An adversarial training method for enhancing the robustness of deep neural networks against adversarial attacks

Mohsin Ali[1], Haider Raza[1], and John Q Gan[1]

School of Computer Science and Electronics Engineering, University of Essex, Colchester, UK.
ma22159@essex.ac.uk

Abstract. Recent studies have highlighted the vulnerability of deep neural networks to adversarial attacks in computer vision. Images generated by adversarial attacks have imperceptible changes to the original image that deceive deep neural networks into producing incorrect outputs despite being visually indistinguishable from humans, which can result in serious security and integrity issues. This work focuses on enhancing the robustness of deep neural networks against adversarial attacks by proposing a novel adversarial training framework. Specifically, ResNet architectures are pre-trained and one of them was selected to generate adversarial samples for augmenting the training dataset to retrain the ResNets. Experiments have been conducted on three different datasets and experimental results show that the proposed method is able to increase the overall robustness of the ResNets against adversarial attacks. It is noteworthy that adversarial samples generated from one deep learning model can be used to increase the robustness of another deep learning model.

Keywords: adversarial attack · adversarial learning · deep neural networks · ResNet · computer vision

1 Introduction

In recent years, deep learning has been successfully applied to the field of computer vision[1] such as in image classification[2], object detection[3], and image segmentation[4], which led to a significant advancement in security applications such as face recognition[5], person re-identification[6], and video surveillance applications[7]. These critical applications require a high degree of security to ensure the integrity of the system and thus it is important to make the deep learning models involved robust and reliable against adversarial attacks.

It has been shown that the output of a convolutional neural network (CNN) can be easily manipulated with slight changes to the input (Image) and these changes are often unidentifiable to the human eye [8]. Furthermore, recent literature [9] related to CNN shows that even the state-of-the-art deep neural networks for image classification can be easily fooled with a slight change to the input, which raises concern about the trustworthiness of these methods.

In recent years, many researchers have investigated adversarial attacks [10] and the defence mechanisms that can be used to make deep neural networks more robust against adversarial attacks [11]. The most common and simple type of adversarial attack against neural networks is called Fast Gradient Signed Method (FGSM). In this type of attack, the input image pixels are changed to the opposite direction (sign) of the gradient, which can easily fool the neural network to misclassify the input image. Although these changes may not be visible to humans and cause serious problems in critical applications of deep neural networks.

This paper proposes a novel adversarial training framework to increase the robustness of deep neural networks against adversarial attacks. Experiments have been conducted to analyze the effect of adversarial attacks and the defence mechanisms for deep neural networks, and demonstrate the effectiveness of the proposed framework. The main contributions of this paper are as follows:

1. A novel framework is proposed to increase the robustness of ResNets against FGSM attacks, which makes use of the adversarial knowledge from one ResNet architecture to increase the robustness of different ResNet architectures. Multiple ResNets[12] architectures were initially trained on MNIST [13], CIFAR10 [14], and Cats vs Dog [15] datasets. Adversarial attacks were than conducted on these ResNets to show their vulnerability in terms their classification accuracy. After these ResNets were retrained using the proposed framework, they were exposed to the adversarial attacks again, and the experimental results show that the average accuracy of the retrained ResNets was increased by 41.54% on MNIST dataset, compared to that of the initially trained ResNets, by 14.96% on CIFAR10 dataset, and by 26.43% on Cat vs Dog dataset.
2. It has been found that CNN models with greater size or number of parameters are more robust to adversarial attacks than CNN models with fewer parameters, which implies that the decision boundaries of adversarial samples are more complex than those of normal input images.

In section 2, related work is reviewed. The proposed method is described in section 3. Experimental results are presented in section 4. Section 5 draws a brief conclusion and elaborates on possible improvements in future work after discussing the results.

