ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 10, October 2019

Performance analysis of misclassification by different neural networks for minimized perturbation

[1] Ritesham, [2] Hemant Saraswat, [3] Yash Vardhan Varshney, [4] Kapil Dev Sharma

***Abstract*—Rapid success results have been noticed by Neural Networks in various learning problems. These models follow the layered architecture to reach the decisions of classifications with a higher prediction with confidence. In this work, we've shown a comparable performance of classification by intentionally designed adversarial perturbation. The differential evolution is used to generate multi pixel to single pixel adversarial perturbation attacks to the colored image dataset of CIFAR-10. We address that the robust image classifiers can be fooled easily with perturbation analysis. The result consists of finding an adversarial perturbation that changed the output of DNN. Further, we've analyzed with misclassification drawbacks that possibly breaking a classifier and many more advantages of such vulnerability.**

***Keywords*—Image perturbation, deep neural network, object misclassification**

I. INTRODUCTION

In recent times, the traditional machine learning-based algorithms are overtaken by DNN based approaches and achieving even human-competitive results. However, several studies have revealed that artificial perturbations on the natural image can easily make DNN based image classifier misclassify the output. A lot of effective algorithms have been proposed by researchers for generating such sample images called "adversarial images". The idea for creating adversarial images is only successful if the image does not look like perturbed but capable to confuse the highly rated objects classifiers. The perturbation of image can be done by adding a tiny amount of pixels over the image, which is expected to be imperceptible to human eyes. Such modification can cause the classifier to label the modified image as a completely different class. Same as CAPTCHA authentication technique take advantages to stop malicious robot activity on a automatic form submission. But in current scenario, the AI machines learn the pattern of CAPTCHA and plain text to attack on a system. It leads to the vulnerability and reliability breaching of the system. Our proposed work is to analyze the performance of DNN with the CIFAR-10 training data set [1] and test the robustness & security. Using small amount of adversarial change in CAPTCHA image will fool the DNN while it will not be a remarkable change in human recognition [2].

The image classifiers select some certain elements to describe the class of the input image. A face classifier usualy designed with the features like pose, illumination, saturation, grains, expression, and image quality [3]. There are many possibilities of perturbation and potential misclassification with machine learning algorithms. As machine learning provides us the flexibility and effective response in situation like spam emails, intrusion detection, automated image extraction and various malicious adversary. They have carried out the amount of randomization in the natural image. The major disadvantage could be the legitimate data which has low influence in the expressivity and constrain of a learning system [4]. The use of machine learning is vulnerability while the attackers may exploit the system by their spam perturbation. The formal structure of classifiers has to be trained against the efficient defenses. The analysis and survey have been suggested in the research to illustrate the methods and guide against the attackers [5]. DNN perform pattern recognition in the most efficient way, but in case of adversarial perturbation, even small amount of data is sufficient to fool the neural network[6]. Szegedy et. al. also shows that the small change in these features may cause for the misclassification of image [7].

Some of the researchers have proposed the techniques to perturb image efficiently without higher impact of noise on image [8][9]. The perturbation is in a way by which image can be easily recognizable by human eyes. To achieve such kind of perturbation, optimization techniques can be applied. Here the target should be the misclassification of object with minimization of number of pixels and their intensity. The particle swarm optimization (PSO) and cukoo-search and artificial bee colony (ABC) can be applied to achieve such goals [10], [11].

In this paper, differential evolution (DE) is proposed to generate a perturbed image that can fool any pre-trained network. An experimental study is performed for misclassification by ResNet [12] & LeNet [13] Object recognition networks. Comparative analysis is performed for image perturbation by the different numbers of pixels (minimized) using differential evolution.

II. ADVERSARIAL IMAGE GENERATION USING DEFERENTIAL EVOLUTION

Differential evolution (DE) is one of the promising evolution algorithms that can be used for solving complex multi-modal optimization problems [Universal adversarial perturbations],. Researchers have found the application of the differential evolution algorithm for the solution of non-differentiable, dynamic, and noisy kind of optimization problems [A Survey on Object Classification using Convolutional Neural Networks.]. In addition, DE uses one-to-one selection holds only between an ancestor and its offspring which is generated through mutation and recombination, rather than the commonly used tournament selection in many other evolutionary algorithms.

Differential Evolution Algorithm

Step 1: Provide the inputs i.e. control parameters of DE: scale factor F, crossover rate Cr, and the population size NP.

Step 2: Start with the generation number$G = 0$ and initialize the population with the given constraints as: $P_G = \{X_{1,G}, \dots\dots, X_{NP,G}\}$ with $X_{i,G} = [x_{1,i,G}, x_{3,i,G}, \dots\dots, x_{D,i,G}]$and each individual uniformly distributed in the range $[X_{min}, X_{max}]$where$X_{min} = \{x_{1,min}, x_{2,min}, \dots\dots., x_{D,min}\}andX_{max} = \{x_{1,max}, x_{2,max}, \dots\dots., x_{D,max}\}$with $i = [1,2, \dots., NP]$.

