Adversarial image detection based on the maximum channel of saliency maps[*]

FU Haoran, WANG Chundong[**], LIN Hao, and HAO Qingbo

Key Laboratory of Computer Vision and System, Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China

(Received 3 October 2021; Revised 25 November 2021)

©Tianjin University of Technology 2022

Studies have shown that deep neural networks (DNNs) are vulnerable to adversarial examples (AEs) that induce incorrect behaviors. To defend these AEs, various detection techniques have been developed. However, most of them only appear to be effective against specific AEs and cannot generalize well to different AEs. We propose a new detection method against AEs based on the maximum channel of saliency maps (MCSM). The proposed method can alter the structure of adversarial perturbations and preserve the statistical properties of images at the same time. We conduct a complete evaluation on AEs generated by 6 prominent adversarial attacks on the ImageNet large scale visual recognition challenge (ILSVRC) 2012 validation sets. The experimental results show that our method performs well on detecting various AEs.

Document code: A **Article ID:** 1673-1905(2022)05-0307-6

DOI https://doi.org/10.1007/s11801-022-1157-z

Although deep learning can solve large-scale complex problems very well, the security threats imposed by adversarial examples (AEs) are increasing[1]. SZEGEDY et al[2] found that deep neural networks (DNNs) can misclassify the images with high confidence, by adding certain hardly perceptible perturbations into the input samples[3].

Many defense strategies are proposed to defend against adversarial-example attacks[4]. SERBAN et al[5] divided them into ‘Guards’ and ‘defenses by design’ based on their places in the processing pipeline. The former does not interact with protected networks and constructs detectors around them. The latter directly modifies DNNs or training data to improve the robustness of models. The change of model architecture may inevitably lose accuracy or reduce generalization. And the change of training data may lose the information. Therefore, more and more researchers adopt the ‘Guards’ against AEs.

‘Guards’ are mainly divided into two categories. One is to utilize the statistical differences between AEs. GROSSE et al[6] applied a model-agnostic statistical test to evaluate the hypothesis that AEs are outside of the training distribution. Recently, KHERCHOUCHE et al[7] determined if the inputs are AEs or not through the natural scene statistics (NSS). Under the assumption that statistics of natural images are different from those of manipulated images, they built a binary classifier that takes as input features parameters of the generalized Gaussian distribution (GGD) and asymmetric generalized Gaussian distribution (AGGD). In contrast to other detection methods based on the statistical properties of inputs, NSS methods can achieve a higher detection rate. The other is to perform some transformations on inputs to restrict the space of AEs. Lots of researches focus on detecting AEs using the prediction inconsistency of the protected models[8]. XU et al[9] reduced the color depth of each pixel to squeeze features of images, and then adopted the prediction inconsistency strategy to detect AEs. Specifically, those inputs with great disagreement among the predictions of models can be detected as AEs.

However, the aforementioned defense methods only perform well in some limited situations. For ease of exposition, we start by classifying AEs into two categories based on the perturbation amount and the perturbation distribution. Some AEs are generated by adding large perturbations into original samples, which uniformly distribute on the whole image. We call them uniform-perturbation AEs (UAEs). The second class of AEs are crafted via introducing small perturbations into original samples, which often concentrate in some pixels of the image with great amounts relative to other pixels. Such AEs are known as the non-uniform perturbations AEs (N-UAEs).

Due to the large amount of perturbations in UAEs, it brings greater statistical differences. Therefore, these

[*] This work has been supported by the Joint Funds of the National Natural Science Foundation of China (No.U1536122), the Science and Technology Commission Major Special Projects of Tianjin of China (No.15ZXDSGX00030), and the Tianjin Municipal Commission of Education of China (No.2021YJSB252).

[**] E-mail: michael3769@163.com

detection models based on the statistical differences often can obtain better detection effects on UAEs. Compared with it, N-UAEs are more limited by the space of AEs. Therefore, the detection techniques based on transformations on inputs to restrict the space of AEs can usually only detect N-UAEs effectively.

Most of existing algorithms only dedicate to achieving the highest detection rate of one certain type of AEs while failing to take the detection rate of the overall ones into account. However, in reality, it is more important to trade off the defensive effects on various AEs and to make the defense system have better defense effects on all kinds of AEs[10].