2 Related Work

In recent years many researchers have contributed to improving the integrity of neural networks. Adversarial attacks are a common type of threat to neural networks. Goodfellow et al. [8] proposed a method to generate adversarial examples with low computation cost, which were powerful enough to fool a variety of state-of-the-art neural networks. They also showed how small permutations to the original examples, which cannot be noticed by human beings can mislead neural networks such as GoogleNet. Some researchers such as Alex et al. [9] focused on generating adversarial examples to fool neural networks, whilst some other researchers [16] worked on making robust neural networks against adversarial attacks. For example, Madry et al. [16] showed that increasing the capacity of a neural network can help it avoid adversarial attacks. Other techniques such as adversarial training have been proposed for making neural networks more robust. Taga et al. [17] have investigated why adversarial attacks are successful on state-of-the-art models. Overfitting was regarded as one of the reasons. Papernot et al. [18] believed that adversarial attacks were successful due to the high non-linearity of the neural network. However, Goodfellow et al[8] disproved these findings. It was found that linear regularized models were insufficient to avoid adversarial attacks. There are two major types of adversarial attacks, black-box attacks and white-box attacks [19]. Black-box attacks[20] can be defined as when an attack is conducted on the model without having the access to the parameters of the model. There are different types of white-box attacks, including Fast Gradient Sign Method or FGSM proposed by Good Fellow et al [8] and its variants. This type of adversarial attacks are generated using the gradient of the loss function related to the input x. It can be mathematically denoted as:

$$X' = X + \epsilon * sign\left(\nabla_x J\left(X, y_{true}\right)\right) \tag{1}$$

Kurakin et al. [21] proposed a variant of FGSM, known as the One-Step Target Class Method, which targets a specific class and moves the features of the input x in the direction of the target class. It can be denoted as:

$$X' = X - \epsilon * sign\left(\nabla_x J\left(X, y_{target}\right)\right) \tag{2}$$

The Basic Iterative Method (BIM) is a further extension of the FGSM attack, in which small steps are taken in the direction of the target class and it generates adversarial examples closer to real examples compared to normal FGSM or One-Step Target Class Method. It can be denoted as:

$$X'_0 = X, X'_{n+1} = Clip_{X,\epsilon}\left\{X'_n - \alpha * sign\left(\nabla_X J\left(X'_n, y_{true}\right)\right)\right\} \tag{3}$$

Dong et al. [22] and Miyato et al [23] conducted experiments to generate adversarial examples using the BIM method and obtained good results. However, the BIM method is more costly than the other methods as it needs to calculate the gradient of the loss function after each step.

3 Method

Figure 1 illustrates a block diagram of the proposed method. In the experiment, the dataset is divided into training, validation, and testing subsets for training and validating CNNs with the conventional method. The adversarial test accuracy was calculated using multiple threshold (ϵ) values. Adversarial examples generated from ResNet18 are used to augment training data with data balancing taken into account to retrain the ResNets to make them more robust against adversarial attacks.

3.1 Problem Formulation

This paper focuses on handling white-box attacks [19] in which the model's parameters are accessible. These types of attacks are more powerful and can easily fool neural networks by analyzing their structures while generating adversarial examples. Suppose x is an input to the neural network, and let f be a trained neural network. In this type of attack, the objective is to generate an adversarial example $\tilde{x} = x + \eta$ by adding some noise η to the original input x so that $f(\tilde{x}) = Y'$ and $Y' \neq f(x)$. White-box attacks are mostly direction-sensitive attacks, in which the direction needs to be calculated for each feature of input x using the model's gradient. After calculating the direction of the gradient, each pixel moves in the opposite direction of the gradient to generate adversarial samples. This paper proposes a new method to improve the robustness of deep neural networks against white-box attacks.

3.2 The proposed method for enhancing the robustness of deep neural networks against adversarial attacks

Multiple ResNet [12] architectures (ResNet18, ResNet50, and ResNet101) were used in this study. To reduce the overfitting of the models the early stopping technique was applied. The trained models were evaluated in terms of testing accuracy as all the datasets in the experiment (MNIST, CIFAR10, Cat vs Dog) are balanced.

