International Journal of Advanced Smart Convergence Vol.8 No.1 141-146 (2019)
http://dx.doi.org/10.7236/IJASC.2019.8.1.141

IJASC 19-1-17

High Representation based GAN defense for Adversarial Attack

Richard Evan Sutanto[1], Suk Ho Lee[2]

[1, 2]*Dept. of Computer Engineering, Dongseo University*
[1]*Richardwenz91@gmail.com,* [2]*petrasuk@gmail.com*

Abstract

These days, there are many applications using neural networks as parts of their system. On the other hand, adversarial examples have become an important issue concerining the security of neural networks. A classifier in neural networks can be fooled and make it miss-classified by adversarial examples. There are many research to encounter adversarial examples by using denoising methods. Some of them using GAN (Generative Adversarial Network) in order to remove adversarial noise from input images. By producing an image from generator network that is close enough to the original clean image, the adversarial examples effects can be reduced. However, there is a chance when adversarial noise can survive the approximation process because it is not like a normal noise. In this chance, we propose a research that utilizes high-level representation in the classifier by combining GAN network with a trained U-Net network. This approach focuses on minimizing the loss function on high representation terms, in order to minimize the difference between the high representation level of the clean data and the approximated output of the noisy data in the training dataset. Furthermore, the generated output is checked whether it shows minimum error compared to true label or not. U-Net network is trained with true label to make sure the generated output gives minimum error in the end. At last, the remaining adversarial noise that still exist after low-level approximation can be removed with the U-Net, because of the minimization on high representation terms.

Keywords: *Neural networks, Adversarial examples, Generative adversarial network, Adversarial attack, Adversarial defense.*

1. Introduction

Nowadays, neural networks have become popular for their wide use in various field. Many applications in different fields are using neural networks to advance their system [1]. However, researchers become concern about the security of the networks regarding the handling of malicious inputs [2]. Some researches have shown that existing neural networks are fragile to adversarial examples where it works as malicious input that makes the classifier in the networks works inaccurately [3]. Meanwhile, adversarial examples can be easily generated by presenting or modifying the image input. Adversarial examples are not distinctive for human where human

Manuscript Received: Feb. 8, 2019 / Revised: Feb. 14, 2019 / Accepted: Feb. 19, 2019
Corresponding Author: petrasuk@gmail.com
Tel: +82-51-320-1744, Fax: +82-51-327-8955
Dept. Computer Engineering, Dongseo University

still can identify the object correctly without any intervention, but not for neural network. There are two parts of adversarial examples research; attack parts and defense parts. There are two categories for each parts; black box and white box based approach. In attack parts, a black box attack means that the attacker has no information at all about the targeted network, while in white box attacks the attacker has full information about the targeted network. One of the most common attack method is Fast Gradient Sign method (FGSM), it is using sign of the gradient loss function to decide the direction where the corresponding pixel value want to be changed [3]. For defense parts, black box category means the defender needs to protect the targeted network from any kind of attack methods without any information about the attack method, while white box category means the defender need to protect the targeted network from a specific attack method. There are many approaches to defend the network, i.e. fine-tuning input images before going through the network [2], adjusting the training data to make the classifier aware with adversarial examples [4] and one of the most popular defense method is Adversarial Training. It uses Adversarial Examples as training dataset together with the clean original dataset to make the classifier more robust against Adversarial Examples [4]. If the attacker use different attack method, Adversarial Training need to include more data into the training dataset and it will cost more. Furthermore, it is better to make sure that input images are regulated to be kept clean as the original clean images.

In this paper, we propose a research that combines the Generative Adversarial Network (GAN) with the U-Net network to apply high representation levels in the classifier in the reducing the effect of adversarial examples. The GAN network consists of two neural networks that collaborate with each other. The first network is called the Generator and the second network the Discriminator [5]. Both the Generator and the Discriminator are trained at the same time, until the Generator can match the data distribution, while the Discriminator becomes able to distinguish between the generated data and the real data. The U-net network is using a U-shaped architecture where it consists of contracting paths and expansive paths that are able to work with few training images and produces better segmentations [6]. The main part in U-net is the up-sampling part where it have many feature channels that allow the network to be able to broadcast the information to upper layers. However, a similar approach has proposed to use the GAN to protect the classifier from Adversarial Examples in [7], which is called the Defense-GAN. It trains the network with original images to make sure the generator able to cleanse input images before they go through to the classifier. By adding the GAN reconstruction loss minimization step, it will help to reduce the Adversarial Examples effect because Adversarial Examples will lead to different distribution than the GAN training examples. It shown that Defense-GAN able to be an effective method to defense Adversarial Examples, by producing cleaner image before it is going through a classifier. However, Adversarial Examples can survive approximation process in the low feature level. The difference between our work and the work in [7] is the adversarial noise can endure the approximation process in low feature levels. While in our approach, the minimizing of the loss function of the high representation terms minimizes the difference between high representation level of the clean data and the approximated output of the noisy data in the training dataset. The detail of our proposed methods is explained in Section 2b.