In this paper, we propose an AE detection method based on the maximum channel of saliency maps (MCSM). This method compresses the space of AEs and preserves the statistical properties of images to some extent. In order to assess the performance of the proposed method, we adopt 6 kinds of adversarial attacks on ImageNet large scale visual recognition challenge (ILSVRC) 2012 to generate the AEs. The experimental results show that the proposed method can achieve well-balanced high detection rates on both the two types of AEs.

The saliency map is built using gradients of the output over the input. It is a visualization technique to measure the importance of each pixel of the image on the model[11].

In order to simulate human visual perception and strengthen the longitudinal connection of each pixel, we map the three channels of the saliency map into a single channel for each pixel using the maximum mapping. We find that the difference between the maximum channel saliency maps of the original images and the AEs is much larger than the difference between the original images and the AEs. In order to compare differences, we first extract the maximum channel difference between the original images and the AEs. The extraction procedure is shown in Fig.1.

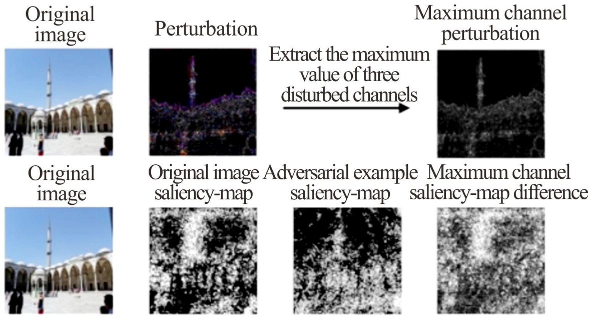
Original image
Perturbation
Extract the maximum value of three disturbed channels
Maximum channel perturbation
Original image
Original image saliency-map
Adversarial example saliency-map
Maximum channel saliency-map difference

Fig.1 Maximum channel of perturbations and saliency map difference extraction

The perturbations can be calculated with the absolute values of differences between the original images and the AEs. The maximum values of the perturbation channels can be calculated by the maximum value of the three channels at the same point in the perturbation images.

The differences in saliency maps can be defined as the differences in the corresponding the maximum channels of the saliency maps in original samples and in AEs. Fig.2 compares the maximum channel differences in perturbations (MDP) with the maximum differences in saliency maps (MDS) using various adversarial algorithms.

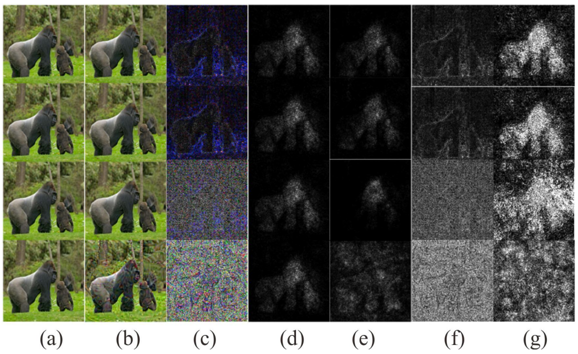

Fig.2 Comparison between MDPs and MDSs (The AEs are generated by Carlini & Wagner 2 (CW_2)[12], DeepFool[13], basic iterative method (BIM)[14] and fast gradient sign method (FGSM)[15] sequentially from top to bottom): (a) Original images; (b) AEs; (c) Differences between original images and AEs; (d) MCSM in the original images; (e) MCSM in the AEs; (f) MDP between the original images and AEs; (g) MDS between the original images and AEs

We generate AEs with four different algorithms on the same 100 images. Then we calculate the mean and variance of their perturbations. Besides, the mean and variance of the saliency maps are also calculated. The statistical results are shown in Fig.3 and Fig.4. In Fig.3, given the fact that there's a large statistical difference in distinct algorithms, we normalized each set of data for ease of comparison.

