A Study on the Optimization of the CNNs for Adversarial Attacks

Hyeongcheol Park, and Jongweon Kim*

Sangmyung University, Seoul, Korea.

Convolutional Neural Networks (CNNs) have shown high accuracy in image classification tasks, including on popular datasets like MNIST and ImageNet. However, these models can be easily fooled by small perturbations added to the input image. To address this issue, the VOneBlock architecture was proposed, which can be added to the frontend of a CNN-based model to improve its robustness to adversarial attacks. In this paper, we compare the performance of CNN models with and without VOneBlock when fine-tuning on adversarial datasets, and show how the inclusion of VOneBlock affects the model's robustness. Additionally, we investigate how the number of Gabor filter kernels used in VOneBlock affects its performance. Through our experiments, we present an optimal way to enhance the robustness of CNN models to adversarial attacks using VOneBlock. Finally, we evaluate whether the classification model with VOneBlock performs well in classifying real-world attacked images, as well as adversarial attacked images. While VOneBlock was developed to improve robustness to small perturbations, we find that the neural network with VOneBlock performs slightly better in classifying real-world attacked images.

CCS CONCEPTS • Computing methodologies • Artificial intelligence • Computer vision

Additional Keywords and Phrases: Adversarial attack, VOneBlock, VOneNet, Real-life attack

1 INTRODUCTION

Since the advent of AlexNet [1], a deep convolutional neural network, in 2012, numerous types of CNN-based networks have been produced and used for object recognition [2], [3]. In addition, network combined with traditional classification algorithm and CNN achieved high classification accuracy [5]. Although advanced networks showed very high performance in classification problems, they could be easily fooled by small perturbations made to deceive the network. These small perturbations are called adversarial attacks, which is not a problem to human since it is not well visible to the human eyes. But CNN-based networks have been shown to be vulnerable to those perturbations [4].

Several methods have been proposed to increase the robustness of the classification network against adversarial attacks [5], [14], [16]. One of the methods is adversarial training in which the network trains an adversarial image. But this kind of method is expensive and performs poorly on original clean data [6].

Therefore, to solve this problem, VOneNet, which combines deep learning technology and a neural network inspired by neuroscience, was presented [6]. As a result, it was possible to create a robust model against white-box attacks [6]. However, there are not only white-box attacks in adversarial attacks, but also black-box attacks that are not affected by the network. In addition, when networks for object classification are

* Hyeongcheol Park is with the Department of Converged Electronic Engineering, Sangmyung University, Seoul, Korea. (e-mail: cchptr204@gmail.com)
Jongweon Kim is with the Department of Intelligent Internet of Things, Sangmyung university, Seoul, Korea. (e-mail: jwkim@smu.ac.kr corresponding author)

used in the real world, objects can be corrupted by damage that is common in outside [15], such as graffiti, dust, rain, etc. For example, in the case of vehicle signs, they can be corrupted by graffiti or weather issue which may cause serious problems for autonomous vehicles that should judge road conditions based on those signs.

In this paper, the ImageNet [7] dataset and MNIST [8] dataset, which are often used as benchmarks for object detection, are used as original data. DAmageNet [9] dataset, which is data that performed black-box attack on ImageNet dataset, and data that performed FGSM attack [10], a kind of white-box attack, on MNIST data are used as adversarial attacked data. Then, using fine-tuning method, train VOneBlock connected network and original network on adversarial data to compare these network's performance. Therefore, we can present some problems that can occur in the original network, which has not connected with VOneBlock. And while reducing the number of kernels of Gabor Filter Bank (GFB), which is a component of the VOneBlock, we present how the number of channels of GFB affects the robustness of the model.

Finally, we made some scribble on the German Traffic Sign Recognition [11] dataset by drawing lines randomly on images to see whether the network with VOneBlock shows better performance than the network that has no VOneBlock. This result will show whether VOneNet is more robust than a network without VOneBlock even for data under real-life attack.

