{L-1} \beta_i \cdot \sum_{j=1}^{h_i} d_{ij}$$

where L - 1 is that the variety of the social control layers of the trained network, hi is that the variety of options within the ith layer, dij is that the distortion of the worth of jth normalized feature within the ith layer, βi may be a constant for equalisation each term. The opposer for adversarial coaching will be any opposer, e.g. the universal first-order opposer PGD attack, but authors have a tendency to use FGSM as our opposer during this paper for process potency.

Result: Accuracy that the trained networks achieve on clean test data and adversarial test data.

Models	MNIST			CIFAR-10		
	Clean	FGSM	PGD	Clean	FGSM	PGD
Standard	0.9939	0.0922	0	0.9306	0.5524	0.0256
Adversarially trained	0.9932	0.9492	0.0612	0.8755	0.8526	0.1043
Our method	0.9903	0.9713	0.9171	0.8714	0.6514	0.3440

[13]. Adversarial Examples Are a Natural Consequence of Test Error in Noise

Author: Nicolas Ford, Justin Gilmer, Nicholas Carlini, Ekin D. Cubuk

Published In: Proceedings of the 36th International Conference on Machine Learning, 2019, Cited 72 times.

Summary: This paper shows that adversarial examples are just a natural consequence of test error in noise. And they should not be taken as bugs. Finally, this paper shows that methods which are going to increase the distance to the decision boundary will also improve robustness towards Gaussian noise, and vice versa. Author states that, given the error rates it is observed in Gaussian noise, small perturbations it is observed in practice appear that roughly the distances would be expected from a linear model, and that therefore there is not much need for invoking any properties of the decision boundary to explain them.

Methodology: Author started by looking into the relation between adversarial and corruption robustness in that case where q consists in the images with Gaussian noise which is additive. For linear models, the rate of

error in the Gaussian noise is going to exactly determine the distance between the decision boundary.Then author compared Neural Networks to the Linear Case. The decision boundary in Deep Learning model is not linear.

Result: This paper finally tries to answer whether we should be focused to find adversarial examples as close as we are currently focusing on, given that the error rates we have observed in the corrupted image distributions. After going through several experiments, the author shows that the answer is no, for the following reasons:

1. The nearby errors which occur show up at the same distance the author expects from a linear model with same robustness.
2. Measure concentration shows that non-zero error rate in Gaussian noise logically implies that the existence of very small perturbations of noisy images.

[14]. Are adversarial examples inevitable

Author: Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, Tom Goldstein

Published In: ICLR 2019 Conference, 2019, Cited 117 times.

Summary : This paper tries to find that if it is possible or not to prevent adversarial perturbations. The author says that the question that if adversarial perturbations are inevitable is wrong.And any model has a limit on correctness to adversarial perturbations that cannot be removed. But, paper proves that these limits depend on fundamentals of the dataset, and also on the power of the adversary and the metric system used to measure different kinds of perturbations. This paper provides great details of these limits and shows us how they are inter-dependent on properties of the distribution of data.

Methodology: The author work the unit cube model to minimize the perturbations to zero but then he failed and said that we cannot completely discard adversarial perturbations.

The idea he used is to show that, if the given class of data takes up enough space,then nearly every unique data point in the class will lie close to the boundary of the class.

Then the next question arises about the sparse adversarial examples. To study this, he investigated adversarial examples under the 0 metric. After much of the work he says that we cannot take it to 0. But all we can do is to reduce the risk of such attacks.

Tighter bounds can only be obtained if there is a guarantee that the adversarial examples that exist for some unique data points in any class, without bounding the probability of the event.

Result: So this paper shows in great detail that it is not possible to prevent adversarial perturbations completely by using any method available. This paper also shows that the adversarial perturbations are the fundamental property of machine learning and to some extent they are going to affect the model.This paper also shows that limit to adversarial perturbation depends on the power of adversary and the system of metrics which is being used to measure different things.

[15]. A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance

Author: Adi Shamir , Itay Safran , Eyal Ronen and Orr Dunkelman.

Published In: 2019, Cited 255 times.

Summary: The paper tries to prove that there exists a Small hamming distance for perturbing any image. In the research made earlier to explain the existence of perturbations they are using a Deep Learning model and an input X whose class is given by the model as belonging to some class C1, and they wanted to find some Y with distance(X,Y) as less as possible which is classified as belonging to some other class C2. In this paper the authorconsidered a better way of attacking, in which the author is taking two class D1 and D2,along with an input X C1, and their goal is to search for some nearby Y which is inside C_2.

Dataset: MINST Dataset

Methodology: Authors used MNIST dataset, where their algorithm failed and did not find any example with Hamming distance of less than or equal to 10 , but what they found is a group of 11 out of the 784 pixels which on manipulating could change the prediction from one digit to other digit.

Algorithm:

Algorithm 1 Basic Algorithm

1: **Input:** $\mathbf{x} \in C_1$, target C_2, neural network NN.
2: Compute $\mathbf{x}' = NN(\mathbf{x})$.
3: Pick an arbitrary $\mathbf{y} \in C_2$, and compute $\mathbf{y}' = NN(\mathbf{y})$ in the output space.
4: Connect $\mathbf{x}$ and $\mathbf{y}$ in a straight line in the input space as defined by $\mathbf{x}_a$.
5: Map the line (using NN) into a piecewise linear line between $\mathbf{x}'$ and $\mathbf{y}'$, denoted by *path*.
6: Set $tmp \leftarrow \mathbf{x}$ and $tmp' \leftarrow \mathbf{x}'$.
7: Choose an arbitrary subset of m (out of n) input variables.
8: **repeat**
9: Describe the linear map at the vicinity of tmp to tmp' as an $m \times m$ matrix M'
10: Find the direction in the reduced input space using M'^{-1} that follows *path*.
11: Advance tmp and tmp' in the direction of the path until a ReLU boundary is found, *path* changes direction, or $\mathbf{y}'$ is reached.
12: **until** $tmp' = \mathbf{y}'$

Result: In this paper authors had developed a new and innovative method to rethink about the adversarial examples, and authors had explained why we find in our neural network adversarial perturbations which contains a Hamming distance of m+1 in Deep Learning models which are used to distinguish between a m number of classes.

Integrated Summary for Section 2

S. No	Methods Used	Dataset	Results	Remarks
4	Linear and Quadratic classifier models have been tested on adversarial perturbations and noise and the results have been plotted out on the graph of their accuracy on training and testing data	-Not Used-	This paper shows how the increase in the level of classification increases the robustness of the model to adversarial perturbations also to noise.	This paper shows how increasing the dimensionality of a system makes it more prone to adversarial perturbations. This paper also shows that system robustness decreases with dimensionality hence perturbations are different from noise.
5	Classify features of the model into robust and non robust features while training a model.	-Not Used-	The previous theory which plainly blames The higher dimensionality of the data set are not completely correct and the adversarial perturbations depends highly on the choice of features.	This paper shows that our thinking about adversarial perturbations is wrong and we should not consider them as bugs but we should think of them as features of a Machine Learning algorithm. This paper states that we need to change our way of machine learning by differentiating features into robust and non robust feature and make the process more human like and less machine like.
6	Find out that if there is an image which can be added to any image and then that image will be misclassified by most of the classifiers.	-Not Used-	Proved that there exist many universal perturbations which can be applied to any image and that image will be majorly misclassified by most of the classifiers.	This paper shows that it is possible to generate an image which when dot produced with any image in the world has a very high change of showing perturbated results by most of the models in the world.
7	The attack proposed in the paper uses a gradient ascent method in which properties of the SVM's optimal solution are the basis of gradient computation.	MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0"	The classification error is overestimated by the validation error due to a smaller sample size. This concludes that this attack can gain higher error rates than	The idea of vulnerability of SVMs has come into view from this paper. And poisoning attacks can easily exploit the working of SVMs.