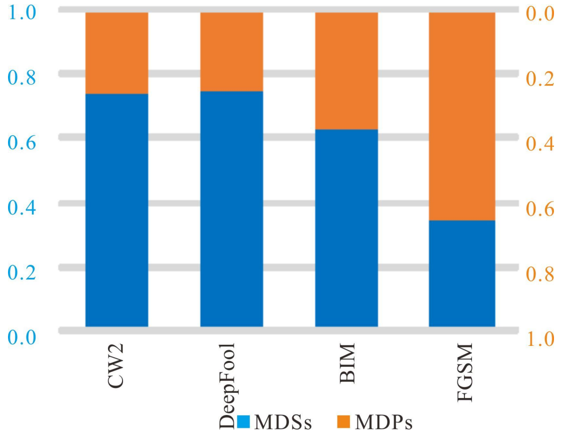

Fig.3 Comparison of average value between MDSs and MDPs

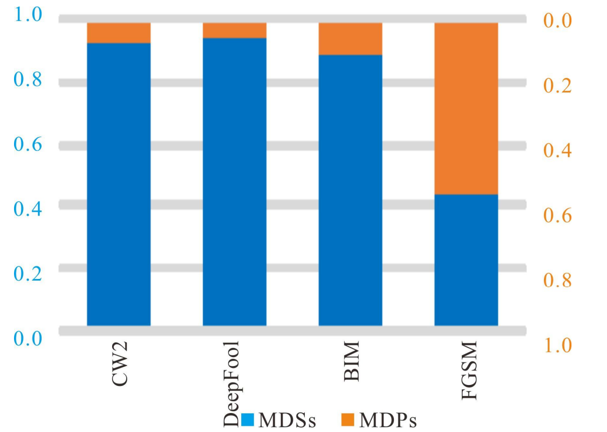

Fig.4 Comparison of variance value between MDSs and MDPs

The results suggest that, except for FGSM with too large single step size to introduce excessive perturbations and even distort images, the average differences of the MCSM are much greater than those of the maximum channel of perturbations, and even the variance may vary by an order of magnitude. The results demonstrate that the saliency maps can magnify the added perturbations and capture the changes in images easily. Hence, the detection effects of the binary-class detector networks on saliency maps should be better than that of detecting the images themselves directly.

Therefore, we propose a new detection method using the MCSM. The architecture of the defense system is shown in Fig.5.

We divide the protected system into preprocessing and image classification. Preprocessed images first enter the MCSM system. After that, the images labeled as 'Normal' are passed into the classification model.

The images entering the detection system have been preprocessed and unified. And we back-propagate each pixel of the image to get the saliency map of the whole image.

