Journal of KIIT. Vol. 19, No. 8, pp. 67-74, Aug. 31, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.8.67

Advanced Generative Adversarial Defense Model using Instance Sparsity Mapping Network

Md Foysal Haque*, Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NO.2017R1D1A1B04030870)

Abstract

The deep learning method has shown tremendous results in different real-life applications, including autonomous cars, video surveillance systems, and medical image examination systems. Deep convolutional neural networks and adversarial learning are attributed to achieving these breakthroughs. However, adversarial networks are vulnerable to adversarial perturbations. The adversarial perturbations affect the unsupervised learning process and considerably decrease the performance. That can be lead to devastating consequences, particularly for critical applications that using for safety and security purposes. Adversarial attacks crafted perturbations that create the optical illusion. It forces the model to misclassify the samples and mistake to accomplish the desired task. For that reason, it is crucial to apply defense solutions that increase the robustness of adversarial networks against adversarial attacks. To overcome this issue, a defense solution introduce by adopting the instance sparsity mapping network. The proposed network is learned from real to distinguish the synthetic images with perturbations, and a perturbation classifier is individually trained to classify the transform of the input-output difference. It performs as a defense model, where the adversarial noise of the generated samples is fused. The defense model combined with a classifier to examine the attacking level. The proposed model provides defense against adversarial attacks to achieve high robustness to fusing adversarial attacks.

요 약

딥 러닝은 자율 자동차, 비디오 감시 시스템 및 의료 이미지 검사 시스템을 포함한 다양한 실제 애플리케이션에서 엄청난 결과를 달성할 수 있음을 보여주었다. 심층 컨볼루션 신경망과 적대적 학습은 이러한 돌파구를 달성하는 데 기여했다. 그러나 적대적 네트워크는 적대적 혼란에 취약하다. 적대적 혼란은 비지도 학습 과정에 영향을 미치며 성능을 크게 감소시킨다. 이는 특히 안전 및 보안 목적으로 사용하는 중요한 애플리케이션의 경우 엄청난 결과를 초래할 수 있다. 적대적 공격은 착시현상을 일으키는 혼란을 만들어 낸다. 이 경우 모델이 샘플을 잘못 분류하고 실수를 통해 원하는 작업을 수행할 수 없다. 그러한 이유로 적대적 공격에 대한 적대적 네트워크의 견고성을 높이는 방어 솔루션을 적용하는 것이 중요하다. 이 문제를 극복하기 위해 적대적 희소성 매핑 네트워크를 채택한 방어 솔루션을 제안한다. 제안된 네트워크는 실제 데이터를 학습하여 공격받아 변형된 이미지를 구별하고, 공격 여부 분류기는 독립적으로 입출력 차이의 변환하여 분류하도록 훈련된다. 생성된 샘플의 적대적인 노이즈가 융합된 경우 방어 모델로 수행한다. 방어 모델이 분류기와 결합되어 공격 수준을 검토한다. 제안된 모델은 적대적 공격의 융합에 대한 높은 견고성을 달성하기 위해 적대적 공격에 대항하는 방어 수단을 제공한다.

Keywords
generative adversarial network, adversarial attacks, image generation

* Dong-A University Electronic Engineering
- ORCID: https://orcid.org/0000-0003-0634-9783
** Dong-A University Electronic Engineering professor
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Jul. 12, 2021, Revised: Aug. 17, 2021, Recepted: Aug. 20, 2021
· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University,37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. Introduction

Adversarial networks are applying to diverse areas to solve complex vision tasks. Recently, deep learning bases supervised, and unsupervised learning models are used to understand human mids to improve the humanoid robot and autonomous applications. Due to the popularity of adversarial models in many real-world applications, mainly for security-sensitive applications, this advances an essential issue as to their strength against adversarial attacks. Adversarial models are vulnerable to quasi-imperceptible perturbations. For example, adversarial examples produced noise or perturbations that can cause inaccurate prediction[1][2].

(a) Orginal image

(b) Adversarial image

(c) Misclassified segmented image

Fig. 1. The source image is perturbed with universal adversarial noise, and the resulting prediction is misclassified due to the noise and unable to segment all the information[3].

Moreover, a small perturbation on inputs can mislead the model. More clearly, considering Fig. 1, the deep learning model unable to segment a specific portion of the image that included adversarial perturbations.