		through "9". There are about 6K training examples of every digit and 1Ktest examples of every digit.	labels flipped randomly, and detects the vulnerability of the support vector machine (SVM) to poisoning attacks.	
8	Two experiments were conducted: a. A toy example from the MNIST handwritten digit classification task. b. Detection of malware in PDF files which shows the effectiveness of the proposed attack.	PDF corpus with 500 malicious samples from the Contagio dataset and 500 gentle samples.	The attack in the case of perfect and limited knowledge of the attacked system, and described that widely used classification algorithms (majorly SVMs and neural networks) can escape with high probability even if the adversary can only detect a copy of the classifier from a small substitute dataset.	Widely used neural networks can be attacked with only little knowledge about the classifiers. So, this is obviously a matter of concern for organizations where such networks are used for various purposes.
9	Approaches are categorised into 4 categories. a. Secondary classification. b. PCA and dimensionality reduction. c. Classical statistical approaches. d. Randomization and Blur.	CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.	Zero-Knowledge Attack Evaluation: Grosse 2017 observed that 98.5% of attacks were adversarial. Perfect-Knowledge Attack Evaluation: none of these approaches are effective on MNIST. Limited-Knowledge Attack Evaluation: Grosse's defense is not effective and can be easily attacked even by an attacker who does not have the knowledge of the model parameters.	Achieving a higher accuracy is useful and interesting result in machine learning tasks but this is not secure or sufficient for secure machine learning. We should consider the attackers mindset like if they even knew about the defense work still defense remains secure.
10	They train a DCGAN generative model on this dataset, with a latent vector dimension d = 100, and think about many neural networks architectures for classification. For every classifier, the empirical lustiness is compared to our boundary. additionally to news the in-distribution and at liberty lustiness, additionally report the lustiness within the latent space	SVHN dataset	Experiments on SVHN dataset. Authors report 25 % of the normalized lustiness at every cell, wherever chances are squared, measured either on paper.	We derive fundamental upper bounds on the robustness to perturbations of any classification function, and prove the existence of adversarial perturbations that transfer well across different classifiers with small risk.
11	Authors apply their approach to two standard datasets, MNIST and	MNIST dataset, LSUN dataset	For MNIST's hand-written digits , author picked up 20	Such malicious examples are not natural as well as not

	LSUN, and generate natural adversaries. They use Δr = 0.01 and N = 5000 with model details.		images, 2 for each digit and generated adversaries against RF and LeNet then observed 13 responses for each of the questions. They also checked adversaries for the LeNet model generated by FGSM and found that 78% of the time the program agrees that adversaries changed to the original images and are more natural.	applicable to complicated domains. In this paper, authors proposed a framework to make natural and reliable adversarial examples by observing in semantic space of dense and continuous data representation which is utilizing the recent findings in generative adversarial networks.
12	To create the options learned by neural networks that are additional sturdy, authors tend to add a distortion term to the initial adversarial objective performance to encourage the distortions to be smaller throughout coaching. Formally, they tend to train neural networks with this objective function	CIFAR-10 Dataset, MNIST Dataset	Accuracy that the trained networks achieve on clean test data and adversarial test data.	They have a tendency to conjointly realize that adversarial coaching against quick gradients sign technique (FGSM) doesn't build the learned options terribly strong, notwithstanding it will build the trained networks terribly proof against FGSM attack
13	For linear models, the rate of error in the Gaussian noise is going to exactly determine the distance between the decision boundary.Then author compared Neural Networks to the Linear Case. The decision boundary in Deep Learning model is not linear.	-Not Used-	This paper finally tries to answer whether we should be focused to find adversarial examples as close as we are currently focusing on, given that the error rates we have observed in the corrupted image distributions.	For given error rates it is observed in Gaussian noise, small perturbations it is observed in practice appear that roughly the distances would be expected from a linear model, and that therefore there is not much need for invoking any properties of the decision boundary to explain them.
14	The idea he used is to show that, if the given class of data takes up enough space,then nearly every unique data point in the class will lie close to the boundary of the class.	-Not Used-	This paper shows in great detail that it is not possible to prevent adversarial perturbations completely by using any method available. This paper also shows that the adversarial perturbations are the fundamental property of machine learning and to some extent they are going to affect	Paper proves that these limits depend on fundamentals of the dataset, and also on the power of the adversary and the metric system used to measure different kinds of perturbations.

			the model.	
15	Authors used MNIST dataset, where their algorithm failed and did not find any example with Hamming distance of less than or equal to 10 , but what they found is a group of 11 out of the 784 pixels which on manipulating could change the prediction from one digit to other digit.	MINST Dataset	In this paper authors had developed a new and innovative method to rethink about the adversarial examples, and authors had explained why we find in our neural network adversarial perturbations which contains a Hamming distance of m+1 in Deep Learning models which are used to distinguish between a m number of classes.	In this paper the author considered a better way of attacking, in which the author is taking two class D1 and D2,along with an input X C1, and their goal is to search for some nearby Y which is inside C_2

Section 3: Defense against Adversarial Attacks

[16]. Standard detectors aren't (currently) fooled by physical adversarial stop signs

Author: Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsy

Published in: University of Illinois at Urbana Champai, 9 Oct 2017, Cited 36 times

Summary: Adversarial examples that exist can be used to fool a detector and create unusual and uncontrollable situations. One such example is the physical adversarial stop sign which is known to fool a large group of classifiers and detectors, but then comes RCNN and YOLO, which was able to be classified as a non-stop sign. An adversarial pattern on a physical object can be detected using a wide family of parameters such as scale, view of angle, etc. Such a pattern is found shall be of great practical and theoretical use. It is difficult to diagnose a misclassifier as compared to a mis-detector unless we get to eliminate the effects of rescaling and resizing.

Dataset: Random videos from youtube having a car driving by a stop sign.

Methodology: The demonstration aims at finding the difficulties observed while classifying and detecting stop sign in moving video using RCNN and YOLO algorithms. Random videos were taken from youtube as a sample set. Various parameters such as brightness, angle of view, blurriness were kept in mind while selecting the videos. Frames from the video were feed into the system.

RCNN and YOLO were applied on the frames and it was found that both the algorithms were able to classify the sign as stop sign well, even after taking adversaries into account. RCNN performed faster compared to YOLO in this context.

To check for detection, relatively low-quality resolution images were feed into the system frame by frame from random video from youtube. RCNN predicts using boxes and then classifies them. YOLO used a grid of cells, where each cell used features computed from much of the image to predict boxes and create labels near the cell. YOLO detects stop sign better compared to RCNN in a moving video window.

Result: They do not claim that detectors are necessarily immune to physical adversarial examples. It can be said that there is no physical anomaly found yet that can fool a detector. An adversarial pattern to fool a detector has to be adversarial in many aspects such as scale, view of angle, illumination, etc. Finding such a pattern is a technical challenge in itself.

[17]. Adversarial Examples: Attacks and Defenses for Deep Learning

Author: Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li

Published In: IEEE Transactions on Neural Networks and Learning Systems, 2019. Cited 564 times.