Step 3: Repeat until the optimum solution is not achieved:

Repeat step 3.1 to 3.3 NPtimes

Step 3.1 ***Mutation Step***

Create a donor vector $V_{i,G} = \left\{v_{1,i,G}, \dots\dots,\right\}$, where $\{v_{D,i,G}\}$ corresponding to the i^{th} target vector $X_{1,G}$via differential mutation scheme of DE as:

$$V_{i,G} = X_{r_1^i,G} + F.(X_{r_2^i,G} - X_{r_3^i,G})$$

Step 3.2 ***Crossover Step***

Generate a trial vector $U_{i,G} = \{u_{1,i,G}, u_{2,i,G,}u_{3,i,G}, \dots, u_{D,i,G}\}$ for the i^{th} target vector $X_{1,G}$ through binomial crossover in the following way:

$$u_{j,i,G} = \{v_{j,i,G} if\ (rand_{i,j}[0,1] \leq Cr\ or\ j = j_{rand}) x_{j,i,G} otherwise$$

Step 3.3 ***Selection Step***

Evaluate the trial vector $U_{i,G}$

IF $f\left(U_{i,G}\right) \leq f\left(X_{i,G}\right)$, THEN $X_{i,G+1} = U_{i,G}$ ELSE $X_{i,G+1} = X_{i,G}$

Step 3.4 Increase the Generation Count

III. ATTACK ON IMAGE CLASSIFIER

To fool to a classifier, the images need to be perturbed. The perturbation problem can be treated as an optimization problem with some specific constraints. The perturbation is usually done by changing the smallest unit of an image i.e. pixel. If a neural network classifier (f) is receiving n −dimensional image$X = (x_1, x_2, x_3, \dots\dots x_n)$, and classification is correctly performed at class$A$ with probability of $f_t(X)$then, the target will be the decreasing the probability of class A and increasing the probability of any other output class. In this case, the vector $e(X) = (e_1, e_2, e_3, \dots. e_n)$ will be the adversarial perturbation to input vector$X$. The goal of attacking the image is to find the optimum solution $e^*(X)$for problem defined in Equation (1)

$$f_{perturbed}(X + e(X)) \qquad subject\ to\ \|e(X)\| \leq L \qquad (1)$$

The solution to the above problem will need:

- the number of pixels to be perturbed.
- Strength of modification for each pixel.

The approach adopted in this work is slightly different from the conventional one, which has been defined as Equation 2-

$$f_{perturbed}(X + e(X)) \qquad subject\ to\ \|e(X)\| \leq d \qquad (2)$$

here d should be minimum to not show the changes in the image by naked eyes. For example, one-pixel attack d=1. In the previous work [2]–[5], the modification in the part of all dimensions was suggested. There, the dimensions for perturbation were considered as d and changes $n - d$ dimensions were kept as $'0'$.

Overall, reduce perturbations are mean to be a low pixel attack and it happens on the low dimensional input space. On the other side, the one-pixel perturbation may diversify an image towards a selected direction out of n possible directions with arbitrary strength that can change the extracted feature value by the neural network layer.

IV. SIMULATION

To check the performance of proposed work, CIFAR-10 data set is used. The dataset were collected by Krizhevsky et. al. [1] that contains 60000 images. The image size is uniform for all i.e. 32*32 pixels. The dataset is created for 10 classes with 6000 images per class.

IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 10, October 2019

There are 50000 training and 10000 test images in this dataset. The dataset is divided into five training batches and one test batch, each with 10000 images.The test batch contains exactly 1000 randomly-selected images from each class. The selected classes for dataset are: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. The classes are completely mutually exclusive. There is no overlap between automobiles and trucks. "Automobile" includes sedans, SUVs, things of that sort. "Truck" includes only big trucks. Figure 1 shows the classes in the dataset, as well as 10 random images from each.

V. RESULTS AND DISCUSSION

Attack is performed in two cases, one is target attack and the other is non-targeted attack. In targeted attack, perturbation is performed is a way that will force classifier to recognize object in targeted class. In case of non-targeted attack, objective is to only misclassify the input image. In both cases, images were perturbed with one to five pixels and tested with classifiers.