Different strategies are adopted to defend against adversarial examples. First, to assure that networks are strong against adversarial examples, second is to transform adversarial examples, and the last one is to distinguish adversarial examples. One of the most effective strategies is to train the adversarial model with adversarial examples to obtain an adversarially trained targeted network concerning the first type. Nevertheless, all of these approaches have particular limitations. The first method is unable to effectively defend against adversarial attacks if it has not been trained about attacks. The second one inescapably has an impact on certain attacks as they would also be reformed. Finally, the third approach may decline the detected adversarial examples, which might be unacceptable in specific scenarios.

To overcome the drawbacks mentioned above, the proposed model introduces a defense framework to fuse the attacks and prevent misclassification.

Ⅱ. Related Theories

The robustness of adversarial networks against adversarial attacks has achieved notable attention in the recent deep learning applications[4]-[8]. As a result, adversarial networks became the core attention in this field to prevent and eliminate attacks.

2.1 Adversarial Attack

Adversarial attacks are two types: first, engineered attacks outlined by hackers, and self-generated attacks occurred during training time. Both attacks are vulnerable for deep learning models, mainly for the adversarial network, because adversarial networks are

constructed with complex architecture to achieve their predefined goal. The network uses latent noise to generate images, and in most cases, adversarial examples are originated during adversarial training time. Furthermore, recently, adversarial models apply in a wide range of social and security applications, and this model is targeted to mislead by adversarial attacks. In Fig. 2 an example of adversarial attacked shown that fooled by adversarial attack.

Adversarial attacks that generate adversarial examples in linear models can be explained due to high-dimensional dot products or noise.

$$x^{adv} = x + \epsilon \cdot sign(\nabla_x j_\theta(x,y)) \quad (1)$$

In Eq. (1), $(\nabla_x j_\theta(x,y)$ is the gradient of the loss function j_θ of the learning model parameterize by Θ in x, and true label is y.

When a perturbed image is input, the output will be the sum of the dot product of the weights and the unperturbed information and the dot product of the weights and specific perturbation. Each weight vector holds average magnitude m, and the input is n-dimensional. Therefore, the activation of the classifier will grow by mn(where is the magnitude of the perturbation). This does not result in a significant difference in classification between the original and attacked data for low-dimensional problems. Despite, if the input has a large dimension n, even weak perturbations can cause large errors in the classifier[9].

Rely on the adversary knowledge attacks can be classified into three categories such as white-box attacks, black-box attacks, and gray-box attacks. These attacks directly access the deep learning model to harm the target task[10].

In this category, the white-box attack[11] conducts on the model architecture and parameter, while in black-box attacks has no access to model parameters. Instead, it utilizes a different model to generate adversarial images assuming that influence on the target model. Finally, the gray-box attack can take over full access to model parameters but unaware of the model defense system.

The proposed network is constructed with an instance sparsity mapping network that is associated with a classifier to fuse the adversarial attacks. The instance sparsity mapping network prevents the attacks and leads the base network to generate a perfect image with exact order.

III. Proposed Algorithm

The proposed network applies a defense strategy to eliminate adversarial attacks. The defense model is constructed with the combination of a classifier and mapping network.

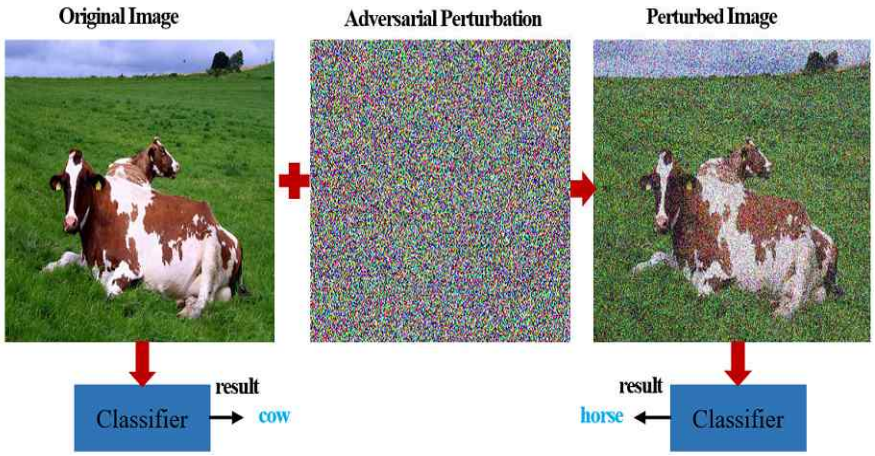


Fig. 2. Adversarial attack to fool neural network

The architecture of the proposed network is shown in Fig. 3. The defense network adopts two core techniques to fuse the internal and external attacks, i.e., Adversarial training and defensive distillation. These techniques are directed to balance the network and improve the robustness against adversarial attacks.