Summary: As rapid progress in a wide spectrum of applications, many safety-critical

applications use deep learning. But, many vulnerabilities have been found in deep neural networks to adversarial examples which are well designed input samples. These types of inputs are not identified by humans but deep neural networks can be fooled easily by these examples. So, this becomes a major issue in a safety-critical environment. In this paper, authors observe some recent theories on adversarial examples for deep neural networks and summarize some attacks of adversarial examples and taxonomy of these examples.

Dataset: CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.

MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

ImageNet dataset consists of 14,197,122 images with 1,000 classes. Most adversarial approaches are evaluated on this dataset because of the large number of images in the ImageNet dataset.

Methodology: Several attacks and defenses on deep learning are explained in the paper. Here I will be discussing attacks only. Some defenses are already discussed in other papers i.e; L-BFGS, Fast Gradient Sign Method , C&W's attack , Universal Perbutations. So, I am going to discuss one more attack here i.e; One Pixel Attack. Su et al. made adversarial examples by changing one pixel to avoid the problem of perceptiveness measurement. The optimization problem becomes:

$$\begin{aligned} \min_{x'} \quad & J(f(x'), l') \\ s.t. \quad & \|\eta\|_0 \leq \epsilon_0, \end{aligned}$$

where E0 = 1 represents one pixel modification. Problems become hard to optimize due to new constraints. CIFAR-10 dataset is used on 3 neural networks : Network in Network, all convolution network , and VGG16. Images that are successfully fooled deep neural networks are 70.97% with confidence of 97.47%.

Result: In this paper, authors observed some findings of adversarial examples in deep neural networks. They checked existing methods for generating adversarial examples. Authors tried to cover study of state-of-the-art for adversarial examples in the deep learning domain.

[18]. Evaluating a Simple Retraining Strategy as a Defense Against Adversarial Attacks

Author: Nupur Thakur, Yuzhen Ding , Baoxin Li

Published In: Arxiv.org, 20 July 2020

Summary: Neural networks are found to be vulnerable on adversarial examples, such inputs which are close to natural inputs but classified wrongly. For better understanding the adversarial examples, authors observed ten recent findings which are designed to detect adversarial examples.They show that all of those can be defeated by making new loss functions. In this paper, authors describe neural networks applied to image classification. As neural networks are the mostly accurate machine learning approach known till now, they are fighting against an adversary who can fool the classifier. For that , a natural image x is given, an adversary produces a visually same image x easily which will be classified differently. But , most of these defences failed to classify adversarial examples correctly.

Dataset: CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.

MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: Many attacking methods are proposed in the paper. So, for being short I am gonna discuss one attacking method here.

Carlini-Wagner Attack: Authors use the L2 attack for our experiments as a result of it's thought-about to be the strongest among the 3 attacks. For each of the datasets, the target label is the label of the smallest amount of probable category. For CIFAR-10 dataset, the initial casual magnitude relation is 100%. Using the whole original information together with 50k adversarial pictures for coaching and 10k adversarial take a look at pictures, take a look at accuracy of 90.8% is achieved on the adversarial pictures. Therefore, the network has learnt the pictures all right. when offensive the retrained classifier, the casual magnitude relation continues to be ninety nine. ten so, the network will be fooled even when learning the adversarial pictures properly.

Result: Retraining the network by the adversarial pictures generated by the Carlini-Wagner rule for CIFAR-10 and TinyImageNet Dataset. The quantity of adversarial pictures used for training is the same because the number of original training pictures.

Dataset	Test Accuracy on Original Images	Test Accuracy on Adversarial Images	Fooling Ratio after Retraining
CIFAR-10	92.37%	90.80%	99%
TinyImageNet	72.11%	70%	99%

[19]. Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser.

Author: Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, Jun Zhu

Published in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, Cited 279 times

Summary: Neural networks are highly sensitive to adversarial examples are therefore poses a threat towards security application. This study proposes a high-level representation guided denoiser (HGD) as a defense towards adversarial image classification. Standard denoiser face problems of error amplification effect, in which small residual adversarial noise is progressively amplified and leads to wrong classifications. Using a loss function, HGD overcomes this problem. The function defines a difference between the target model's outputs activated by the clean image and denoised image.

On comparing with the state-of-the-art classifier, HGD has few advantages over it. The target model is more robust to either white-box or black-box attacks with HGD as a defense. HGD can be trained with a few image sets to perform well on other classes. HGD can transform from guiding a model to defending it when needed.

HGD won the first place in NIPS competition on defense against adversarial attacks and also outperformed other models by a huge margin.

Dataset: 30K images from the ImageNet training set (30 images per class).

Methodology: None of the models are so efficient to completely defend against adversarial inputs and hence a small residual perturbation is amplified to a large magintude in top layers of classification of the target model.

This effect is well understood as the error amplification effect, which leads to wrong prediction. To solve this problem, the authors have tried to implements a loss function on the basis of difference between the top-level output of the target model induced by original and adversarial examples, rather than implementing the original state-of-the-art system which uses a pixel-level reconstruction as standard denoiser.

They introduced a pixel guided denoiser which is mapped to work with the Imagenet dataset. A potential problem with this pixel guided denoiser is the amplification effect of adversarial noise in the topmost layers. HGD overcome this problem, where the supervised signal comes from certain high-level layers of the target model. HGD uses the same U-net structure as DUNET. The activities of this layer are feed to the linear classification layer after the global average pooling. The loss function used here is a perceptual loss or feature matching loss. Another set of HGD uses the index of layer before the final softmax function. Here , the loss function is the difference between the two logits activated by the adversarial image and the denoised image.

They considered both the HGD structures for the denoiser as the convolutional feature maps provide richer supervised information, while the logits represents the classification results. Another alternative for the loss function is to use the classification loss of the target model as denoising loss function, which is supervised learning.

Result: From the study it is found that DUNET has much lower denoising loss than DAE and NA which represents structural advantage of DUNET. DAE does not perform well with encoding of high-resolution images and hence the accuracy drops significantly. For white-box attacks, DUNET has much lower denoising loss than DAE but the classification accuracy is significantly worse.

They also discovered that error amplification effect of adversarial examples and proposed it to guide the training of a denoiser by using the error in the top layers of neural networks as loss functions. Also the denoising ability of HGD depends on the representability of the training set and there is scope for improvement in HGD by incorporating other different attacks.

[20]. APE-GAN: Adversarial Perturbation Elimination with GAN

Author: Shiwei Shen, Guoqing Jin, Ke Gao, Yongdong Zhang

Published in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ,Cited 65 times

Summary: Neural Networks have achieved the desired state-of-the-art performance on recognizing images. It is found that these networks often suffer defeat from samples involving perturbation on samples from the datasets. Finding defense mechanisms that are effective enough and capable to protect the model from such adversarial attacks is still a vast field for research. People in the area have made a few advancements and the techiniques are growing with implementation. This study proposes an idea based on Generative

Adversarial Networks named APE-GAN is targeted to defend against these adversarial examples. An experimental study is also conducted to find out the efficiency of the implementation on MNIST, CIFAR-10 and ImageNet indicate that APE-GAN is effective to resist adversarial examples.

Dataset: MNIST, CIFAR-10, ImageNet

Methodology: Misclassification of images happen due to intentionally imperceptible perturbations to some parts of the images or precisely some pixels of the images. The study proposes an algorithm to apply defense against adversarial examples and eliminate the adversarial perturbation from the input set. GAN or Generative adversarial network proposed by Goodfellow et al is able to generate images that are similar to the training set with an addition of a little noise. Therefore, the authors have used GAN to reverse generate clean images from noised images.