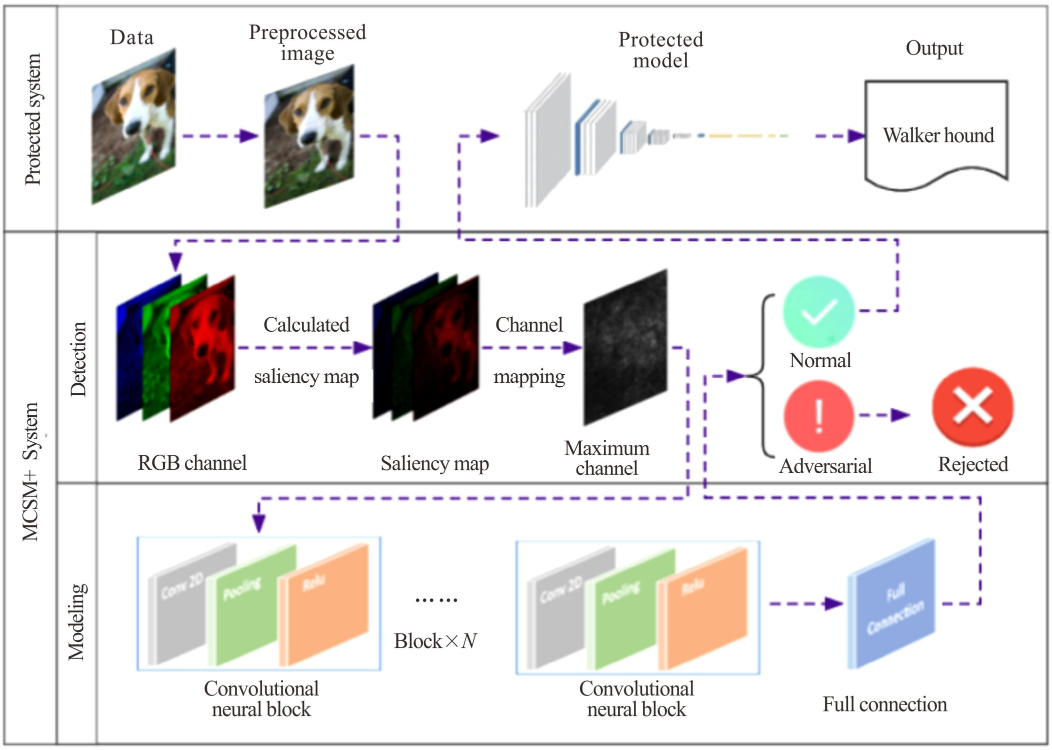

Fig.5 MCSM-based defense system

Suppose that the whole complex classifier $f(\cdot)$ can be approximated by applying a linear function in the neighbor of x_n using a first-order Taylor expansion $f(x)\approx\omega^{\mathrm{T}}+b$, where $\omega=\partial y/\partial x$ is the saliency map we need to compute. The larger value of ω means that the pixel has a greater impact on the result of the classifier.

The shape of the saliency map is also (3, H, W) for a three-channel RGB image. Then we perform the absolute-maximum mapping on the three channels of the saliency map, and keep the maximum absolute values in 3 channels. We can get the maximum channel of the saliency map (MCS) shaped like (H, W), which can be formulated as

$$MCS = x_{HW}, x_{HW} = \max\left\{\left|R_{HW}\right|,\left|G_{HW}\right|,\left|B_{HW}\right|\right\}. \quad (1)$$

We find that the MCSM algorithm is not suitable for complex models in experiments. It works poorly both on the modified Resnet50 and InceptionV3. This may be due to few features or insufficient data available. Therefore, we use a combination of multiple convolutional neural blocks to construct a light weight detection model. Each convolutional neural block contains a 5×5 convolutional kernel which uses the same padding, a rectified linear unit (ReLU) activation function, and a Max-Pooling layer with a kernel size of 2. After multiple convolutional blocks, we use a fully connected layer to classify data at the end. The samples that output from the defense system can be divided into two classes: 'Normal' and 'Adversarial'. If the samples belong to 'Normal', the defense system would pass them to the protected network to make predictions. If the samples belong to the 'Adversarial', the defense system would reject their requests.

Specifically, the pre-trained visual geometry group 16 (VGG16) architecture has been used as our protected network in this experiment. This network produces 74% accuracy in precision top-1 and 92.7% in top-5 on ILSVRC 2012. The reason why we choose VGG16 is that existing research has demonstrated that the AEs generated by the VGG16 network are more transferable. This means that choosing VGG16 as the protected model would make our model more universal. Besides, the reason why we use the pre-trained network is that it would hugely reduce the cost of training. Additionally, it has better estimates of the parameters and stronger authority.

We use the ILSVRC 2012 validation set (50 000 images) as our dataset. By taking 5 000 images as a group, we adopt different algorithms to generate AEs, 3 UAEs algorithms and 3 N-UAEs algorithms.

As for the UAEs, the FGSM, proposed by GOODFELLOW et al[14], generates the AEs by adding a small perturbation ε proportionally to the input images along the gradient direction $\operatorname{sign}\left(\nabla_x J\left(\theta, x, y\right)\right)$. Subsequently, a variant of the FGSM is referred to as the BIM. It is the iterative version of the FGSM attack with smaller ε.

In terms of the N-UAEs, CW is superior to other white-box algorithms. This algorithm adopts some transformations to map the AEs into an infinite interval, which converts the constrained optimization into an unconstrained one. In addition, similar to FGSM, DeepFool is also a gradient-based algorithm. It employs the vertical approximation method to estimate the distance from the sample to the classification boundary. However, unlike FGSM, the perturbations generated by DeepFool are extremely small.

In this experiment, we select two sets of AEs generated by the FGSM with step size $\varepsilon_{\mathrm{FGSM}}$=0.01/0.001 and the BIM with step size $\varepsilon_{\mathrm{BIM}}$=0.005 to compose our UAEs dataset. And the AEs, included in the N-UAEs dataset, are generated by CW_2, CW_∞ and DeepFool.