In the initial stage, the model is constructed by considering the adversarial training. In this block, the network is incorporated with the approach of a conventional adversarial network. The adversarial network conducts the operation by balancing the adversarial distribution between the generator block and discriminator block.

The adversarial training method includes adversarial examples from the attacks during the training phase, so the discriminator is trained either on a mix of both clean and adversarial examples.

However, despite the adversarial examples generated with a specific type of attack, the discriminator remains weak to other types of attacks. Therefore, to eliminate the issue and improve the defense method, the proposed network included a defense distillation technique.

$$x^{adv} = \hat{x} + (\epsilon - f\phi)\nabla_{\hat{x}} j_\theta(x, y) \tag{2}$$

From Eq. (2), ϕ is the noise factor and $\hat{x}$ is a random vector.

The model combined with a mapping network and a classifier to improve the robustness of the proposed network. In this part, the classifier examines the generated samples to justifies the sample is attacked or not. It assumes a probabilistic classifier that maintains the network to develop a defense against adversarial attacks.

The instance sparse mapping network trained with both clean and attacked image samples that to ensure that the adversarial attacks do not bias the change learned by the network.

$$j_\theta = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - f\phi(x_i^{'})\right) \tag{3}$$

In Eq. (3), n is the number of samples and x' is the perturbed version of x.

The proposed instance sparsity mapping network detects the classifier noise and sends the generator network to feedback to regenerate the sample again along with association with the network. This process continues until the generated sample completely matches the source data.

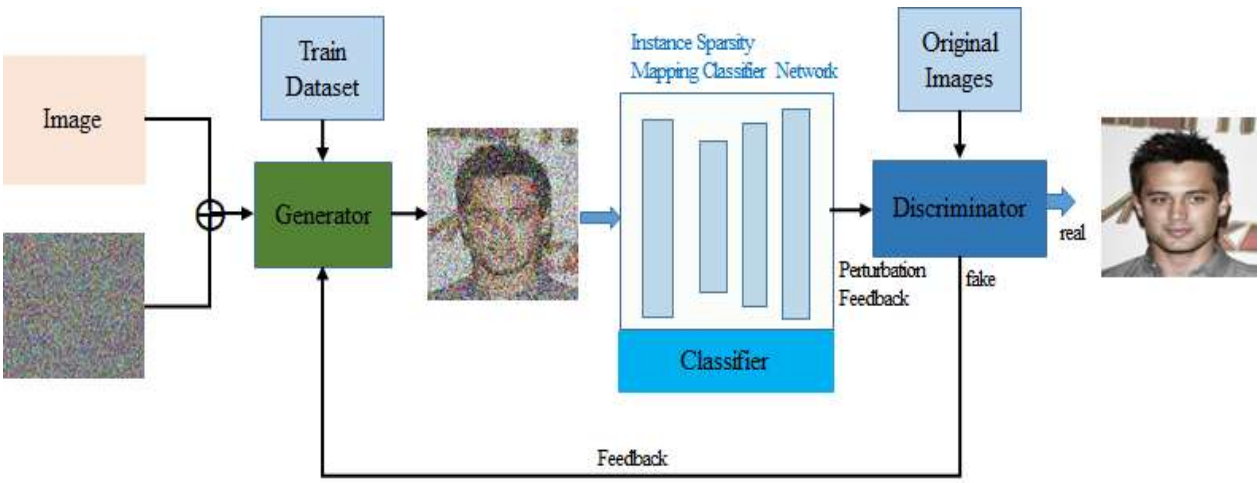


Fig. 3. Network architecture of instance sparsity mapping network

IV. Experiments

4.1 Baseline Architecture

The proposed network is constructed with an adversarial network and a defense block known as the instance sparsity mapping network. Improved MagNet [12] architecture use to construct the instance sparsity mapping network. In addition, the network adopts adversarial learning and defense distillation model to fuse the adversarial attacks the defense model constructed with the combination of classification block and autoencoder. Eventually, both of model combines to generate image samples by easing influence of adversarial attacks. The architecture of the classifier is shown in Table 1, and the architecture of the instance sparsity mapping network is shown in Table 2.

