Review on Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey

Neelu Verma (MP19AI002)
verma.14@iitj.ac.in
Course Instructor: Dr. Richa Singh

Department of Computer Science &
Engineering,
IIT Jodhpur, India

Overview of Reviewed Papers

Adversarial attack has been an important research problem in Artificial Intelligence. It also play a significant role in the Deep learning in Computer vision field and in almost every related field have effect of adversarial attack. This review show the comparative analysis of various papers in the field adversarial attack, its detection, and mitigation. There is lots of reported attacks but comparative study into a single paper has been missing in previous works by the researchers for a long time. Recently, researchers have started exploring this survey comparisons, for the effect of adversarial attack in the every field like self driving car and security . So, The goal of this review writing will be reviewing the relevant reported literature on the topic "Threat of Adversarial Attacks on Deep Learning in Computer Vision", which are understanding that system can be vulnerable to these adversarial attacks in which subtle perturbations can completely fool the deep learning models and also adding some other paper in the same field.

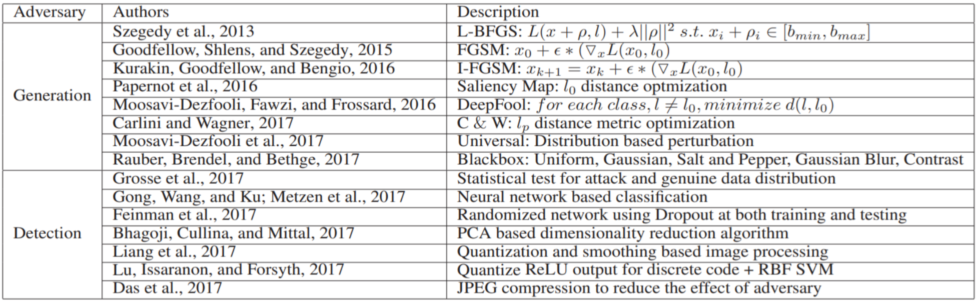
Akhtar at el.[1] presents the first comprehensive survey on adversarial attacks on deep learning in computer vision. Their review work consists design adversarial attacks, analysis, existence of such attacks and propose defenses against those attacks. This paper starts with the definitions of related terms like Adversarial example/image, Adversarial perturbation, Adversarial training, Adversary,Black-box attack,Detector, Fooling ratio/rate, One-shot/one-step methods,Quasi-imperceptible, Rectifier, Targeted attacks,Threat model, Transferability, Universal perturbation, White-box attacks. There are some attack on deep learning model discussed below beyond the classification or recognition like attack on autoencoders and generative models, attack on recurrent neural network, attack on deep reinforcement learning, attacks on semantic segmentation and object detection, attack on face attributes. Apart from these attack there is some attack which can be on real world like cell-phone camera attack, road sign attack, attack using generic adversarial 3D objects, cyberspace attacks, robotic vision and visual QA attacks. These all attack discussed in this paper are task specific. these are designed for different purpose and have its own importance and working. In a same way this paper consist of lot of literature review work which cannot fully summarise but a beautiful contribution to make findings easy. So the conclusion is that even if the deep neural network perform well on any computer vision task with high accuracy but still these models are vulnerable to perturbation which can lead to incorrect prediction. This paper resulted in contribution of recent, most influential and interesting , and devise adversarial attack and its mitigation for deep learning. It has been concluded from this paper that in mostly safety and security specific application, real threat to deep learning is adversarial attack and can also affect physical world. The future work according to this paper can be robustness of these models against adversarial attack.

