A Non-Global Disturbance Targeted Adversarial Example Algorithm Combined with C&W and Grad-Cam

Yinghui Zhu
Hanshan Normal University
Yuzhen Jiang (jiangyz@hstc.edu.cn)
Hanshan Normal University

Research Article

Keywords: adversarial example, targeted attack, category explanation, salient region

Posted Date: September 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-865960/v1

A Non-Global Disturbance Targeted Adversarial Example Algorithm Combined with C&W and Grad-Cam

ZHU Yinghui, JIANG Yuzhen*
(School of Computer and Information Engineering, Hanshan Normal University, Chaozhou 521041, China)
ZHU Yinghui email: zyh366@163.com
* JIANG Yuzhen email: jiangyz@hstc.edu.cn

Abstract: Adversarial examples are artificially crafted to mislead deep learning systems into making wrong decisions. In the research of attack algorithms against multi-class image classifiers, an improved strategy of applying category explanation to the generation control of targeted adversarial example is proposed to reduce the perturbation noise and improve the adversarial robustness. On the basis of C&W adversarial attack algorithm, the method uses Grad-Cam, a category visualization explanation algorithm of CNN, to dynamically obtain the salient regions according to the signal features of source and target categories during the iterative generation process. The adversarial example of non-global perturbation is finally achieved by gradually shielding the non salient regions and fine-tuning the perturbation signals. Compared with other similar algorithms under the same conditions, the method enhances the effects of the original image category signal on the perturbation position. Experimental results show that, the improved adversarial examples have higher PSNR. In addition, in a variety of different defense processing tests, the examples can keep high adversarial performance and show strong attacking robustness.
Keywords: adversarial example; targeted attack; category explanation; salient region

1. Introduction

Derived from the development of deep neural network classification and recognition, adversarial example overlays a well-designed disturbing noise on the original image, and the system makes a wrong decision. In the generation graph of an adversarial example, the low superimposed disturbing noise cannot be recognized by human eyes or detection software. However, a Class A picture is likely to be recognized as a Class B picture by the system, which poses a great threat to the reliability and security of the application of network model. Proposed by Szegedy et al. ,[1,2] the concept of adversarial example has quickly attracted the attention of scholars in the field of deep learning, and gradually formed a new study direction. Many scholars have tried to make more scientific explanation for the existence of adversarial examples. According to Goodfellow et al., it is caused by the linear characteristics of deep network [3], while more scholars such as Papernot and Su believe that adversarial examples are the product of inaccurate classification decision boundaries of the existing deep neural network models [4,5]. However, the deep neural network can give the image recognition result with the highest confidence, which may be wrong but reasonable in the inference process. The inference process of deep learning consists of numerous judgments. If an input graph is inferred from the decision route of the decision result, and the important factors influencing the decision can be determined, then the basis of the decision result of deep network can be revealed. To this end, Selvaralju et al. , proposed Grad-Cam algorithm [6], which made the visual interpretation of the category judgment of CNN reasonable. The visual interpretation of network decision has injected new vitality into the study of deep learning. In section 2.2, this study also explains the reasons for wrong decisions: since disturbing noise is added, the attention position, target range and category intensity of network model have changed greatly in image recognition, which leads to the change of prediction results.

With noisy and robust large-class image adversarial example as the study object, this study introduced Grad-Cam method, and proposed a non-global noise targeted attack algorithm. Based on the combination of C&W algorithm and category visual interpretation method, the disturbing noise range is narrowed, and the position of the target category is closer to the original category position, so as to strengthen the target disturbing signal and improve the adversarial performance. In the experiment, the algorithm is competent to

attack any input graph, that is, the attack rate is 100%. Compared with other large-class adversarial example algorithms under the same conditions, this algorithm can effectively reduce the global sensitivity of disturbing noise, and the generated examples have higher peak signal-to-noise ratio, stronger concealment, and better adversarial performance after receiving attack-defense processing, such as geometric transformation, JPEG compression, and blurring.