2 BACKGROUND

2.1 VOneNet

VOneNet was developed through the discovery that CNN with hidden layers that fit V1 well, which means primary visual cortex, is more robust to adversarial attacks [6] [12]. VOneNet includes a fixed weight neural network on the front-end of the network that acts like a primate V1 and a state-of-art CNN on the back-end of the network. VOneBlock is the front-end model in VOneNet, and it's a linear-nonlinear-Poisson model (LNP) [6].

VOneBlock consists of three layers: a Gabor Filter Bank (GFB) layer that performs both linear and nonlinear operations, a Noise layer that gives stochasticity, and an output layer that returns identity. Figure 1 shows us the architecture of the VOneBlock. GFB layer allows VOneBlock to better approximate primate V1 with the mathematically parameterized GFB.

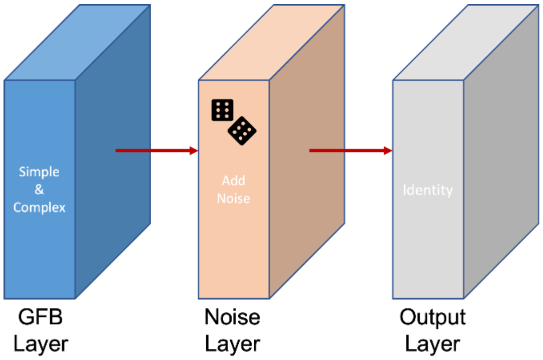


Figure 1: VOneBlock architecture.

VOneNet can be created by connecting a VOneBlock which made of 3 layers, and a CNN based network. Input image goes through VOneBlock and become [batch_size, 512, 56, 56] size activation, then pass the 1×1 BottleNeck layer which compress the depth of activation to [batch_size, bottleneck_connection_channel, 56, 56] and enter as input of CNN based network. Figure 2 shows us the VOneNet architecture as we explained.

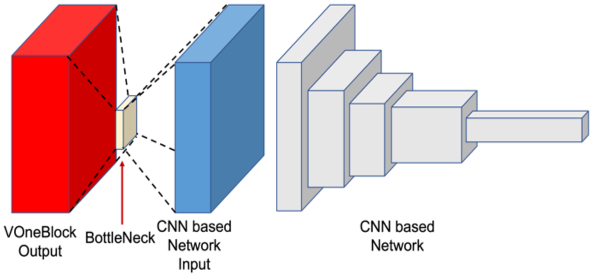


Figure 2: VOneNet architecture.

2.2 Adversarial attack

An adversarial attack is the operation of adding a small perturbation to dataset for the purpose of confusion the model. Adversarial sample generated by the adversarial attack cause the network to infer incorrectly [4]. To a human eye, there is not much difference between an attacked image and a clean image, but the network gets fooled by attacked image with high certainty.

There are two types of adversarial attacks method: black-box attack and white-box attack. In this paper, we use two adversarial datasets: DAmageNet [9] dataset, which is a dataset subjected to black-box attack on the ImageNet dataset, and the dataset performed FGSM [10] attack on the MNIST dataset.

2.2.1 White-box attack: FGSM

FGSM is a technique for generating adversarial images using the gradient from a pre-trained model. If the input to the model is an image, the gradient of the cost function for the input image is calculated to generate an image that maximize the loss. This newly created image is called an adversarial image and equation (1) can be applied to generate an adversarial image.

$$adv_x = x + \varepsilon * sign(\nabla_x J(\theta, x, y)) \quad (1)$$

2.2.2 Black-box attack: DAmageNet

Unlike white-box attack, black-box attack refers to an attack technique that create an adversarial sample by adding perturbation to an image without prior information about pre-trained model [9]. DAmageNet is a dataset that performed black-box attack on the ImageNet and contains a total of 96,020 adversarial samples with 1000 labels. DAmageNet has a high error rate of approximately 90 to 100% when entered a pre-trained network [9].

3 EXPERIMENT SETUP

In this paper, a total of four datasets, MNIST, ImageNet, DAmageNet, GTSRB, were used to compare the accuracy between VOneNet and the plane network. Resnet [3] was used as a CNN-based network, which is a plane network. Before we are fine-tuning the model, we first trained the model 10 epochs with non-attacked dataset and did 50 epoch fine-tuning with adversarial dataset.