Models like DNN process sensitive and proprietary dataset which can raise a serious issues regarding privacy. Zhang at el.[8] present a work on model- inversion attacks because It has been reported that successful model-inversion attack is only demonstrated on simple model like linear and logistic regression. Model-inversion attack on publicly available dataset with generative adversarial network which is generic to learn distributional prior.This paper focus on image data and using that MI-Attack method which include inversion of DNNs and synthesis of private training data with high fidelity. Experiments were performed on the three most popular dataset, MNIST handwritten digit data, X-ray database, celeb faces attribute dataset. This experiments results in improving the accuracy to 75% from the exixting work This paper compares their proposed Generative Model-Inversion (GMI) attack with existing models inversion attack (EMI) and conclude that the adversary only can exploit the identity loss and can return pixel value which minimises the identity loss. Comparison with PIL indicates that attack proposed in paper reveals private information.

with the several attack discussed in above papers, Kolouri at el[7] contribute a technique for detecting backdoor attacks on Convolutional Neural Network. Introduces a concept of Universal Litmus Pattern(ULPs) in which they feed these universal patterns to the model and than analyzing that network is clean or corrupt. dataset used for this purpose is German Traffic Sign Recognition Benchmark (GTSRB), MNIST, CIFAR10, and Tiny-ImageNe. This paper stated that backdoor attack are more stealthy. They present an approach to detection of this backdoor attack on convolutional neural network independent of training data accesss, test running. set of universal test pattern have been use for backdoor attack is present. Lot of attacks like Generating Backdoor Attacks, Evading Backdoor Attacks, Detecting Backdoor Attacks have been repoted in literature work. This paper mainly focus on detection of backdoor attack which don't have need of infected training data. leaning of universal and transferable set of pattern mentioned as a litmus test for identifying networks to detect network is poisoned or not. It means a visual trigger on training a model can misclassify images. ULP so called universal litmus pattern are images such that these images are optimised on small set of poisoned and cleaned network. Both input and output need to access for these models requires. It has been analyses that results over the ULPs small set can able to detect malicious network with relatively high accuracy. ULPs are better than Neural Cleanse in sense that ULPs for CNN require only one forward pass. In future can find a way to harden ULP-based detection against adaptive attack.

[6]This paper is a important contribution for the detection of attack. This paper stated three major contribution that is effect of deep architectures is evaluated for face recognition, detection of abnormal filter response behavior of singularities characterizing in the deep networks hidden layers and make the corrections in the processing of pipeline to the problem alleviate. This papers experimental evaluation is using open source Deep Neural Network,including

OpenFace and VGG-Face for facial recognition, and It has been demonstrated that on these two publicly available database (MEDS and PaSC) the performance can be affected by the presence of distortion. Automatically detection and correction of adversarial sample. this can be done at run time and deployement is done for real world applications.in adversarial attacks on deep learning based face recognition the proposed adversarial distortion that are able to degrade the performance of deep learning face recognition algorithms.they evaluate the robustness of deep learning based recognition in the presence of image processing based distortions.the two types of image distortions like grid based occlusion, and most significant bit based noise, along with tree face -level distortion;a.forehead and eye brow occlusion, b.eye region occlusion,and c.beard-like occlusion. In future need to build more complex mitigation frameworks which can restore a normal level of performance.



Adversary	Authors	Description
Generation	Szegedy et al., 2013	L-BFGS: $L(x+\rho,l)+\lambda\|\|\rho\|\|^2\ s.t.\ x_i+\rho_i \in [b_{min}, b_{max}]$
	Goodfellow, Shlens, and Szegedy, 2015	FGSM: $x_0+\epsilon * (\nabla_x L(x_0,l_0)$
	Kurakin, Goodfellow, and Bengio, 2016	I-FGSM: $x_{k+1}=x_k+\epsilon * (\nabla_x L(x_0,l_0)$
	Papernot et al., 2016	Saliency Map: l_0 distance optmization
	Moosavi-Dezfooli, Fawzi, and Frossard, 2016	DeepFool: $for\ each\ class, l \neq l_0, minimize\ d(l,l_0)$
	Carlini and Wagner, 2017	C & W: l_p distance metric optimization
	Moosavi-Dezfooli et al., 2017	Universal: Distribution based perturbation
	Rauber, Brendel, and Bethge, 2017	Blackbox: Uniform, Gaussian, Salt and Pepper, Gaussian Blur, Contrast
Detection	Grosse et al., 2017	Statistical test for attack and genuine data distribution
	Gong, Wang, and Ku; Metzen et al., 2017	Neural network based classification
	Feinman et al., 2017	Randomized network using Dropout at both training and testing
	Bhagoji, Cullina, and Mittal, 2017	PCA based dimensionality reduction algorithm
	Liang et al., 2017	Quantization and smoothing based image processing
	Lu, Issaranon, and Forsyth, 2017	Quantize ReLU output for discrete code + RBF SVM
	Das et al., 2017	JPEG compression to reduce the effect of adversary

Figure 1: Literature review of adversarial attack generation and detection algorithms.