They are generating adversarial examples from 5 algorithms namely: L-BFGS Attack, Fast Gradient Sign Method Attack (FGSM), DeepFool Attack, Iterative Gradient Sign, Jacobian-based Saliency Map Attack. The fundamental idea is to eliminate the trivial perturbation from the input before sending them for recognition by the model.

The general idea behind this method is to allow one to train a generative model with the aim of deceiving a differentiable discriminator that is targeted to classify reconstructed images from original images. Hence, the model can be used to produce revamped images that are highly similar to original clean images. Basically, it is like adding noise to an already noised data to lower the effect of the latter and hence produce images that are near to clean images.

To differentiate clean images from reconstructed images, they devised a discriminator network. The discriminator network solves the maximization problem and also contains some convolutional layers to get some high-level feature maps. They used 4 losses to define their loss function namely: Discriminator loss, Generator loss, Content loss, Adversarial loss.

Result: The error rates of adversarial inputs are significantly decreased after its perturbation is reduced by APE-GAN. The error rate of FGSM is much larger as compared to L-BFGS. The aggressivity of adversarial examples can be eliminated by APE-GAN so is the perturbation whether regular or irregular, can also be eliminated. The combination of various defenses give rise to a much better and effective defense mechanism. Such combinations are a good field for research in near future for further study.

[21]. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Author: Kimin Lee, Kibok Lee, Honglak Lee, Jinwoo Shin

Published in: Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018 ,Cited 239 times

Summary: Identifying test samples for image data which sufficiently diverse when compared with the training distribution statistically or adversarially is a basic requirement for deploying a good classification model. Deep neural networks are capable of producing methods to detect any abnormal samples which are applicable to all the softmax classifiers. Most prior methods have been reported for detecting either out-of-distribution or adversarial samples, but not both, the proposed methods achieves state of the art performances for both cases in various experiments conducted. The proposed methods is more robust in certain tough scenarios. It is shown that the proposed idea enjoys broader application by applying it to class-incremental learning. That signifies whenever out-of-distribution samples are detected, the method is able to create new classification classes without further training.

Dataset: CIFAR-10, ImageNet, ResNet

Methodology: The idea is to measure the probability density of test sample on the spaces of features of DNNs utilizing the concept of a generative (distance-based) classifier. Contrary to the conventional beliefs, they found that using a generative classifier does not hampers the softmax accuracy. On the other hand, it's confidence score outperforms softmax-based ones very easily on various specified tasks.

They demonstrated the effectiveness of the idea by implementing it using deep convolutional neural networks such as resnet, trained as image classifier on multiple datasets. For detecting out-of-distribution samples, the proposed idea outruns the state-of-the-art implementation. For detecting the adversarial samples, generated using four of the attack algorithms that are FGSM, BIM, DeepFool and CW algorithms, the idea outruns the state-of-the-art implementation measurements.

They found that the proposed method is more robust when it comes to making choice of its hyper-parameters and also against extreme scenarios. e.g. when training have some noise or distortion or random labels etc. Hyperparameters of the proposed method can be tuned only using in-distribution samples maintaining the performance alongside.

Then they applied the method to class-incremental learning. It means that new classes are added progressively to a pre-trained classifier without any retraining. The new classes are obtained from the out-of-training distribution. It is naturally expected that one can classify them using proposed metric.

Result: They proposed a simple yet effective method for detecting abnormal test samples including both out-of-distribution and adversarial ones. The main idea was to induce a generative classifier and define new confidence scores based on it. They believe that the approach has the potential to apply to many other related machine models and learning tasks.

Integrated Summary for Section 3

S. No	Methods Used	Dataset	Results	Remarks
16	Finding the difficulties observed while classifying and detecting stop sign in moving video using RCNN and YOLO algorithms.	Random videos from youtube having a car driving by a stop sign.	It can be said that there is no physical anomaly found yet that can fool a detector. An adversarial pattern to fool a detector has to be adversarial in many aspects such as scale, view of angle, illumination, etc.	This paper aimed at making the reader understand the preventive measures against the faulty machine model, if one is. It has made clear points about the factors like distance, angle and illumination which can be made use of to prevent faulty classification.
17	One Pixel Attack. Su et al. made adversarial examples by changing one pixel to avoid the problem of perceptiveness measurement.	CIFAR-10 dataset, MNIST dataset, ImageNet	They checked existing methods for generating adversarial examples. Authors tried to cover study of state-of-the-art for adversarial examples in the deep learning domain.	In this paper, authors observed some findings of adversarial examples in deep neural networks.
18	Authors use the L2 attack for our experiments as a result of it's thought-about to be the strongest among the 3 attacks. For each of the datasets, the target label is the label of the smallest amount of probable category.	CIFAR-10 dataset, MNIST dataset	Retraining the network by the adversarial pictures generated by the Carlini-Wagner rule for CIFAR-10 and TinyImageNet Dataset. The quantity of adversarial pictures used for training is the same because the number of original training pictures.	They show that all of those can be defeated by making new loss functions. In this paper, authors describe neural networks applied to image classification.
19	They introduced a pixel guided denoiser which is mapped to work with the Imagenet dataset. A potential problem with this pixel guided denoiser is the amplification effect of adversarial noise in the topmost layers. HGD overcome this problem,	30K images from the ImageNet training set	From the study it is found that DUNET has much lower denoising loss than DAE and NA which represents structural advantage of DUNET. DAE does not perform well with encoding of high-resolution images	HGD won the first place in NIPS competition on defense against adversarial attacks and also outperformed other models by a huge margin.

	where the supervised signal comes from certain high-level layers of the target model. HGD uses the same U-net structure as DUNET. The activities of this layer are feed to the linear classification layer after the global average pooling.		and hence the accuracy drops significantly. For white-box attacks, DUNET has much lower denoising loss than DAE but the classification accuracy is significantly worse.	
20	The study proposes an algorithm to apply defense against adversarial examples and eliminate the adversarial perturbation from the input set. GAN or Generative adversarial network proposed by Goodfellow et al is able to generate images that are similar to the training set with an addition of a little noise.	MNIST, CIFAR-10, ImageNet	The error rates of adversarial inputs are significantly decreased after its perturbation is reduced by APE-GAN. The error rate of FGSM is much larger as compared to L-BFGS. The aggressivity of adversarial examples can be eliminated by APE-GAN so is the perturbation whether regular or irregular, can also be eliminated.	This study proposes an idea based on Generative Adversarial Networks named APE-GAN is targeted to defend against these adversarial examples. An experimental study is also conducted to find out the efficiency of the implementation on MNIST, CIFAR-10 and ImageNet indicate that APE-GAN is effective to resist adversarial examples.
21	The idea is to measure the probability density of test sample on the spaces of features of DNNs utilizing the concept of a generative (distance-based) classifier. Contrary to the conventional beliefs, they found that using a generative classifier does not hampers the softmax accuracy. On the other hand, it's confidence score outperforms softmax-based ones very easily on various specified tasks.	CIFAR-10, ImageNet, ResNet	They proposed a simple yet effective method for detecting abnormal test samples including both out-of-distribution and adversarial ones. The main idea was to induce a generative classifier and define new confidence scores based on it. They believe that the approach has the potential to apply to many other related machine models and learning tasks.	The proposed methods is more robust in certain tough scenarios. It is shown that the proposed idea enjoys broader application by applying it to class-incremental learning. That signifies whenever out-of-distribution samples are detected, the method is able to create new classification classes without further training.

Section 4: Application Of Adversarial Attacks

[22]. No Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles

Author: Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth

Published in: University of Illinois at Urbana Champaign, Jul 2017, Cited 142 times

Dataset: 180 photographs of stop sign at a highway, from various angles and distances.