2. Method

Our work focuses on utilizing the GAN Network with the U-Net by minimizing the loss function of the high representation term, in order to get the minimum difference between high representation levels of clean data and approximated output of the noisy data in the training dataset. One of the advantages of the Generative Network (GAN) is that the generator learns to produce a new data that is close to the important feature from real data, while the discriminator learns to differentiate a bad generated data that looks not like the real data. Because of that advantage, the GAN network is able to produce images which exclude the adversarial noise in some extent. By using a similar architecture as in Defense-GAN explained in [7], the network finds the most suitable seed (z^*) among all random seed (z) to make the generator (G) produce a generated data as close to the original clean images (x) as possible by minimizing the following distance:

$$\min||G(z) - x||_2^2 \tag{1}$$

After the most suitable seed (z^*) is acquired by the minimization problem in Eq. (1), the Generator (G) produces a generated data $(\hat{x})$ that is close enough with the original clean images (x), i.e.:

$$\hat{x} = G(z^*) \tag{2}$$

At the second stage, the discriminator network is trained by using the original clean image to reduce the effect of Adversarial Examples. The Discriminator network checks whether a generated data $(\hat{x})$ from generator network as shown in Eq. (2) looks similar to the real data or not $(\hat{x} \approx x)$. Then, the classifier receives the generated data $(\hat{x})$ and passes it through the discriminator and classifies them by comparing with the true label (x_t). In order to make sure that the adversarial examples no longer exist after the approximation in low representation level, each generated data is passed through a classifier are going through U-Net network (U). In the U-Net network (U), contracting side of the network is downsampling every input to make sure there is no more adversarial examples exist in the data. Then, the expansive side of the U-Net network (U) is going to upsample each data to pass through classifier once again to check with the true label with an approach:

$$\min||f(U(\hat{x})) - f(x_t)||_2^2 \tag{3}$$

In Eq. (3), $\hat{x}$ defines the generated data from generator where it is become the input of U-Net network $(U(\hat{x}))$. The main goal in our approach is to minimize the different of classification result $(f(U(\hat{x})))$ from U-Net network $(U(\hat{x}))$ with the classification of true label $(f(x_t))$. By applying this method, the distance from output of U-Net network and the true label (x_t) can be as small as possible so the output still look similar with original data and it can further reduce the adversarial examples that still exist after the low-level approximation. The new generated output that comes from the U-Net network will be checked again through classifier to re-check the result. Figure 1 shows the whole diagram of the proposed method.

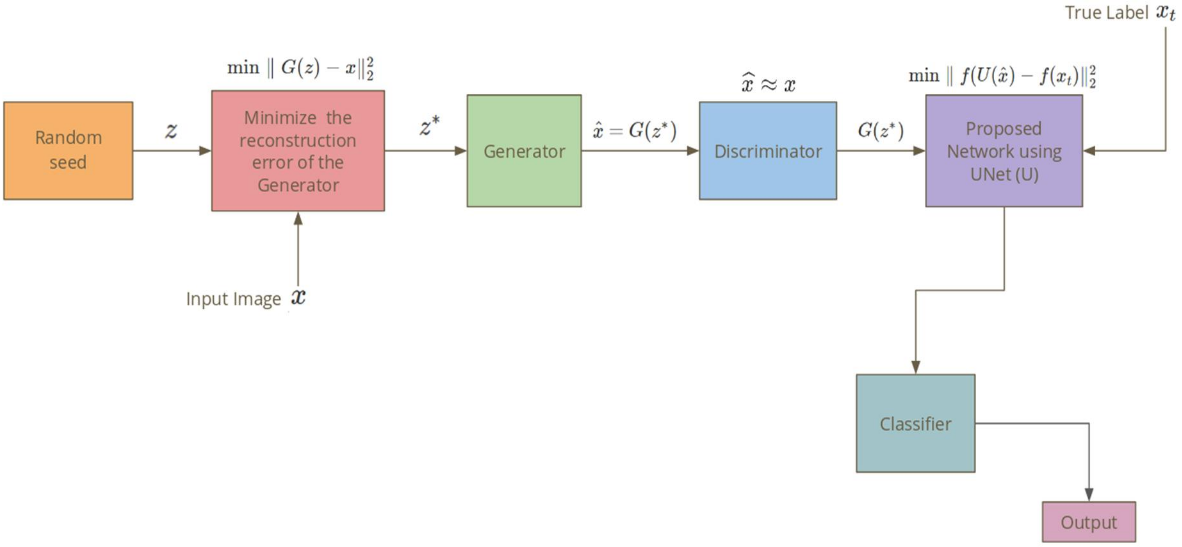

Figure 1. Proposed Method Flowchart

3. Experiments

In this research, the dataset that used is MNIST dataset. A dataset that consist of handwritten digits with 60,000 images of training set and 10,000 images of test set [8]. It is using black and white images, which has only one channel with 28x28 width-height dimension, and has 10 classes in the dataset. We used the Fast Gradient Sign Method (FGSM) as the adversarial attack method in this experiments. This method is using a sign of the loss function to decide how much values want to be changed, and there is a variable used in FGSM called *epsilon* where it works as a number that affect the noise [9]. For this experiment, in our approach we used *epsilon*= 0.3 for the MNIST dataset.