DAmageNet [9] was used as a black-box attacked sample and attacked MNIST with FGSM [10] was used as a white-box attacked sample.

The accuracy of the model for the adversarial data and the original data was measured by reducing the number of Gabor Filters in VOneBlock to 256, 128, 64, and 32, respectively. Through this work, we could find out how the number of filters affects the performance of the VOneNet.

German Traffic Sign Recognition Benchmark [11] (GTSRB) dataset was used as a real-life attacked sample. We scribbled on this dataset's image to express the real-life attack which can occur in real life, by displaying white or black lines at random locations in the traffic sign area.

3.1 ResNet

In this paper, ResNet is used as a CNN-based model. ResNet was presented in [3] and was developed with the aim of facilitating the training of deep neural network.

Especially, ResNet showed higher performance than other traditional CNN-based models [1], [2] in image classification problem and that's why we used this model in this paper.

For other reason, in this paper, we use fine-tuning method to compare the performance on adversarial image between models with and without VOneBlock, but Figure 3 shows that during fine-tuning, accuracy of those four different models, AlexNet, Basic-CNN, ConvNet, Basic linear regression, generally converges before reaching 30 epoch. In this paper, since we want to check the gradual change of accuracy of model with higher epoch number, we choose ResNet which its accuracy converges at relatively higher epoch numbers.

The problem of accuracy convergence in a small number of fine-tuning epoch became more pronounced when VOneBlock attached on the front-end of the model. Figure 4 shows that the accuracy of those models with VOneBlock converges much faster.

So, we set the highest number of fine-tuning epochs to 50, to see the gradual change of model's accuracy. And during fine-tuning, the accuracy of ResNet typically converge when the fine-tuning epoch reaches 50.

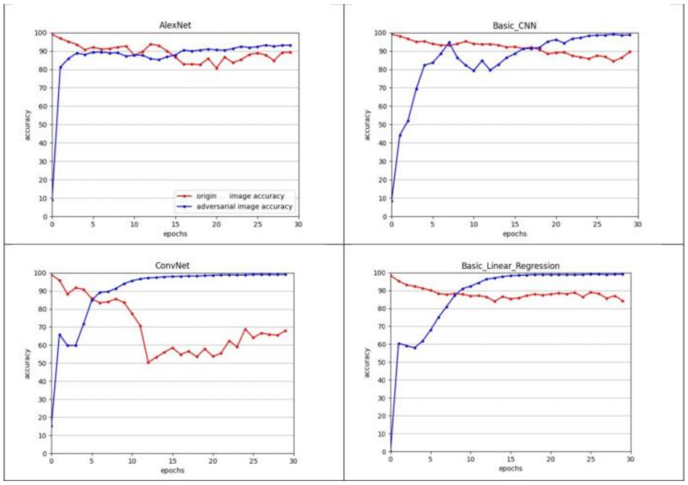


Figure 3: Accuracy of plane models when fine-tuning. Blue line stands for model's accuracy to original image, and red line stands for model's accuracy to adversarial image.

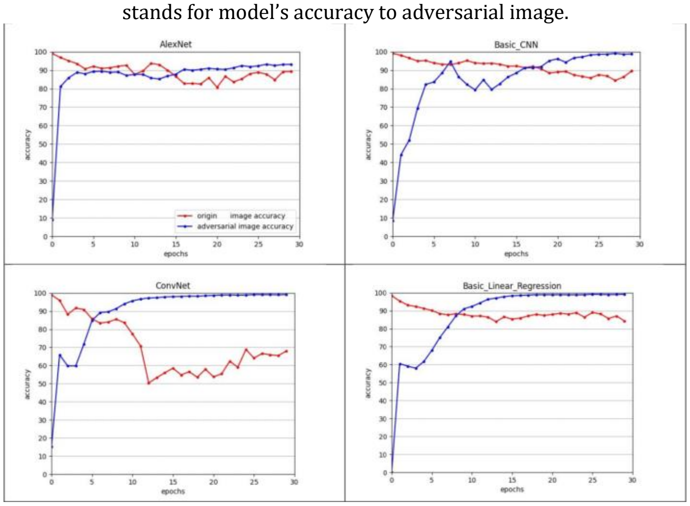


Figure 4: Accuracy of VOneBlock contained models when fine-tuning. Blue line stands for model's accuracy to original image, and red line stands for model's accuracy to adversarial image.