The adversarial attack method proposed by Goodfellow et al. [8] was applied to generate adversarial samples as follows:

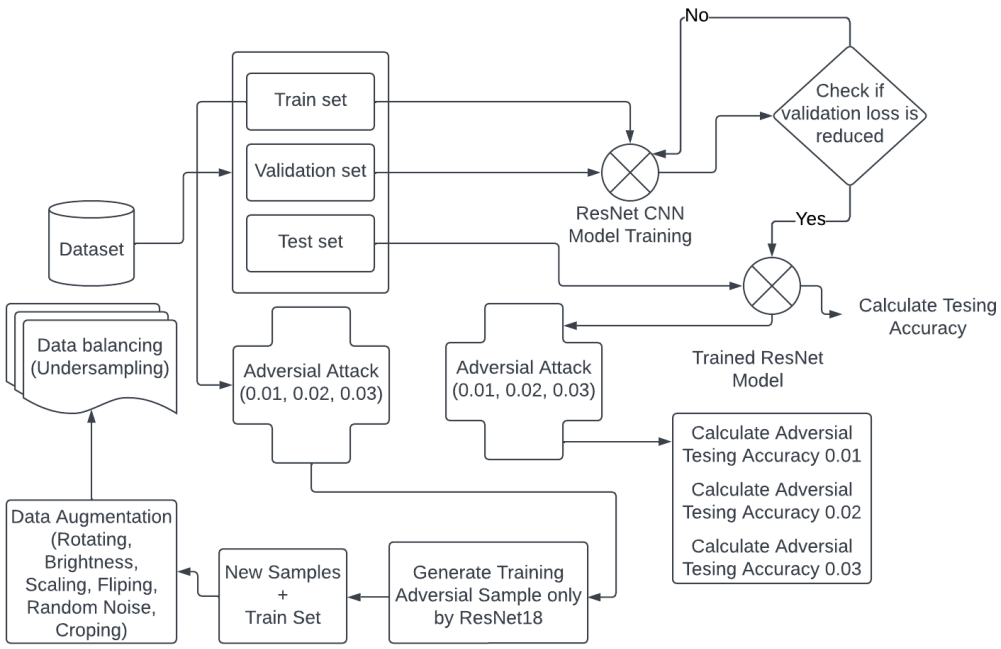

Fig. 1. A proposed method based on adversarial training framework to enhance the robustness of deep neural networks against adversarial attacks.

$$\tilde{x} = x + \eta \tag{4}$$

where $\widetilde{x}$ denotes the adversarial example, x denotes the original image input, and η represents the adversarial noise that can be represented mathematically as follow:

$$\eta = \epsilon sign\left(\nabla_x J(\theta, x, y)\right) \tag{5}$$

where θ are the parameters of the CNN model, x is the input image, y is the target variable that the CNN model should return ideally, $J(\theta, x, y)$ is the loss function used to train the CNN model, and ϵ is the threshold determining how much change is made to the pixel value of the input image. Note that changes to the pixel value can be positive or negative depending on the gradient. During the experiment, it was observed that this method can generate adversarial examples that can be easily misclassified by a variety of CNN models.

After conducting the adversarial attacks on the ResNets the testing accuracy of each model was calculated with different ϵ values (0.01, 0.02, 0.03). It was noted that ResNet18 suffered the most from adversarial attacks. To enhance the robustness of the ResNets against adversarial attacks, the adversarial samples generated from ResNet18 are added to the training set for each dataset respectively, and all the ResNets were retrained using the augmented training data. Furthermore, to diversify the training datasets data augmentation techniques such as rotating, flipping, scal-

ing, cropping, brightness, and adding random noise were applied. To rebalance the datasets, data balancing techniques were applied, in particular undersampling.

4 Experimental Setup and Results

4.1 Experimental Setup

Three benchmark datasets, MNIST [13], CIFAR10 [14], and Cats vs Dog [15], were used in the experiment. These datasets were split into 3 parts: 70% for training, 10% for validation, and20% for testing. The experiment was conducted on Nvidia T4 which contains 2560 CUDA cores and has graphic memory of 16 GB and 32GB of RAM, with Linux Ubuntu 20 as the operating system. The software platform was based on Python 3.8 along with deep learning and computer vision libraries such as PyTorch and Torch vision.

4.2 Results

First of all, the ResNets were trained and tested on the three datasets without adversarial attacks. ResNet101 achieved a maximum accuracy of 98.65% on MNIST and ResNet10 achieved a minimum accuracy of 75.39% on CIFAR10. Detailed results are shown in table 1.