Summary: It is shown in various researches that machine learning algorithms are prone to adversarial perturbations. There are cases where physical adversaries are possible by printing malicious images and taking a picture of the same. But a major factor that hasn't been given weightage in calculations is the physical aspects

of the object. The camera can view objects from different angles and different distances. This paper shows that the current physical adversaries are not enough to create perturbations for object detection from a moving platform. It is believed that perturbed images can exhibit malicious behavior within a range of distances. Thus, the practical impact of these perturbations can be reduced when it comes to observation from a moving platform.

Methodology: Physical adversaries are hard to achieve targets. Previous researches show that the authors created perturbations in the image set and printed them. They captured photographs of these anomalies and feed them into the system, which was then not able to classify them correctly. Another research showed that anomalies can be created using white-box methods with which they printed sunglasses that can cause a state-of-the-art face recognition system to misclassify the attacker's face to a specific face. Both of these experiments raised serious safety and security concerns. Both of these papers are summarized above.

The idea is to fetch for an adversarial image and a target image as conceived by human sense and check for the entropy of difference using 3D tensor objects of the images. The cross-entropy cost and true class will be used to train the new neural set.

Methods which are considered to create adversarial images are:

a. Fast Sign Method
b. Iterative Methods
c. L-BFGS Method
d. Attacking a detector

The experiment included capturing 180 photos of stop signs along the highway using Iphone7 under various lighting conditions from different angles. They used the YOLO detector pre-trained for stop signs, which is then treated with 100 images taken from the camera and passing them using the methods mentioned above to create adversarial images. They also trained a VGG16 traffic sign classifier trained using the 150 images captures and tested with 30 adversarial images created from the remaining images. After that, they attached these malicious images to stop signs in the area and tested for detection using an autonomous car.

Result: The following observations were made from the experiments:

a. The printed stop signs, including the original and adversarial) were detected from a closer distance at first.
b. The printed perturbed signs were rarely misclassified as when the car got closer, the camera angle changes, and eventually, the sign fell out of frame.
c. In one case, the perturbed sign was misclassified as a ball only for 2 frames, the rest went undetected.

Hence, if the perturbation generated by a method requires a camera to be at a specific angle, it would most likely not form a significant threat to the vehicle or device. The device takes images from different camera angles as it approaches the sign, so some of the frames will be detected correctly. This paper shows that even if the sign possesses some kind of perturbation, it will go undetected when parameters like distance, angle, illumination, blurriness are taken into account.

[23]. Explaining and harnessing adversarial examples.

Author: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy

Published in: ICLR 2015 Conference, 2015, Cited 5649 times

Summary: This paper tries to explain the basic reason for occurrence problems due to adversarial perturbations in any model. The paper states that the problem becomes more prominent as we have models of higher dimension. It states that as humans live only in 3 dimensions so we cannot perceive the effect of small changes in every dimension. This paper clearly shows how very small changes in all the dimensions can change the end result of the model.

Dataset: CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.

MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: This paper majorly focuses on how linear model and non-linear model behaves on encountering adversarial perturbations and how they can be reduced. It has shown how in a linear model it is very very difficult to reduce adversarial problems as the dimensions of the model increase the more prone the model becomes to adversarial perturbations. It also shows that almost linear neural networks are also prone to adversarial perturbations.

This paper further shows that a deep learning model with higher training can reduce the risk of adversarial perturbations to a very great extent using extensive training. After doing adversarial training on MNIST with fine tuning for DBM with dropout the error rate fell to 17.9 % from 40.9 % on the original model. Average

confidence was also increased to 81.4 % of the model. The adversarial training procedure can be seen as minimizing the worst case error when the data is perturbed by an adversary. That can be interpreted as learning to play an adversarial game, or as minimizing an upper bound on the expected cost.

Result: As a summary, this paper has made the following observations:

a. Adversarial perturbations are nothing but high-dimensional dot products of different vectors. They are a result of models not being nonlinear.
b. The generalization across different models is caused majorly because adversarial perturbations tend to the weight vectors of a model.
c. The direction in which perturbation is a dot product with the image matters most, rather than the specific point in space.
d. Because direction matters most so the adversarial perturbations show generalization across various examples.
e. Models which are easy to optimize during training and testing are also easy to perturbate.

[24]. Synthesizing Robust Adversarial Examples

Author: Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok

Published in: Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018, Cited 602 times.

Summary: This paper shows how adversarial examples can be generated in the real time world as the adversarial examples generated using common algorithms like FGSM and CW have a very limited success. Prior work has shown that adversarial examples generated using these standard techniques often lose their adversarial nature once subjected to minor transformations. This paper uses a new algorithm called Expectation over transformation.

Methodology: In this algorithm they have tried to minimize the perceived distance as seen by the classifier. So they have tried to minimize the visual difference between t(x) and t(x").

To solve the problem of optimization, EOT algorithm requires the ability to differentiate between 3D render functions with respect to texture. Given any pose and choices for all transformation parameters possible, any simple 3D rendering process can be used as a matrix multiplication and addition: every pixel is nothing but some linear combination of pixels in the texture plus constant.

Result: Their work demonstrates the existence of robust adversarial examples, adversarial inputs that remain adversarial over a chosen distribution of transformations. By introducing EOT, a general-purpose algorithm for creating robust adversarial examples, and by modeling 3D rendering and printing within the framework of EOT, we succeed in fabricating three-dimensional adversarial objects. With access only to low-cost commercially available 3D printing technology, we successfully print physical adversarial objects that are classified as a chosen target class over a variety of angles, viewpoints, and lighting conditions by a standard ImageNet classifier. Their results suggest that adversarial examples and objects are a practical concern for real world systems, even when the examples are viewed from a variety of angles and viewpoints.

[25]. Robust Physical-World Attacks on Deep Learning Visual Classification.

Author: Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song

Published In: CVPR 2018 (Conference on Computer Vision and Pattern Recognition), Cited 581 times.

Summary: This paper is about hiding in plain sight. This approach just makes innocuous changes that "hide in the human psyche," rather than attempt to make imperceptible changes. Choosing road signs as an attack vector is a good approach as signs are visually simple, so it is difficult to hide perturbations. They are merged with a noisy, complex environment. And there are real-world safety effects, especially as autonomous vehicles come into major use.

Dataset: LISA, a U.S. traffic sign dataset which contains 47 different road signs. Another one is German Traffic Sign Recognition Benchmark (GTSRB) which contains a single image, multi-class classification problem. It has more than 40 classes and more than 50k images in total.

Methodology: This involves actually taking images of the real physical target object from several angles, distances, and lighting conditions. Inputs are augmented with analytic changes to brightness. Since there is not any involution of 3-D renderer, authors cover the object to avoid considering physically impossible manipulation to the background. They further identify that output class is mostly impacted by which target class. Later, the masked perturbation is converted with an alignment function before being merged into the input.

Result: This paper tested the problem using LISA-CNN, GTSRB-CNN which had 99.4% accuracy on stop sign dataset. Two types of attack are there, one is poster-printing in which print-out covers the entire sign and sticker attacks, with graffiti-like. Depending upon the experimental setup, the results vary a bit ranging from 65% success rates to 100%.

[26]. Practical Black-Box Attacks against Machine Learning.

Author: Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami

Published In: ASIA CCS '17, April 02 - 06, 2017, Cited 1408 times.