3.2 Dataset

3.2.1 ImageNet

ImageNet is a representative large-scale dataset and has been publicized in [7]. It contains a total of 1000 classes of images, and there are 1300 samples in each class.

3.2.2 MNIST

MNIST is a simple computer vision dataset consisting of 70,000 binary images of (28×28) size [8]. Each image is a numeric handwriting image between 0 and 9, and this dataset is currently a “hello world” in the field of computer vision.

3.2.3 DAmageNet

DAmageNet [9] is a set of samples subjected to black-box attack on ImageNet [7] dataset, with a high error-rate for pre-trained classification networks. Since it is a dataset with black-box attack on ImageNet, it has 1000 classes the same as ImageNet, but contains 96,020 samples in total because it is a sample that performed an attack on ImageNet's validation set.

3.2.4 GTRSB (German Traffic Sign Recognition Benchmark)

GTSRB dataset was publicized in [11] as a large-scale multi-class dataset. It contains more than 50,000 samples in 43 classes of German traffic road sign. Traffic signs are objects that can easily be corrupted in real-life and can affect autonomous vehicles that control vehicle operation by recognizing road signs, so we used this dataset for our experiment

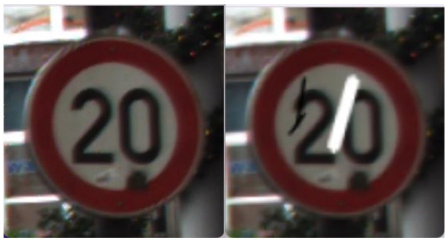
3.3 Number of kernels in Gabor Filter

In this paper, when the number of Gabor Filters, one of the components of VOneBlock, changes to 256, 128, 64, and 32 for simple & complex channels, we compare accuracy for adversarial image and clear image while we fine-tune both VOneNet and non-VOneBlock network with FGSM attacked data.

3.4 Real-life attack method

In this paper, there are two attack methods to be performed on the dataset: an adversarial attack that is invisible to the human eye, and a real-life attack that is visible to the human eye. In the case of a real-life attack, a dataset of objects with corruption that may occur in real life was required. Therefore, using GTSRB dataset, we could consider graffiti that could occur on the traffic road sign as a real-life attack, and performed random graffiti on the dataset image to generate a real-life attacked dataset.

The images in the GTSRB dataset is all different in size, but in the image, the traffic road sign is usually located in the middle. So, we resized the image to [90×90] and fixed the traffic road sign area as a center area of the image which is our ROI. In ROI, we draw some random numbers of straight lines with random size, color, and location to generate the real-life attacked data. Figure 5 shows original GTSRB image and scribbled image that we created.



(a) (b)

Figure 5: (a) Original GTSRB image, (b) Attacked GTSRB image.

3.5 Fine-tuning with attacked data

The pre-trained network was trained once again on adversarial images with smaller learning rate. This procedure is called fine-tuning. Using this method, we could maintain the accuracy for the original image and increase the accuracy for the adversarial image of network.

Both the network to which VOneBlock is connected to the front-end and the original network were pre-trained on the original image by 10 epochs with learning rate was 1e-3.

After training, we fine-tuned both networks with adversarial dataset. During fine-tuning, we recorded the accuracy of the network for the original dataset and the adversarial dataset in each epoch

4 EVALUATION

4.1 Fine-tuning with DAmageNet

4.1.1 ResNet18

Table 1 shows how the accuracy of the ResNet18 network for adversarial and clear images changes as the fine-tuning epoch increases.