4. Results and Discussion

For each output data that denoised from U-Net network use true label, original label, noisy label as comparison. True label become a comparison in order to check the classification result of denoised data with the true class of original dataset. Original label and noisy label also become a comparison to check the classification result between de-noised data with clean data and noisy data. There are three classification percentages used to represent the result of our experiments. The percentage success of original (PSO) is a measure of percentages of successful classification of original images against the true label. This percentage show how much original images classified correctly by the classifier against with the true label. The percentage success of denoised (PSD) is a percentage of successful classification of denoised images. This percentage show how much denoised images classified correctly by the classifier against with the true label. Finally, the percentage comparison with original (PCO) is a measure of percentage of successful classification of denoised images against the original label. For this percentage, it shows how much denoised images classified against the original images in order to compare the quality of denoised images whether the classifier still able to recognize them similarly with original images.

From Table 1, it shows that percentage success of denoised (PSD) is not much different with percentage success of original (PSO). That means the result of denoised images still can be used where the classifier able to classified them correctly with accuracy 90.69%. With that amount of accuracy, our result also present another comparison between denoised images and original images. Percentage comparison with original (PCO) is 96.63%, this percentage represent how much denoised images classified equally with original images. That percentage shows that denoised images from our proposed method able to be classified accurate enough compared with original images, and it also still represent similar feature with original images.

Table 1. Percentage of classification accuracy

PSO	92.73182957
PSD	90.69896965
PCO	96.63046505

Figure 2 shows the visualization comparison between each images; original, noisy, generated data from GAN network, and generated data from U-Net network. The denoised images are a little different from the original images. However, these images give the same classification result as the original images, while the images in row (b) give different classification results. This is due to the fact that we trained the neural network to have similar classification results and not to have visually similarities. So, even though the final results looks noisy, the accuracy is improved.

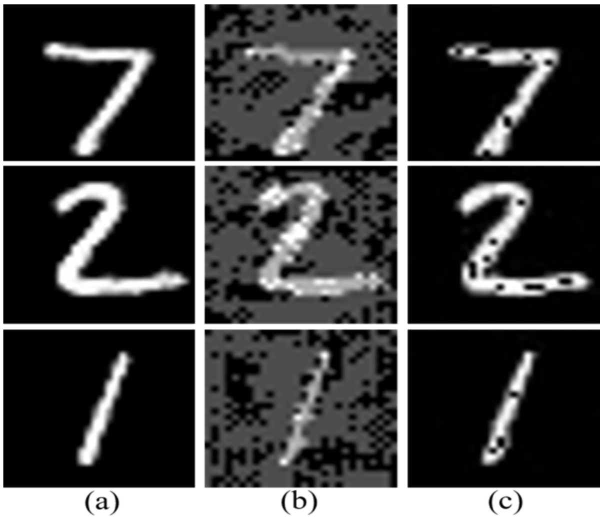

Figure 2. Image Comparison: (a) Original images (b) Images with adversarial noise (c) Denoised with proposed method

5. Conclusion

We propose a defense method for adversarial attack by using high representation based. This approach pay attention to high representation level in order to eliminate any adversarial example that still exist after low representation level approximation. By utilizing the Generative Adversarial Network (GAN) and optimize the advantage of it where the network consist with two networks that work simultaneously to create new data that look similar as original dataset. In addition, we proposed to combine the high representation based denoising with the U-Net network as an additional denoising platform which further transforms the output of the GAN network so that the effect of the noise in the output of the GAN gets further removed by the UNet network. Experimental results show that the proposed method is effective in denoising the adversarial noise and enhances the accuracy of the network against adversarial examples.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (No. NRF-2016R1D1A3B03931875) and by the Institute for Information and Communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2018-0-00245, Development of prevention technology against AI dysfunction induced by deception attack).

References

[1] A.S. Rakin, Z. He, B. Gong, and D. Fan, “Blind Pre-Processing: A Robust Defense Method Against Adversarial Examples”. arXiv preprint arXiv:1802.01549, 2018.

[2] C. Guo, M. Rana, M. Cissé, and L.V. Maaten, “Countering Adversarial Images using Input Transformations”. arXiv preprint arXiv:1711.00117, 2017.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv preprint arXiv:1412.6572, 2014.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks”. arXiv preprint arXiv:1312.6199, 2013.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative Adversarial Nets”. Neural Information Processing System (NIPS), 2014.

[6] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation”. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015.

[7] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models”. arXiv preprint arXiv:1805.06605, 2017.

[8] L. Deng, “The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]”. IEEE Signal Processing Magazine, Vol 29, pp 141-142, 2012.

[9] Perhaps the Simplest Introduction of Adversarial Examples Ever.
https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d