Summary: Papernot et al designed an attack that gets rid of a defence for an adversarial example that has been created previously. Adversarial examples transfer well between neural classifiers which have been trained on the same data but till then these types of attacks were limited to either white-box attacks. In this paper that limitation was shattered with a new querying heuristic that effectively takes out information about a classifier's decision boundaries only by checking its label outputs.

Dataset: MNIST dataset which is a handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9". There are about 6K training examples of every digit and 1K test examples of every digit.

Methodology: Jacobian-based Dataset Augmentation is the basis of this paper's attack.

The working of the attack goes like: We have given a point x, we have to get the target label T(x). After that find the Jacobian J of the substitute model then use the column of the Jacobian corresponding to T(x). Now, The

perturbed point is the x + sign(J[T(x)]).

Training algorithm is as:

a. Initially acquire a small dataset.
b. Select a domain sensible architecture of the substitute neural network.
c. Target Network is queried for labels to any data point without labelling.
d. Now, Network is trained on current data points.
e. Then, JbDA perturbations augments the dataset.

Result: Deep Neural Network attack results in working against logistic regression models, decision trees, SVMs, KNN and distilled networks. Authors also successfully tested the attack against the defensive distillation from Papernot 2015.

[27]. Parseval Networks: Improving Robustness to Adversarial Examples

Author: Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, Nicolas Usunier

Published In: ICML'17: Proceedings of the 34th International Conference on Machine Learning, 2017, Cited 21 times.

Summary: This paper focuses on finding methods which are going to help us in increasing the robustness to adversarial perturbation. In this paper the author introduced Parseval network, a regularization method which works layerwise for reducing the sensitivity of network to small perturbations by controlling various global constants including the Lipschitz constant. Since the deep learning neural network is a composition of various functions which are represented by its different layers, author tries to achieve higher level of robustness by constantly trying to maintain a small Lipschitz constant (e.g., 1) at every underlying layer; be it fully-connected, convolutional or residual.

Dataset: CIFAR and SVHN

Methodology: Author's main idea is to control the Lipschitz constant by using parameterization in the network with a very tight parseval frame, a generalization of orthogonal matrices.

Algorithm 1: Parseval Training

$\Theta = \{W_k, \boldsymbol{\alpha}_k\}_{k=1}^{K}, e \leftarrow 0$
while $e \leq E$ **do**
  Sample a minibatch $\{(x_i, y_i)\}_{i=1}^{B}$.
  for $k \in \{1, \ldots, K\}$ **do**
    Compute the gradient:
    $G_{W_k} \leftarrow \nabla_{W_k} \ell(\Theta, \{(x_i, y_i)\})$,
    $G_{\boldsymbol{\alpha}_k} \leftarrow \nabla_{\boldsymbol{\alpha}_k} \ell(\Theta, \{(x_i, y_i)\})$.
    Update the parameters:
    $W_k \leftarrow W_k - \epsilon \cdot G_{W_k}$
    $\boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_k - \epsilon \cdot G_{\boldsymbol{\alpha}_k}$.
    if *hidden layer* **then**
      Sample a subset S of rows of W_k.
      Projection:
      $W_S \leftarrow (1+\beta)W_S - \beta W_S W_S^\top W_S$.
      $\boldsymbol{\alpha}_k \leftarrow \text{argmin}_{\boldsymbol{\gamma} \in \Delta^{K-1}} \|\boldsymbol{\alpha}_K - \boldsymbol{\gamma}\|_2^2$
  $e \leftarrow e + 1$.

Result: Author introduced new type of neural network Parseval networks, this is a new approach in the learning of a neural network that is more robust by nature to most kinds of adversarial noise. Author proposed an algorithm which will allow us to make better optimization in the model and in a very efficient manner. Empirical results of this new learning technique on three datasets used in this paper with a fully connected, wide residual neural networks illustrates the very performance of our new approach. Parseval Networks can be treated as a byproduct of the regularization the author

proposes, the model which trains faster and is useful in making a better use of its capacity.

[28]. Boosting Adversarial Attacks With Momentum

Author: Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li

Published In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, Cited 536 times.

Summary: There are a lot of algorithms which are vulnerable to attacks by adversarial abnormalities, especially the deep neural networks. Most of the existing attacks are capable of fooling a black box model. The study proposes a broad class of momentum-based iterative algorithms. By connecting it with a momentum into an iterative process for attacks. For the improvement of the success rates for black box attacks, they apply a momentum iterative algorithm which ensemble a model and show that the adversarial model with a strong defense are also vulnerable to the black box attacks.

Methodology: This paper basically aims add implementing the following methods:

They plan to introduce a new class of attacks where they accumulate the gradients of the loss function after each iteration and then use it to stabilize optimization and try to divert from the poor local maxima.

They are also aiming at implementing a approach to attack multiple models at the same time, which shows a powerful capability of transferability of adversarial properties and preserving the high success rates of the attacks.

They have extended their approach to the following attacks for studying:

1. Momentum iterative fast gradient sign method.
2. Attacking ensemble of models.

Result: This paper introduces a braod class of momentum based attacks, which are iterative in nature and boots adversarial attacks. These can effectively fool a white-box attacks as well as black-box attacks. These methods can outperform the one-step gradient-based methods. Aiming at improving the transferability property of these attack agents, these propose to attack an ensemble of models and fuse the logits together.

[29]. Adversarial Attacks on Neural Network Policies.

Author: Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abbeel

Published in: Conference at ICLR 2017, 2017, Cited 302 times

Summary: Studies in the field have shown great advancements in the designing algorithms that hampers the raw input resulting into a misclassified objects. Researches have shown how these algorithms plays with arcade games like Atari, etc. With every neural network, there are some policies associated that parameterise the neural network. For example, for a CNN model designed to classify images, perturbations added on the training input side can cause complete fail of the trained model. There are multiple scenarios available for the study of the effect of these adversaries. Supervised learning and unsupervised learning have their own course of vulnerabilities. An adversarial model effective on one training model, is applicable on various other models as well due to property of transfer-ability in adversaries. Such vulnerabilities can target any machine model either during learning by tampering with the training data or during inference by manipulating inputs on which model is making predictions.

Methodology: The study has summarized various subjects to plant adversaries in the input as well as in the training policy. The primary mentions are given below:

1. Adversarial Example Crafting with the Fast Gradient Sign Method
2. Deep Reinforcement Learning
    - 2.1 Deep Q-Networks
    - 2.2 Trust Region Policy Optimization
    - 2.3 Asynchronous Advantage Actor-Critic

The paper shows the use of FGSM as a white-box attack to compute the adversarial perturbations on a trained network, and as a black-box attack by computing the gradients for a separately trained policy enabling the transferability property.

1. Applying FGSM to policies: When computing adversarial perturbations with FGSM for a trained model, we assume the action with the maximum weight to be the appropriate action to take and them we try to apply the adversaries to reduces the chances of output ending in the defined class.
2. Changing a Norm Constraint: It applies changing the input element by a slight n amount, or maybe change only a small number of input features resulting into a misclassified class. The change is done using a cost function dependent on N which is the amount of perturbation to be applied.

Study ends with an experimental setup to support the given argument about the application of perturbation on unsupervised and supervised models and process them such that the classifier are no longer able to define a class accurately.

The experiment involves playing the Atari game using three deep reinforcement algorithms. For each game, they trained five policies having different initial conditions. The results of each iteration which taken a cumulative so as to average out the effect of perturbations.

Result: For the experiment being conducted as two parted.