We use the 2, 4 and 6 convolutional blocks to conduct the experiment respectively. And we found the optimal hyperparameters by this way. The results of the experiment are shown in Tab.1. MCSM-2, MCSM-4 and MCSM-6 mean the MCSM with 2, 4 and 6 convolutional neural blocks, respectively. The number after FGSM and BIM means the step size we used in the corresponding two attacks.

Tab.1 Comparison with multiple MCSMs using convolutional neural blocks

Method	FGSM 0001	BIM 0005	FGSM 001	CW_2	CW_∞	Deep-Fool	Average scores
MCSM-2	0.908	0.866	0.891	**0.917**	**0.936**	**0.958**	**0.913**
MCSM-4	0.912	0.884	0.915	0.619	0.602	0.950	0.814
MCSM-6	**0.926**	**0.901**	**0.923**	0.541	0.556	0.941	0.798

Tab.1 summarizes the prediction performance of different numbers of convolutional neural blocks. In terms of the number of convolutional neural blocks, the detection ability of the model against UAEs has slightly improved with the increase of DNN models, while the detection ability against N-UAEs decreased a lot. In terms of the average scores, we recommend using the neural networks with fewer blocks.

We reimplement three representative detection algorithms against AEs for comparison. The results are shown in Tab.2, where AEs generated by FGSM and BIM are UAEs, while AEs generated by CW and DeepFool are N-UAEs.

Tab.2 Comparison between MCSM and existing methods

Method	FGSM 0001	BIM 0005	FGSM 001	CW_2	CW_∞	Deep-Fool	Average scores
Gauss[16]	0.420	0.379	0.362	0.519	0.629	0.582	0.482
NSS[7]	0.912	**0.953**	**0.974**	0.803	0.812	0.847	0.884
FS[9]	0.603	0.694	0.558	**0.925**	**0.959**	0.865	0.767
MCSM-2	**0.913**	0.866	0.891	0.917	0.936	**0.958**	**0.913**

Among the three groups of control experiments we conducted, the Gaussian perturbation method is a relatively traditional defense, which works by adding some Gaussian perturbations randomly to the AEs. However, this algorithm may not perform very well under various situations due to its randomness. Since the NSS detection relies on statistical properties of images, the detector is more effective for UAEs than N-UAEs. Feature squeezing (FS) squeezes inputs by reducing the color bit depth and spatial smoothing. It has better detection performance on the AEs generated by CW_2 and CW_∞, but it performs relatively poorly on the UAEs.

Although our method outperforms than the state-of-the-art methods only on the AEs generated by the DeepFool and the small-step FGSM, it can achieve relatively balanced effect on both UAEs and N-UAEs. In principle, the saliency map extraction and the maximum channel mapping in the MCSM breaks the perturbation amount and its structure. They also preserve the statistical properties of original samples to some extent. Therefore, they retain the advantages of the NSS and FS. Furthermore, based on the high average and small variance in Tab.2, the MCSM has stronger universality and more stable detection rate than any other algorithm.

In addition, like the Gaussian random perturbation method and FS, many other detection methods are constrained by the prediction accuracy of the protected model itself. That is only when the protected model itself can correctly predict the original images that these methods have a higher detection rate of AEs. However, our method is not affected by the prediction accuracy of the protected model. In fact, among the 200 AEs we randomly select that are misclassified by the detector network, the samples whose original images could not be correctly classified by the protected network account for about 53.5%. It demonstrates that our method is not affected by the accuracy of the protected network.

To verify the effectiveness of the MCSM, we conduct a two-step ablation experiment. First, we remove the maximum mapping in the MCSM. We call this method CSM. Second, we also remove the calculation of saliency maps in the CSM. We call this method CM. The results are shown in Tab.3.

Tab.3 Ablation experiment of MCSM

Method	FGSM 0001	BIM 0005	FGSM 001	CW_2	CW_∞	Deep-Fool	Average scores
CM	0.609	**0.912**	**0.930**	0.520	0.542	0.525	0.673
CSM	0.861	0.888	0.910	0.802	0.812	0.835	0.851
MCSM-2	**0.913**	0.866	0.891	**0.917**	**0.936**	**0.958**	**0.913**

Results in Tab.3 show that only when the perturbation amount is relatively large, the CSM method has a relatively small improvement in detection rate compared with the CM method.