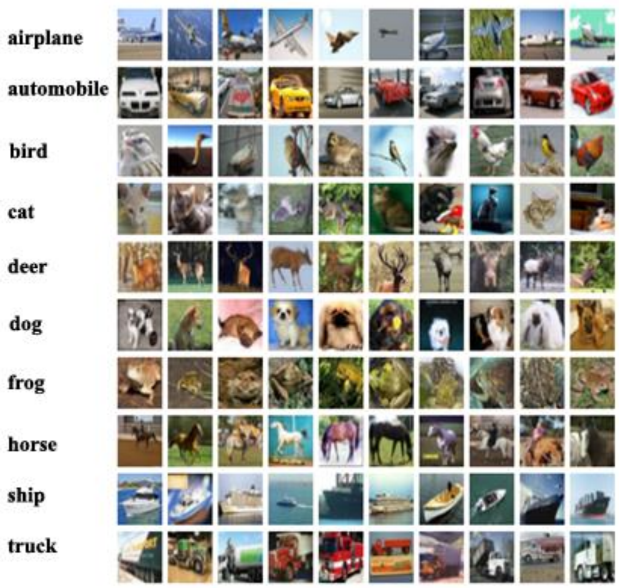

Figure 1: CIFAR-10 dataset classes with 10 random images with each class

All obtained results are summarized to compare the performance of image perturbation against the LeNet and ResNet (Figure 2). Results are showing the performance of ResNet is better in correct image classification for one, two and three pixels attack as compare to LeNet. As the number of pixels increases for perturbation purpose, the ResNet's start to more misclassify the image.

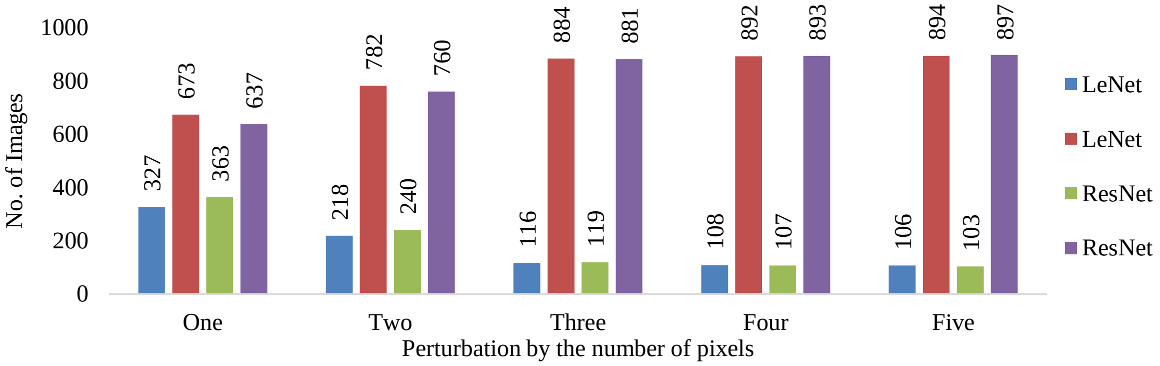

Figure 2: Comparative analysis for image perturbation by different number of pixels using differential evolution

The average performance of perturbation for targeted misclassification by LeNet and ResNet is shown in Figure 3. ResNet shown more robust performance for one, two and three pixels attack as compare to LeNet. However, the better recognition capability of ResNet turn the results for high perturbation i.e. for four and five pixel attack. The perturbation succeeded in better way for targeted misclassification in these cases.

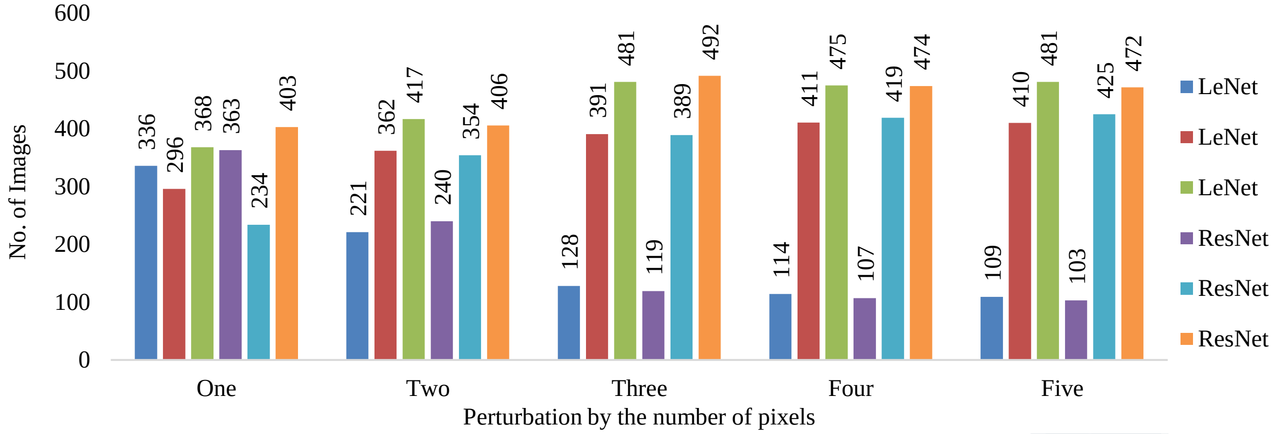

Figure 3: Comparative analysis for image perturbation by different number of pixels using differential evolution