1. Vulnerability to White-box Attacks: It was found that policies trained with the three deep reinforcement algorithms were all susceptible to adversarial inputs.
2. Vulnerability to Black-box Attacks: Various properties were observed in this phase. The transferability property was prominent during the observations across both policies and training algorithms.

It is observable that there is a need to develop defenses against the adversarial attacks. This might involve adding adversarially-perturbed inputs during training of model to avoid possibilities of misclassification.

[30]. Simple Black-box Adversarial Attacks.

Author: Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, Kilian Q. Weinberger

Published in: ICML Conference, 2019, Cited 90 times.

Summary: The study is proposing a method to construct adversarial images in a black-box setting. In contrast to the white-box scenario, constructing a black-box adversarial image has a constraint on the computation cost, hence efficient attacks still remain a goal to achieve.

Taking few assumptions about the confidence values, the algorithm proposed is highly query-efficient and uses an iterative principle: they are taking a random vector on an orthonormal basis and adding or subtracting it from the target image. The proposed method can be used for both targeted and untargeted attacks, giving pretty efficient querying processing in both the scenarios.

The study shows demonstration on various real world settings including the Google Cloud Vision system. The study system stands string for becoming a baseline for future innovation in black-box attacks.

Dataset: ImageNet sample

Methodology: White-box attacks require the knowledge of the target model whereas black-box attacks merely requires querying to the target model that may result into complete or partial information about the insides.

Online services like Google Cloud Vision, etc provide access through API calls only and hence the target model is not available for study but the API returns certain confidence scores which provide an option to exploit the prediction algorithm. These confidence scores are a lot for the proposed algorithm to design a simple black-box attacks.

Here, the authors are repeatedly picking up random vectors from the orthogonal space of search directions. Using the confidence score obtained in each response to check if it is pointing towards the decision boundary or not and then perturb the image by adding or subtracting the vector from the image. Each update moves the image further from the original information and towards the decision boundary.

They are also providing some theoretical insight on the efficacy of their approach. They have observed that restricting the searching near the lower frequency end of the DCT (Discrete cosine transform) basis is pretty efficient considering the querying.

Their approach is observing similar results and success rates as compared to the state-of-the-art black-box attacks but at a fairly lower number of black-box queries. On the other hand, this algorithm is so simple to implement that it can actually be implemented in less than 20 lines of python code and hence they strongly consider their methods to become the new baseline for study on black-box attacks and adversarial image attacks.

Result: As a conclusion, the study believes that the attack proposed can be used with any orthonormal basis, provided the vectors are sampled correctly. High resolution dataset like Imagenet, etc provide certain challenges due to the vector dimensionality. They choose to evaluate the attacks on the basis of standard vectors and DCT basis vectors for their efficiency.

Given the real world applicability, the algorithm can be used to develop defense against malicious adversaries under this more realistic threat model. Also the method requires very few specifications and hence is more suitable when it comes to applications terms. While they intentionally avoided sophisticated techniques in designing this algorithm in the favor of it's simplicity. They believe that additional modifications can drastically decrease the number of queries made. There is deep field to explore in this direction.

Integrated Summary for Section 4

S. No	Methods Used	Dataset	Results	Remarks
22	Methods which are considered to create adversarial images are: a. Fast Sign Method b. Iterative Methods c. L-BFGS Method d. Attacking a detector	180 photographs of stop sign at a highway, from various angles and distances.	This paper shows that even if the sign possesses some kind of perturbation, it will go undetected when parameters like distance, angle, illumination, blurriness are taken into account.	This paper explores the region of research in the area of preventive measures against faulty classifications. It shows how a model can be made to avoid misclassification by using a few methods describe therein.
23	Monitoring the behaviors of linear model and non-linear model.	CIFAR-10 dataset - This dataset consists of 60k 32*32 colour images classified in 10 sections with 6k images in a section. Among these, 50k are training images and 10k are test images.	Adversarial perturbations are nothing but high dimensional dot products of different vectors. They are a result of models not being nonlinear.	This paper explains how the perturbations are caused and it shows that they are nothing but dot product of 2 vectors. And as they are dot product so the direction of the vectors matters most. Hence the images taken in the real world applications are less prone to perturbations as specific angle can not be maintained in the real world images.
24	Minimize the perceived distance as seen by the classifier. EOT algorithm requires the ability to differentiate between 3D render functions with respect to texture.	-Not Used-	Adversarial examples and objects are a practical concern for real world systems, even when the examples are viewed from a variety of angles and viewpoints	This paper shows that using some advanced algorithms like EOT we can generate images which are effective irrespective of the direction in which the image is taken. Hence the adversarial perturbations can cause real trouble to the mankind with increasing using of AI in day to day life.
25	Taking images of the real physical target object from several angles, distances, and lighting conditions. Inputs are augmented with analytic changes to brightness.	a. LISA, a U.S. traffic sign dataset which contains 47 different road signs. b. German Traffic Sign Recognition Benchmark (GTSRB).	Two types of attack are there, one is poster-printing in which print-out covers the entire sign and sticker attacks, with graffiti-like.	Generating physical adversarial examples robust to largely varying range is possible. This shows that defenses that came in view in future should not ase on physical sources of noise as defense against these adversarial examples.
26	Jacobian-based Dataset Augmentation	MNIST dataset which is a	Deep Neural Network attack results in	This paper clears that what humans see and

		handwritten digit recognition dataset. It contains 8-bit grayscale images of "0" through "9".	working against logistic regression models, decision trees, SVMs, KNN and distilled networks.	what algorithms see can be exploited. Humans can't make any difference between the original sign and adversarial sign which makes it difficult to identify the attack.
27	Author's main idea is to control the Lipschitz constant by using parameterization in the network with a very tight parseval frame, a generalization of orthogonal matrices.	CIFAR and SVHN dataset.	Author introduced new type of neural network Parseval networks, this is a new approach in the learning of a neural network that is more robust by nature to most kinds of adversarial noise. Author proposed an algorithm which will allow us to make better optimization in the model and in a very efficient manner.	Since the deep learning neural network is a composition of various functions which are represented by its different layers, author tries to achieve higher level of robustness by constantly trying to maintain a small Lipschitz constant (e.g., 1) at every underlying layer; be it fully-connected, convolutional or residual.
28	They plan to introduce a new class of attacks where they accumulate the gradients of the loss function after each iteration and then use it to stabilize optimization and try to divert from the poor local maxima.	-Not Used-	This paper introduces a braod class of momentum based attacks, which are iterative in nature and boots adversarial attacks. These can effectively fool a white-box attacks as well as black-box attacks.	For the improvement of the success rates for black box attacks, they apply a momentum iterative algorithm which ensemble a model and show that the adversarial model with a strong defense are also vulnerable to the black box attacks.
29	The study has summarized various subjects to plant adversaries in the input as well as in the training policy. The paper shows the use of FGSM as a white-box attack to compute the adversarial perturbations on a trained network, and as a black-box attack by computing the gradients for a separately trained policy enabling the transferability property.	-Not Used-	It is observable that there is a need to develop defenses against the adversarial attacks. This might involve adding adversarially-perturbed inputs during training of model to avoid possibilities of misclassification.	There are multiple scenarios available for the study of the effect of these adversaries. Supervised learning and unsupervised learning have their own course of vulnerabilities. An adversarial model effective on one training model, is applicable on various other models as well due to property of transfer-ability in adversaries.
30	The authors are repeatedly picking up random vectors from the orthogonal space of search directions. Using	ImageNet sample	Given the real world applicability, the algorithm can be used to develop defense	The study shows demonstration on various real world settings including the

	the confidence score obtained in each response to check if it is pointing towards the decision boundary or not and then perturb the image by adding or subtracting the vector from the image.		against malicious adversaries under this more realistic threat model. Also the method requires very few specifications and hence is more suitable when it comes to applications terms.	Google Cloud Vision system. The study system stands string for becoming a baseline for future innovation in black-box attacks.