TABLE 1: Resnet18 Accuracy

Fine-tuning epoch	Adversarial image accuracy	Clear image accuracy
0	5.87%	57.68%
10	25.46%	34.39%
20	31.17%	35.02%
30	34.83%	35.75%
40	37.73%	36.43%
50	40.18%	36.99%

The accuracy of adversarial images steadily increases by progressing fine-tuning, but the accuracy for clear images decreases by up to 23.39% and increases by 2.6% by repeating fine-tuning.

4.1.2 VOneResnet18

Table 2 shows how the accuracy of the VOneResNet18 network for adversarial and clear images changes as the fine-tuning epoch increases

TABLE 2: VOneResNet18 Accuracy

Fine-tuning epoch	Adversarial image accuracy	Clear image accuracy
0	11.12%	54.13%
10	24.37%	43.56%
20	29.04%	43.23%
30	32.04%	43.41%
40	34.11%	43.83%
50	35.83%	44.31%

The accuracy of adversarial images steadily increases by progressing fine-tuning, and the accuracy for clear images decreases by up to 10.9% and increases by 1.08% by repeating fine-tuning.

Figure 6 is a graph of accuracy per epoch during fine-tuning of VOneResnet18 and Resnet18 and Figure 7 is a graph of loss per epoch during fine-tuning of VOneResNet18 and ResNet18. The blue line represents the Resnet 18, and the red line represents the VOneResnet18.

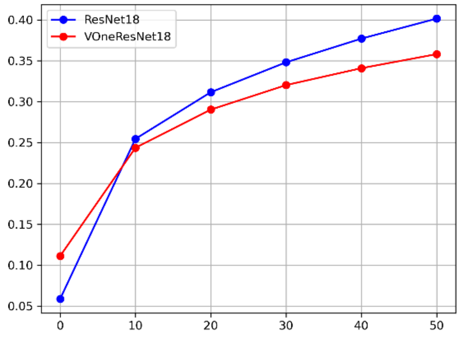


Figure 6: Accuracy to Adversarial Image per epoch

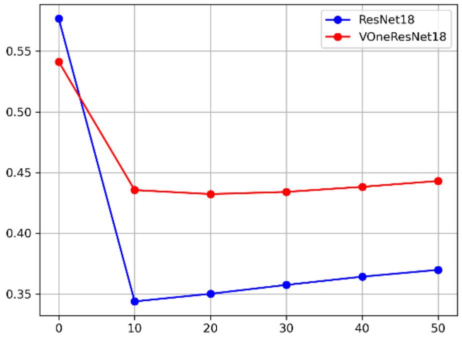


Figure 7: Loss to Adversarial Image per epoch

For fine-tuning, the network that has been connected with VOneBlock can be expected to be more robust.

4.2 Fine-tuning with MNIST FGSM attacked data

For MNIST dataset, validation was performed with the basic image size which is (28×28). VOneResNet18 was used, and fine-tuning was performed on the FGSM attacked MNIST. We progressed experiment reducing number of Gabor Filters from 256 to 32.

We set the number of VOneBlock's simple & complex channels of Gabor Filters to 256, 128, 64, 32 and the accuracy for the test dataset was measured for each epoch during fine-tuning with FGSM attacked MNIST. Figure 8 shows the change in each model accuracy according to the fine-tuning epoch.

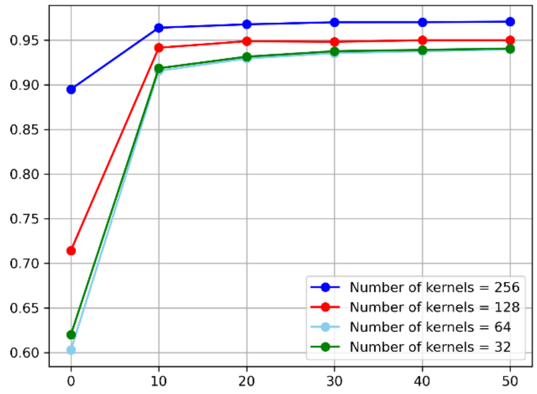


Figure 8: Accuracy about adversarial image on different number of kernels

4.3 Real-life attack with GTSRB

By drawing a random line on the images of GTSRB dataset, we created a dataset that received a real-life attack. The Original GTSRB dataset was used for ResNet50 and VOneResNet50 training, respectively, 50 epochs. After 50 epochs training, we measured accuracy of original GTSRB and attacked GTSRB for both trained networks. Images were all resized to (32×32) and used as inputs of the network. Table 3 shows the accuracy of ResNet50 and VOneResNet50 after training.