The presented results show that the perturbation by high number of pixels, confuse the classifier network with high probability which is obvious. However the performance of perturbation using one, two and three pixels is lower as compare to LeNet. Still significant number of images can be misclassified by the attacking on right areas. The perturbation fool the LeNet 67.3%, 78.2%, 88.4%, 89.2% and 89.4% time for one, two, three, four and five pixel attack where the ResNet misclassify the input images by 63.7%, 76.0%, 88.1%, 89.3%, 89.7% for same attacks.

CONCLUSION

In this work, we've analyzed the performance of image classifiers on the minimized adversarial perturbation with the training dataset. The differential evolution is the proposed algorithm to generate the minimal perturbation that is sufficient to alter the classifier's output. Various experiments are done to move data points such that classification results in false outputs. To do this many data points are moved which may or may not be near the decision boundaries, result in images in which change easily detectable by human eyes. Through DE algorithm models we've minimized the pixel from a five-pixel attack to one pixel to test the vulnerability of the DNN.

REFRENCES

[1] [1] A. Krizhevsky, "CIFAR-10 database." [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html.

[2] U. A. C. Attack, N. Yu, and K. Darling, "A Low-Cost Approach to Crack Python CAPTCHAs Using AI-Based Chosen-Plaintext Attack," 2019.

[3] Y. Taigman, M. A. Ranzato, T. Aviv, and M. Park, "DeepFace : Closing the Gap to Human-Level Performance in Face Verification," Computer Vision and Pattern Recognition (CVPR), pp. 1–8, 2014.

[4] M. Barreno, A. D. Joseph, and J. D. Tygar, "Can Machine Learning Be Secure ?," ACM Symposium on Information, Computer, and Communication Security, no. March, 2006.

[5] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," Machine learning, no. November, pp. 121–148, 2010.

[6] A. Fawzi and P. Frossard, "DeepFool : a simple and accurate method to fool deep neural networks," Computer Vision and Pattern Recognition (CVPR), pp. 2574–2582, 2019.

[7] C. Szegedy, I. Goodfellow, J. Bruna, and R. Fergus, "Intriguing properties of neural networks," ICLR, pp. 1–9, 2013.

[8] O. Fawzi and P. Frossard, "Universal adversarial perturbations," preprint arxiv, 2016.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "EXPLAINING AND HARNESSING," ICLR, pp. 1–11, 2015.

[10] P. Civicioglu and E. Besdok, "A conceptual comparison of the Cuckoo-search , particle swarm optimization , differential evolution and artificial bee colony algorithms," Artificial intelligence review, pp. 315–346, 2013.

[11] P. Sabarinath, N. B. K. Babu, M. R. Thansekhar, and R. Saravanan, "Performance Evaluation of Differential Evolution and Particle Swarm Optimization Algorithms for the Optimal Design of Closed Coil Helical Spring," International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, no. 3, pp. 1423–1428, 2014.

[12] H. A. Abbass, "The Self – Adaptive Pareto Differential Evolution Algorithm," congress on evolutionary computation, 2002.

[13] R. K. O. Bayot, "A Survey on Object Classification using Convolutional Neural Networks."

AUTHOR’S PROFILE

Ritesham Shastri is a Master’s student at IET, Alwar. He is also a mentor and trainer at NITI Ayog, Govt. of India. He is currently involved in exploring the field of Neural Network and image recognition. He is very fond of IoT, Embedded System Design and cloud computing. He had worked on microwave-based ground penetration radar, Wi-Fi enabled walky-talky device and various other innovations.

Hemant saraswat has completed his bachelors from Hindustan college of science & technology, Mathura and masters from AMU, Aligarh in communication and information system. His area of interest is machine learning, deep neural network and image processing.

Yash Vardhan Varshney is a Research Assistant at IIT Bombay, Powai. He has completed his Ph.D. from AMU, Aligarh in speech signal processing. He has a 5 year experience of teaching to a graduate students. His area of interest is machine learning, biomedical and speech signal processing and neuropsychology. Click the link for his publication details https://orcid.org/0000-0001-9254-8986.

Kapil Dev Sharma is involved in researches related to image and communication. He is faculty in the CS Department and taking active participation in technical research publications in the field of NS2 protocol, cybersecurity and wireless network. He has been a member of several conferences and workshop committees. Design analysis and algorithm is his favorite area of work.