Finally from above papers it is noted that small imperceptible perturbations can lead to incorrect predictions. SmartBox is a novel paper [4] in the field of computer vision. It is python base toolbox for benchmarking the performance of adversarial detection and mitigation algorithm for face recognition. SmartBox plateform can evaluate newer attack using different attack algorithm like Gradient methods, DeepFool, L2 attack, and Elastic-Net. Moreover, this paper summarises the adversarial examples generation, detection, and mitigation algorithms. As DeepFool which calculates orthogonal distance from nearest separating hyperplane. For mitigation this method proposed gaussian blurring SmartBox is the one which contain modules for all the three task:attack generation, detection, and mitigation. Short summary of adversarial examples generation, detection, and mitigation algorithms implemented in the SmartBox is given in figure 2.
In future this smartbox tool can be used to fool other biometric modality like iris and fingerprint.

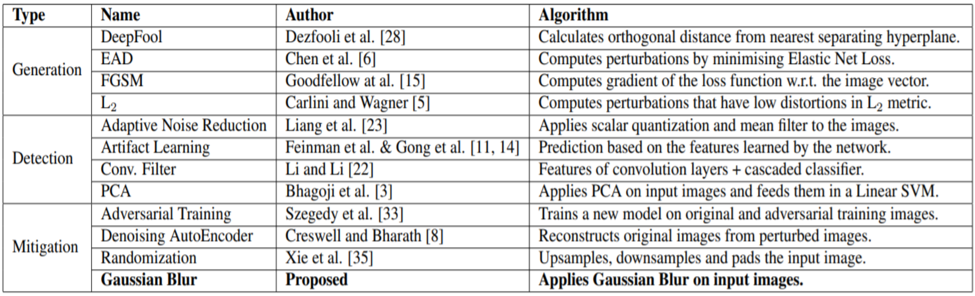
[5] This paper is combination of both Detection and Mitigation of Adversarial perturbation's for Robust Face Recognition. It is dealing with aspect related to robustness of Deep Neural Network for facial recognition. Experimentation is done using multiple open source DNN models. The proposed approach has been evaluated on four quasi-imperceptible distortions. these distortions are DeepFool, Elastic-Net (EAD), l2, and Universal adversarial perturbations. Method proposed in this approach is able to find two type of attack with great accuracy, This is done because of classifier is suitably designed by using response of hidden

Method	Black White box	Targeted Non-targeted	Image-specific Universal	Perturbation norm	Learning	Strenght
L-BFGS	White box	Targeted	Image specific	ℓ_∞	One Shot	***
FGSM	White box	Targeted	Image Specific	ℓ_∞	One Shot	***
BIM ILCM	White box	Non Targeted	Images Specific	ℓ_∞	Iterative	****
JSMA	White box	Targeted	Image specific	ℓ_0	Iterative	***
One-pixel	Black box	Non Targeted	Image specific	ℓ_0	Iterative	**
C W attacks	White box	Targeted	Image specific	$\ell_0, \ell_2, \ell_\infty$	Iterative	*****
Deep Fool	White box	Non Targeted	Image specific	ℓ_2, ℓ_∞	Iterative	****
Universal perturbations	White box	non Targeted	Universal	ℓ_2, ℓ_∞	Iterative	*****
UPSET	Black box	Targeted	Universal	ℓ_∞	Iterative	****
ANGRI	Black box	Targeted	Image specific	ℓ_∞	Iterative	****
Houdini	Black box	Targeted	Image specific	ℓ_2, ℓ_∞	Iterative	****
ATNs	White box	Targeted	Image specific	ℓ_∞	Iterative	****
GMI[8]	Black box	Targeted	Image specific	ℓ_2	Iterative	*****
EAD[3]	White box	Targeted	Image specific	ℓ_1, ℓ_2	Iterative	****

Table 1: Summary of the attributes of diverse attacking methods: The 'perturbation norm' indicates the restricted 'p-norm of the perturbations to make them imperceptible. The strength (higher for more asterisks) is based on the impression from the reviewed literature

layer in network. For mitigation of impact of adversarial attack a effective countermeasures presented in this paper which improves the overall robustness of DNN.