TABLE 3: Network Accuracy for GTRSB

Dataset	ResNet50 Accuracy	VOneResNet50 Accuracy
Original Image	99.65%	99.82%
Attacked Image	60.13%	68.08%

We can see that the VOneResNet50 shows a slightly better performance on attacked image than ResNet50.

5 DISCUSSION AND FUTURE WORKS

5.1 Discussion

Result of fine-tuning ResNet18 with DAmageNet, without VOneBlock attached, shows that repeat of fine-tuning increases the accuracy for adversarial image but not for clear image. On the other hand, result of fine-tuning VOneResNet18 shows that repeat of fine-tuning not only increases the accuracy for adversarial image but also does the same for clear image. In other words, fine-tuning is sufficient to enhance the robustness of the network for adversarial images, but the base network is adapted to the adversarial image thus confuses the clear image. And as you can see that VOneBlock prevent this causing problem.

Figure 8 shows the performance of the network that varies depending on the number of simple and complex Gabor Filters of VOneBlock when fine-tuning with FGSM attacked MNIST data. It is noteworthy that when the number of epochs of fine-tuning is 0, the performance of VOneResNet18 with the number of filters set to 256 is significantly higher than the performance of VOneResNet18 with the number of filters reduced to 128, 64, and 32. That is, the higher number of Gabor Filters in VOneBlock, the more robust the model is for adversarial images even before fine-tuning with adversarial images. And as you can see in Figure 8, the accuracy of model does not change when the fine-tuning epoch reaches 50.

5.2 Future works

Although the performance of VOneNet on real-life attacked data was better than the non-VOneBlock network, the difference in performance was not as high as 8%. There is clearly room for performance improvement with algorithms such as fine-tuning or etc. In addition, in the case of real-life attacks, performance between human and AI network should also be compared because there are some data corrupted so much that is difficult to recognize even in human eyes.

ACKNOWLEDGMENTS

This work was conducted with the support of the National Research Foundation of Korea (NRF-2020RIA2C1101938)

REFERENCES

[1] Krizhevsky, A., Sutskever, I. and Hinton, G.E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM. Association for Computing Machinery (ACM).

[2] Simonyan, K. and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv.

[3] He, K., Zhang, X., Ren, S. and Sun, J. 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[4] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. 2013. Intriguing properties of neural networks. arXiv.

[5] Nguyen Minh Trieu and Nguyen Truong Thinh, "A Study of Combining KNN and ANN for Classifying Dragon Fruits Automatically," Journal of Image and Graphics, Vol. 10, No. 1, pp. 28-35, March 2022.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv.

[7] Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D.D. and DiCarlo, J.J. 2020. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. Cold Spring Harbor Laboratory.

[8] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li and Li Fei-Fei 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE.

[9] Li Deng 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. IEEE Signal Processing Magazine. Institute of Electrical and Electronics Engineers (IEEE).

[10] Chen, S., Huang, X., He, Z. and Sun, C. 2019. DAmageNet: A Universal Adversarial Dataset. arXiv.

[11] Goodfellow, I.J., Shlens, J. and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. arXiv.

[12] Stallkamp, J., Schlipsing, M., Salmen, J. and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural Networks. Elsevier BV.

[13] Hubel, D.H. and Wiesel, T.N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology. Wiley.

[14] Patrick, D., Geyer, M., Tran, R. and Fernandez, A. 2022. Reconstructive Training for Real-World Robustness in Image Classification. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). IEEE.

[15] Hendrycks, D. and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv.

[16] Samangouei, P., Kabkab, M. and Chellappa, R. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. arXiv.