After that, adversarial attacks were conducted on ResNet18, ResNet50, and ResNet101 respectively, by generating adversarial samples from the original test data as new test data. The accuracy of the ResNets on these adversarial samples was calculated. It was clear that adversarial attacks using any threshold value (ϵ) decreased the accuracy of ResNet18, ResNet50, and ResNet101 significantly, showing that all the ResNets were highly vulnerable to adversarial attacks. However, it was noticed that models with higher complexity (having more parameters) were slightly less vulnerable to attacks. It was also noticed that the ResNets were more robust on some datasets than others. The most significant drop from the average original accuracy was found on the Cats vs Dog dataset, which is (95.35 - 31.08) 64.27%, and the least drop in average accuracy was recorded on the MNIST dataset at about (97.93 - 39.46) 58.47%, as shown in table 2.

To demonstrate the effectiveness of the proposed framework for enhancing the robustness of deep neural networks, the ResNets were retrained using training data augmented by adversarial samples. The key idea behind the proposed framework is to use the knowledge of the adversarial samples generated from one neural network architecture, which is ResNet18 in this experiment, to improve the robustness of other neural network architectures such as ResNet50 and ResNet101. The adversarial samples generated by ResNet18 using multiple thresholds (0.01, 0.02, and 0.03) were added to the original training sets of the three datasets. Furthermore, data augmentation techniques such as random rotation, random shifting, random brightness, random skew, etc. were applied to diversify the training data, and data-balancing techniques, specifically undersampling, were applied to make the training data balanced. After retraining using the proposed method, the testing accuracy of ResNet18, ResNet50, and ResNet101 on adversarial test samples was significantly increased, as shown in table 3. It can be clearly seen that the proposed method resulted in the most significant improvement on the MNIST dataset using ResNet50, by up to 52.25%. Moreover, it can be observed that more improvement in testing accuracy was archived in deeper models than in shallow models or models with a smaller number of free parameters.

It was also noticed that when the ResNets were retrained using the proposed framework their accuracy on the original test data without adversarial samples was also increased, as shown in Table

Table 1. Improvement in test accuracy of ResNet by adversarial training

	Before adversarial training			**After adversarial training**		
	ResNet18	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**
MNIST	97.43%	97.72%	98.65%	99.20%	99.22%	99.26%
CIFAR10	75.39%	83.69%	81.69%	76.63%	80.41%	78.48%
Cat VS Dog	96.93%	95.05%	94.08%	97.16%	95.83%	94.34%

1, in which the most significant improvement can be found on the MNIST dataset using ResNet18 with an overall test accuracy improvement of 1.8%.

Furthermore, to establish that the observed improvement in the results presented in Table 4 is attributable to the proposed adversarial training framework and not solely due to conventional data augmentation and data balancing techniques, additional experiments were conducted. Specifically, an experiment was carried out to calculate the accuracy solely with the use of data augmentation techniques, as presented in Table 3. The results reveal that there was an enhancement in the adversarial testing accuracy with the use of conventional data augmentation techniques. However, the proposed adversarial training framework further improved the results, as demonstrated in Table 4.

5 Discussion and Conclusion

In critical applications such as security-based deep neural networks must be robust against adversarial attacks. This paper proposes a framework to train a deep neural network by using the information generated by another deep neural network through an adversarial attack. It also investigated how the size of a deep neural network affects its robustness.

The experimental results show that ResNet18 suffered more from adversarial attacks than ResNet50 and ResNet101, which indicates that a neural network with fewer parameters is more vulnerable than a neural network with more parameters [8]. Similarly, it was also noted that some datasets such as Cat vs Dog suffered more from adversarial attacks than others.

This study shows how vulnerable deep neural networks are against adversarial attacks. Even state-of-the-art models can be easily fooled by adversarial examples that are similar to original examples and cannot be distinguished by the human eye. Further, a novel framework is proposed in this paper to make deep neural networks more robust against adversarial attacks. Experimental results have shown that the proposed method is able to significantly improve the performance of deep neural networks in terms of test accuracy with adversarial test samples. On average 41.54% improvement in test accuracy was achieved on the MNIST dataset, 14.96% improvement on CIFAR10, and 26.43% improvement on Cat vs Dog.