Furthermore, the advantages of MCSM become more obvious and the detection rate on UAEs has been increased even more. In comparison, The MCSM algorithm achieves a higher detection rate on the entire data set. This verifies the effectiveness of our methods further.

In summary, we propose a novel approach to detect the AEs using the MCSM. This method has a high detection rate both for UAEs and N-UAEs. Furthermore, it wouldn't affect the accuracy of the protected model. Additionally, whether more complex models, much larger datasets or more comprehensive combinations of saliency maps could further increase the accuracy of the model? How to preserve the statistical properties of samples when faced with the problem of the example perturbations dilution? And how to get more clear decision boundaries to make the detection model more targeted according to the protected network? These important questions open some promising research paths that are worth studying.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//2015 International Conference on Learning Representations (ICLR), May 7-9, 2015, San Diego, CA, USA. CoRR, 2015：abs/1409.1556.

[2] SZEGEDY C, ZARERBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//2014 International Conference on Learning Representations (ICLR poster), April 14-16, 2014, Banff, Canada. CoRR, 2014：abs/1312.6199.

[3] ZHANG S S, ZUO X, LIU J W. The problem of the adversarial examples in deep learning[J]. Chinese journal of computers, 2019, 42(08)：1886-1904.

[4] WANG X M, LI J, KUANG X H, et al. The security of machine learning in an adversarial setting：a survey[J]. Journal of parallel and distributed computing, 2019, 130：12-23.

[5] SERBAN A, POLL E, VISSER J. Adversarial examples on object recognition：a comprehensive survey[J]. ACM computing surveys, 2020, 53(3)：1-38.

[6] GROSSE K, MANOHARAN P, PAPERNOT N, et al. On the (statistical) detection of adversarial examples[EB/OL]. (2017-02-21) [2021-11-12]. https：//arxiv.org/pdf/1702.06280.pdf.

[7] KHERCHOUCHE A, FEZZA S A, HAMIDOUCHE W, et al. Detection of adversarial examples in deep neural networks with natural scene statistics[C]//2020 International Joint Conference on Neural Networks (IJCNN), July 19-24, 2020, Glasgow, UK. New York：IEEE, 2020：9206956.

[8] LIANG B, LI H C, SU M Q, et al. Detecting adversarial image examples in deep neural networks with adaptive noise reduction[J]. IEEE transactions on dependable and secure computing, 2021, 18(1)：72-85.

[9] XU L, EVANS D, QI Y J, et al. Feature squeezing：detecting adversarial examples in deep neural networks[C]//2018 Conference on Network and Distributed System Security, February 18-21, 2018, San Diego, CA, USA. CoRR, 2018：abs/1704.01155.

[10] CAI P, QUAN H M. Face anti-spoofing algorithm combined with CNN and brightness equalization[J]. Journal of Central South University, 2021, 28(1)：194-204.

[11] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep

inside convolutional networks：visualising image classification models and saliency maps[C]//2014 International Conference on Learning Representations (ICLR poster), April 14-16, 2014, Banff, Canada. CoRR, 2014：abs/1312.6034.

[12] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy, May 22-24, 2017, San Jose, CA, USA. New York：IEEE, 2017：39-57.

[13] MOOSAVI S M, FAWZI A, FROSSARD P. Deepfool：a simple and accurate method to fool deep neural networks[C]//2016 IEEE International Conference on Computer Vision (CVPR), June 26-July 1, 2016, Las Vegas, NV, USA. New York：IEEE, 2016：16526893.

[14] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//2017 International Conference on Learning Representations (ICLR), April 24-26, 2017, Toulon, France. CoRR, 2017：abs/1607.02533.

[15] GOODFELLOW I L, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//2015 International Conference on Learning Representations (ICLR), May 7-9, 2015, San Diego, CA, USA. CoRR, 2015：abs/1412.6572.

[16] CARLINI N, WAGNER D. MagNet and "efficient defenses against adversarial attacks" are not robust to adversarial examples[EB/OL]. (2017-11-22) [2021-11-12]. https：//arxiv.org/abs/1711.08478.