[2] The critical issues of adversarial attack, this paper propose a mitigation of Evasion Attack to Machine Learning cyber Detectors. The problem with previous approaches is that even if the detectors are highly affected by small perturbations but the proficient detector is also highly affected to malicious samples, and noticed that existing countermeasures are immature. So to resolve this problem upto a mark this paper present a AppCon. AppCon



Type	Name	Author	Algorithm
Generation	DeepFool	Dezfooli et al. [28]	Calculates orthogonal distance from nearest separating hyperplane.
	EAD	Chen et al. [6]	Computes perturbations by minimising Elastic Net Loss.
	FGSM	Goodfellow at al. [15]	Computes gradient of the loss function w.r.t. the image vector.
	L_2	Carlini and Wagner [5]	Computes perturbations that have low distortions in L_2 metric.
Detection	Adaptive Noise Reduction	Liang et al. [23]	Applies scalar quantization and mean filter to the images.
	Artifact Learning	Feinman et al. & Gong et al. [11, 14]	Prediction based on the features learned by the network.
	Conv. Filter	Li and Li [22]	Features of convolution layers + cascaded classifier.
	PCA	Bhagoji et al. [3]	Applies PCA on input images and feeds them in a Linear SVM.
Mitigation	Adversarial Training	Szegedy et al. [33]	Trains a new model on original and adversarial training images.
	Denoising AutoEncoder	Creswell and Bharath [8]	Reconstructs original images from perturbed images.
	Randomization	Xie et al. [35]	Upsamples, downsamples and pads the input image.
	Gaussian Blur	**Proposed**	**Applies Gaussian Blur on input images.**

Figure 2: Summary of adversarial examples generation, detection, and mitigation algorithms implemented in the SmartBox

is a original approach against adversarial evasion attacks to harden the intrusion detection. AppCon is a integration of ensemble learning for the realistic network environments which combines the layers of detectors to monitor the applications behaiviour. This paper resulted effectiveness of AppCon for the mitigation of attack and this cannot affected by limitations of existing remedy. This is a important contribution toward the security of cyber defence in reference to machine learning-based network intrusion. The quality of this approach is demonstrated on a botnot detection scenario. In future more secure cyber detector can be proposed based on the work in this paper.

Note: To view references use adobe pdf reader browser might not support references

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR*, abs/1801.00553, 2018. URL http://arxiv.org/abs/1801.00553.

[2] Giovanni Apruzzese, Mauro Andreolini, Mirco Marchetti, Vincenzo Giuseppe Colacino, and Giacomo Russo. Appcon: Mitigating evasion attacks to ml cyber detectors. *Symmetry*, 12(4):653, Apr 2020. ISSN 2073-8994. doi: 10.3390/sym12040653. URL http://dx.doi.org/10.3390/sym12040653.

[3] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11302.

[4] Akhil Goel, Anirudh Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Smartbox: Benchmarking adversarial detection and mitigation algorithms for face recognition. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*, pages 1–7. IEEE, 2018. doi: 10.1109/BTAS.2018.8698567. URL https://doi.org/10.1109/BTAS.2018.8698567.

[5] Agarwal A. Ratha N. et al Goswami, G. Detecting and mitigating adversarial perturbations for robust face recognition. int j comput vis 127, 719–742 (2019). https://doi.org/10.1007/s11263-019-01160-w.

[6] Gaurav Goswami, Nalini K. Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. *CoRR*, abs/1803.00401, 2018. URL http://arxiv.org/abs/1803.00401.

[7] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. *CoRR*, abs/1906.10842, 2019. URL http://arxiv.org/abs/1906.10842.

[8] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. *CoRR*, abs/1911.07135, 2019. URL http://arxiv.org/abs/1911.07135.