Table 1. Neural network architectures used for classifiers

Classifier
Conv2D(128, 3,3)
ReLU
Conv2D(256, 3,3)
ReLU
Maxpolling(2,2)
BN
Conv2D(128, 3,3)
ReLU
Conv2D(64, 3,3)
ReLU
Dense(1024)
ReLU
Softmax

Table 2. Neural network architectures of instance sparsity mapping network

Layer
Conv(64, 7 × 7,2)
LeakyReLU(0.2)
Conv(128, 3 × 3, 2)LeakyReLU(0.2)
Conv(256, 3 × 3, 2)
LeakyReLU(0.2)
Conv(256, 5 × 5, 2)
Fully Connected (128)

4.2 Experimental Environment

Evaluating the network performance against adversarial attacks to generate the multi-resolution image trained with two different datasets for low-resolution image CIFAR-10[13] used and high-resolution image CelebA dataset[14] used. The CelebA dataset including 19999 images for the test set, test dataset includes 2000 images randomly collected from the dataset. At first, preprocess the images to 512×512.

Moreover, the defense module implied tested by different adversarial attacking methods and levels. For test conditions, adversarial attacks externally transfer to the base network during the adversarial learning to evaluate the defense ability.

V. Results

The evaluation of the proposed defense network is conducted by varying the attacking levels. The adversarial model is aware of the defense framework and incorporates them in generating attack-free examples. We consider two types of dataset to test the defense performance for the multi-resolution image generation task. The evaluation result on the result on CIFAR-10 is shown in Fig. 4, and the result the CelebA dataset is shown in Fig. 5.

In terms of the proposed adversarial model defense itself, the success rate of a fusing attack is higher than the robust defense model. Low-resolution images are vulnerable to attack, and it is tough to generate images by fusing adversarial attacks. On the other hand, fusing attacks in the high-resolution image is also tricky because if the classifier network and discriminator network are unable to trace the difference between the original image and adversarial examples, then the quality of generation examples will be decreased.

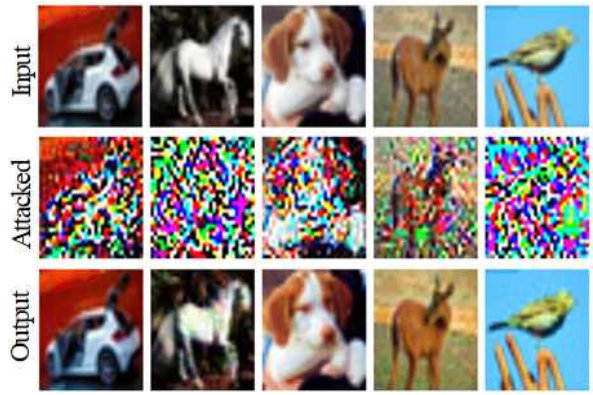


Fig. 4. Evaluation results on CIFAR-10 dataset

Fig. 5. Evaluation results on CelebA dataset

In CelebA dataset evaluation task, 2000 images were used in the test dataset, with the perturbation σ= 0.2 to 0.4. Analogously, the proposed model tested to fuse adversarial examples with perturbations [0.1, 0.2, 0.3, 0.4] to attack methods with full access to the adversarial model. The defense performance of the proposed method is shown in Tables 3, and 4. The conventional defence models are only specified for low or high resolution dataset. Hence, the proposed model trained and evaluated in multi-resolution images to fuse the adversarial attacks and the successive rate is higher than the conventional defence model.

As shown in Table 4, the proposed model showed an enormous rate to fuse the attacks on σ = 0.4 from 86.02% to 89.04% and on σ = 0.3 from 92.7% to 93.6%. Moreover, the success rate is showed more prominent in low-level attacks for both dataset.

Table 3. Classification accuracy based on different attack level on CIFAR-10 dataset

Attacking level	No defence (Proposed model)	Proposed model	MagNet [12]	DDSA [15]
0.2	0.482	**0.936**	0.792	0.895
0.3	0.374	**0.894**	0.688	0.853
0.4	0.268	**0.871**	0.564	0.816

Table 4. Classification accuracy based on different attack level on CelebA dataset

Attack level	Dataset	No defence (Proposed model)	CelebANet [16]	Proposed model
FGSM (0.3)	CelebA	0.624	0.912	**0.927**
PGD (0.3)		0.597	0.923	**0.936**

VI. Conclusions

We introduce a defense framework containing three modules: adversarial example identification, adversarial perturbation fusing, and adversarially trained targeted network. Specifically, if an input sample is detected to be adversarial, the sample is cleaned by an instance sparsity mapping network and then analyzed by the adversarially targeted network. As a result, the proposed model achieved successive scores to protect the target model against adversarial attacks. In future work, the proposed adversarial defense framework extends to apply real-world applications, such as face recognition.

Acknowledgement

This paper is an extension of the "Adversarial

Sparsity Mapping Network to Defense Against Adversarial Attacks" published at the Proceedings of Korean Institute of Information Technology Comprehensive Academic Conference in 2021.