For future research, the work in this paper could be extended toward the direction of explainability, where the effect of adversarial attacks and improvement in the performance of deep neural networks by adversarial training can be visually analyzed. It is also worthwhile to analyze the reasons why some datasets are more vulnerable to adversarial attacks than others.

Table 2. Test accuracies of the models on the generated adversarial examples or after the adversarial attack.

	ϵ=0.01			ϵ=0.02			ϵ=0.03		
	ResNet18	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**
MNIST	52.30%	60.21%	64.09%	36.50%	31.53%	35.91%	34.90%	17.46%	22.32%
CIFAR 10	27.52%	16.04%	12.21%	21.52%	14.86%	10.38%	19.56%	12.00%	9.88%
Cat VS Dog	15.76%	24.97%	4.73%	34.94%	42.44%	18.95%	46.98%	49.74%	41.21%

Table 3. Test accuracies of the models on the adversarial examples after training using multiple image augmentation techniques.

	ϵ=0.01			ϵ=0.02			ϵ=0.03		
	ResNet18	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**
MNIST	61.86%	77.55%	73.10%	46.86%	69.40%	64.25%	37.47%	62.66%	57.56%
CIFAR 10	30.06%	22.24%	27.72%	25.62%	17.96%	25.86%	24.35%	16.54%	23.52%
Cat VS Dog	25.67%	30.47%	20.84%	20.32%	49.99%	30.54%	17.94%	43.87%	31.58%

Table 4. Test accuracies of the models on adversarial examples after retraining using the proposed framework.

	ϵ=0.01			ϵ=0.02			ϵ=0.03		
	ResNet18	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**	**ResNet18**	**ResNet50**	**ResNet101**
MNIST	91.53%	93.49%	93.50%	78.88%	83.78%	83.34%	64.63%	67.53%	69.40%
CIFAR 10	37.77%	34.73%	37.89%	29.27%	22.99%	28.88%	26.67%	27.81%	27.86%
Cat VS Dog	54.04%	50.81%	59.94%	54.10%	57.81%	60.99%	54.14%	59.00%	66.81%

References

1. Yingjie Tian, Duo Su, Stanislao Lauria, and Xiaohui Liu. Recent advances on loss functions in deep learning for computer vision. *Neurocomputing*, 2022.
2. Yinglong Li. Research and application of deep learning in image recognition. In *IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 994–999, 2022.
3. Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee. A survey of modern deep learning based object detection models. *Digital Signal Processing*, page 103514, 2022.
4. Narges Kheradmandi and Vida Mehranfar. A critical review and comparative study on image segmentation-based techniques for pavement crack detection. *Construction and Building Materials*, 321:126162, 2022.
5. Prateek Singhal, Prabhat Kumar Srivastava, Arvind Kumar Tiwari, and Ratnesh Kumar Shukla. A survey: Approaches to facial detection and recognition with machine learning techniques. In *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021*, pages 103–125. Springer, 2022.
6. Nikhil Kumar Singh, Manish Khare, and Harikrishna B Jethva. A comprehensive survey on person re-identification approaches: various aspects. *Multimedia Tools and Applications*, 81(11):15747–15791, 2022.
7. Devashree R Patrikar and Mayur Rajaram Parate. Anomaly detection using edge computing in video surveillance system. *International Journal of Multimedia Information Retrieval*, 11(2):85–110, 2022.
8. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
9. Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018.
10. Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. 5 2016.
11. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. pages 506–519. Association for Computing Machinery, Inc, 4 2017.
12. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
13. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
14. Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
15. Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C.V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.
16. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
17. Kei Taga, Keisuke Kameyama, and Kazuo Toraichi. Regularization of hidden layer unit response for neural networks. In *IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM 2003)(Cat. No. 03CH37490)*, volume 1, pages 348–351, 2003.
18. Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017.
19. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2016.
20. Marius Kloft and Pavel Laskov. Online anomaly detection under adversarial impact. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 405–412. JMLR Workshop and Conference Proceedings, 2010.

21. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.
22. Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.
23. Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.