Significance of Work

From the above study, we learned about the adversarial networks and their working. How they hamper the efficiency of an image classifier and how it is harmful on physical scale. We had the following observation after completing the study on various topics and research papers related to the former subject.

- We were able to understand the logics behind these adversarial vulnerabilities.
- We got to learn how images are read by these adversarial algorithms and broken down to add noise to disable the quality of the classifier.
- We learned about the various algorithms which are expected to get replaced by another research topics.
- We learned about the algorithms such as FGSM, APE-GAN, Normal constraints etc.
- We learned about various logical advancements aimed towards defending these adversarial attacks at low cost.
- We learned about the properties of advesarial models, their transferability properties.
- Various methods describing ideas to prevent these attacks have been discussed and it has been found that majority of the ideas focused on training the training models with all kinds of adversarial sample subjects.
- Adversarial attacks is a very vast field and is still open to contribution. There is a lot of scope of deployment and research in this area of science.
- We learned about the advantages and disadvantages of adversarial networks.
- Few studies have show how the adversarial models hamper the performance of google cloud API and other real world settings.
- We learned about the course of adversaries in the field of supervised and unsupervised learning.
- Basic implementation of black-box attacks have been perfectly defined in few of the researches.
- We learned how these adversarial principles can be used as a defense mechanism against the very own origin and how these can be employed to secure application and data.
- We got to know about the future scope of this field of research.

These were few outcomes of the extensive study conducted as a part of this report study.

Gap Analysis

The topic of Adversarial study on deep networks and machine learning is a trending topic in the field of research. Various studies are being conducted around the world trying to find out an effective defense against this vulnerability.

In the preceding study, we did not found any acceptable gaps in the research part. Papers on trending research and hot topics were made available from around the decade. Most recent studies were also available and a few papers have also been used to compose this report.

Few of the observation regarding the future work in this field are as follows:

- The principles of adversarial networks have tremendous application on both online and real-world deployment. It is possible to apply adversarial perturbation to real-world obejcts and that can be a new source of study and research.
- APE-GAN research suggests that implementing various defense mechanism together to develop layered prevention can be a direction for research in future.
- Research by Guo et al. suggest that their observation on simple black box attacks defining a new type of attacks can be string baseline for fututre work and references. The efficiency and application provides a strong basis to implement various new ideas.
- Few real world services like speech recognition can also be targeted for research under the adversarial research category.
- A simple model based on iteration that can modify the input at random to hamper the classification capabilities of a classifier is still an area to explore.
- Work by Liao et al. suggests scope of research in the field of increasing the efficiency of their HGD denoiser.
- 2016-2020 have shown alot of increase in the research considering adversarial attacks and defense mechanisms, Adversarial attacks and defense being a new field and robust field to explore has got a lot to go through.
- Different kinds of attacks and vulnerabilities appearing everyday requires a ready to go defense mechanism for ensure security.
- Various researches have come up with different kinds of adaptation of former researches showing potential to be applied to a wide range of applications.

References

[1] Patrick McDaniel, Nicolas Papernot, and Z. Berkay Celik (2016). Machine Learning in Adversarial Settings. IEEE Security & Privacy, May/June 2016.

[2] Alexey Kurakin, Ian J. Goodfellow and Samy Bengio. (2017). Adversarial Machine Learning at Scale. ICLR Conference, 2017

[3] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. (2016). Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, CCS'16, Vienna, Austria

[4] Alhussein Fawzi, Omar Fawzi, Pascal Frossard (2015), Fundamental limits on adversarial robustness, ICML 2015 Workshop on Deep Learning, Lille, France.

[5] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry, (2019), Adversarial Examples are not Bugs, they are Features, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)

[6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, (2017), Universal adversarial perturbations, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[7] Battista Biggio, Blaine Nelson, Pavel Laskov, (2019), Poisoning Attacks against Support Vector Machines, 29th International Conference on Machine Learning, Edinburgh, 2013

[8] Battista Biggio, Davide Maiorca, Igino Corona, Nedim Srndic, Blaine Nelson, Pavel Laskov, Giorgio Giacinto, and Fabio Roli, (2013), Evasion attacks against machine learning at test time, ECML PKDD 2013, Part III, LNAI 8190, pp. 387–402, 2013

[9] Nicholas Carlini, David Wagner, (2017), Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods, AISec'17.

[10] Alhussein Fawzi ,Hamza Fawzi, Omar Fawzi (2018), Adversarial vulnerability for any classifier, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).

[11] Zhengli Zhao,Dheeru Dua, Sameer Singh, (2018), Generating Natural Adversarial Examples, ICLR 2018.

[12] Shuangtao Li , Yuanke Chen, Yanlin Peng, Lin Bai, (2018), Learning More Robust Features with Adversarial, AISec'18.

[13] Nicolas Ford, Justin Gilmer, Nicholas Carlini, Ekin D. Cubuk, (2019), Adversarial Examples Are a Natural Consequence of Test Error in Noise, Proceedings of the 36th International Conference on Machine Learning, 2019.

[14] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, Tom Goldstein, (2019), Are adversarial examples inevitable, ICLR 2019 Conference

[15] Adi Shamir , Itay Safran , Eyal Ronen and Orr Dunkelman, (2019), A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance.

[16] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. (2017). Standard detectors aren't (currently) fooled by physical adversarial stop signs. University of Illinois at Urbana Champai

[17] Xiaoyong Yuan, Pan He, Qile Zhu, Xiaolin Li, (2019), Adversarial Examples: Attacks and Defenses for Deep Learning, IEEE Transactions on Neural Networks and Learning Systems, 2019.

[18] Nupur Thakur, Yuzhen Ding , Baoxin Li, (2020), Evaluating a Simple Retraining Strategy as a Defense Against Adversarial Attacks, Arxiv.org, 2020.

[19] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, Jun Zhu, (2018), Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

[20] Shiwei Shen, Guoqing Jin, Ke Gao, Yongdong Zhang, (2019), APE-GAN: Adversarial Perturbation Elimination with GAN, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[21] Kimin Lee, Kibok Lee, Honglak Lee, Jinwoo Shin, (2018), A Simple Unified Framework for Detecting Out-of-Distribution

Samples and Adversarial Attacks, Advances in Neural Information Processing Systems 31 (NeurIPS 2018)

[22] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. (2017). No Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles. University of Illinois at Urbana Champaign

[23] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy (2015), Explaining and harnessing adversarial examples, ICLR 2015 Conference.

[24] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, (2018), Synthesizing Robust Adversarial Examples, 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80.

[25] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, (2018), Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR 2018 (Conference on Computer Vision and Pattern Recognition)

[26] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami, (2017), Practical Black-Box Attacks against Machine Learning, ASIA CCS '17

[27] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, Nicolas Usunier, (2017), Parseval Networks: Improving Robustness to Adversarial Examples, ICML'17, Proceedings of the 34th International Conference on Machine Learning,

[28] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, Jianguo Li, (2018), Boosting Adversarial Attacks With Momentum, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018,

[29] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abbeel, (2017), Adversarial Attacks on Neural Network Policies, Conference at ICLR 2017.

[30] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, Kilian Q. Weinberger, (2019), Simple Black-box Adversarial Attacks, ICML Conference, 2019,