References

[1] N. Papernot, et al., "The Limitations of Deep Learning in Adversarial Settings", In Proc. IEEE Eur. Symp. Secur. Privacy(EuroS&P), pp. 372-387, Mar. 2016, http://dx.doi.org/10.1109/EuroSP.2016.36.

[2] Ian J. Goodfellow, Jonathan Schlens, and Christian Sczegedy, "Explaining and Harnessing Adversarial Examples", Machine Learning (stat.ML), Vol. 1, Dec. 2014, https://arxiv.org/abs/ 1412.6572.

[3] J. H. Metzen et al., "Universal Adversarial Perturbations Against Semantic Image Segmentation", In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 2755-2764, Oct. 2017, http:// dx.doi.org/10.1109/ICCV.2017.300.

[4] Naveed Akhtar, Jian Liu, and Ajmal Mian, "Defense Against Universal Adversarial Perturbations", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, Vol. 3, pp. 3389-3398, Jun. 2018, https://doi.org/ 10.1109/CVPR.2018.00357.

[5] F. Liao, et al., "Defense Against Adversarial Attacks using High-level Representation Guided Denoiser", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, Vol. 2, pp. 1778-1787, Jun. 2018, https://doi.org/10.1109/CVPR.2018.00191

[6] G. Sun, et al., "Complete Defense Framework to Protect Deep Neural Networks against Adversarial Examples", Mathematical Problems in Engineering 2020, https://doi.org/10.1155/2020/8319249.

[7] J.H. Kim, et al., "Inconspicuous Adversarial Perturbation Post-processing Method with Image Texture Analysis", Journal of KIIT, Vol. 19, No. 2, pp. 89-95, Feb. 2021, https://doi.org/10.14801/ jkiit.2021.19.2.89.

[8] I. Park, et al. "Design of Robust Hyperspectral Image Classifier Based on Adversarial Training Against Adversarial Attack", Journal of Institute of Control, Robotics and Systems, Vol. 27, No. 6, pp. 389-400, 2021, http://dx.doi.org/10.5302/ J.ICROS.2021.21.0012.

[9] S. A. Fezza, et al., "Perceptual Evaluation of Adversarial Attacks for CNN-based Image Classification.", In 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), Berlin, Germany, pp. 1-6, Jun. 2019, https://doi.org/10.1109/QoMEX.2019.8743213.

[10] F. Vakhshiteh, et al., "Adversarial Attacks Against Face Recognition: A Comprehensive Study", Computer Vision and Pattern Recognition, Vol. 1, Jul. 2020, https://arxiv.org/abs/2007.11709.

[11] J. Kim,, "A White-box Implementation of SEED", Journal of Advanced Information Technology and Convergence, Vol. 9, No. 2, pp. 115-123, Feb. 2019, http://dx.doi.org/10.14801/ jaitc.2019.9.2.115.

[12] Dongyu Meng and Hao Chen, "Magnet: a two-pronged defense against adversarial examples", In Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, Dallas Texas USA, pp. 135-147, Oct. 2017, https://doi.org/10.1145/ 3133956.3134057.

[13] A. Krizhevsky, et al., "Learning Multiple Layers of Features From Tiny Images", Apr. 2009.

[14] Z. Liu, et al., "Large-scale Celeb Faces Attributes (celeba) dataset", 2018.

[15] Y. Bakhti, et al., "DDSA: A Defense Against Adversarial Attacks using Deep Denoising Sparse Autoencoder", IEEE Access, Vol. 7, pp. 160397-160407, Nov. 2019, https://doi.org/10.1109/ACCESS. 2019.2951526.

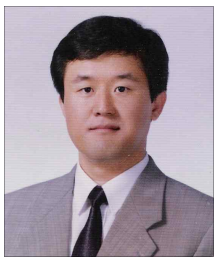
[16] Xiaowei Zhou, Ivor W. Tsang, and Jie Yin., "Latent Adversarial Defence with Boundary-guided Generation", Machine Learning (cs.LG), Jul. 2019, https://arxiv.org/abs/1907.07001.

Authors

Md Foysal Haque

2018 ~ 2020 : M.Sc. in Electronic Engineering, Dong-A University.
2020 ~ present : Ph.D. in Electronic Engineering, Dong-A University.
Research Interests : Digital Image Processing, Computer Vision and Pattern Recognition

Dae-Seong Kang



1994 : Ph.D. in Electrical Engineering, Texas A&M University.
1995 ~ present : Professor at Dong-A University.
Research Interests : Image Processing and Compression