2. Related work

2. 1 Adversarial example algorithm

The existing studies of adversarial-attack/adversarial example of deep learning classifier have proposed many algorithms with strong attack power. From the aspect of attacking targets, adversarial example algorithms can be divided into targeted algorithms and non-targeted algorithms. The targeted algorithm needs the classifier to identify the input image as a specified category to realize successful attack, while the non-targeted algorithm only needs the classifier to identify the image as other categories. In other words, if it is different from the original category, the attack is successful. The existing classical non-targeted algorithms include FGSM [3], BIM[7], One-Pixel[5], etc., and the targeted attack algorithms include JSMA[4], L-BFGS[2], PGD[8], C&W[9], etc. Fig. 1 shows different adversarial-attack effects of a digital image. The first line is the original image and adversarial examples generated by different algorithms, the second line is the enlarged effect of the disturbing noise image, and *C* is the recognized category.

From the aspect of the added disturbing information, the adversarial example algorithm can be divided into global noise algorithm and local noise algorithm. The global noise algorithm superposes disturbing information to the original image, and the added information can be regarded as a noise image with the same resolution as the original image. Among different algorithms, FGSM, BIM, L-BFGS, PGD and C&W are global noise algorithms. Local noise algorithm modifies a small part of pixel information of the image to achieve the purpose of category misjudgment. JSMA and One-Pixel are local noise algorithms, which try to modify as few pixels as possible to achieve category attacks. Although few pixels are modified, this algorithm will produce obvious noise, and the computational power is too large to be suitable for large images. The disturbing noise maps of PGD and C&W algorithms in Fig. 1 are similar with that of local algorithms, both of which belong to global noise algorithms because the disturbing amount of black background area is too small and the display is not obvious.

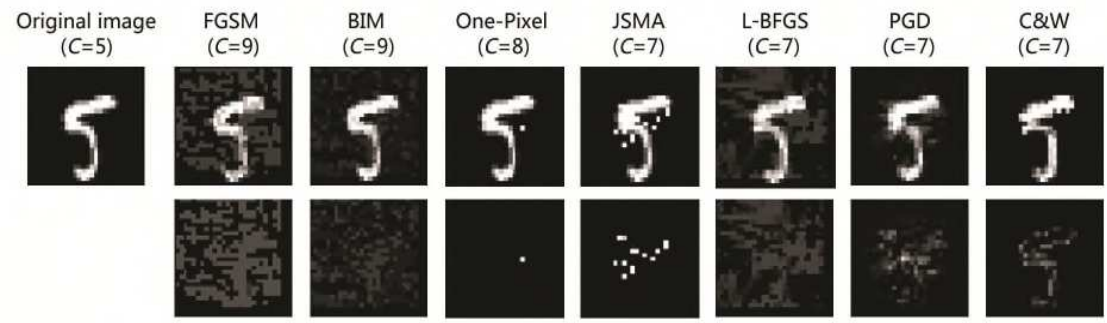

Fig. 1 Adversarial examples and disturbing noise maps of digital maps

2. 2 C&W algorithm

Improved on the basis of L-BFGS, C&W algorithm[9] has less sample noise, high image quality, strong anti-concealment and special immunity to defensive distillation algorithm[10]. The algorithm is feasible and controllable, and it can attack any target with high confidence in images. In [1], it is evaluated as a 5-star algorithm. The characteristics of the algorithm are as below:

(1) In the selection of multiple objective functions and distance standards, an optimal targeted attack scheme based on logits layer is combined. Taking L2 attack with the lowest noise as an example, the optimized loss function is as below:

$$loss = \text{minimize} \left\| I' - I \right\|_2^2 + c \cdot f(I') \tag{1}$$

Formula (1) seeks the lowest loss cost of disturbing noise on the basis of realizing the attack target. I is the original image, and I' is the adversarial example image. L2 normal form limits the disturbance data of the image. Parameter c is the adjustment coefficient of attack quality and attack success rate. $f(I')$ is the objective function proposed by C&W algorithm:

$$f(I') = \max(\max\{Z(I')_i : i \neq t\} - Z(I')_t, -\kappa) \tag{2}$$

$f(I')$ is used to guide the generation of adversarial examples along the specified target class. t is the target category specified by the attack, and parameter κ is used to adjust the balance between attack quality and confidence. The larger κ is, the higher the generalization and confidence of the algorithm will be, but the effect of the confrontation graph is worse (because its L2 normal form value will also become larger). The defensive distillation algorithm can be attacked 100% when $\kappa \geq 40$ [9,10]. The logits output layer of the network model refers to the input layer of the normalized exponential function softmax, and the logits output layer of I' is represented by $Z(I')$ in equation (2).

(2) C&W algorithm proposes an optimization method based on tanh function in formula (3) in the optimization process of iterative steps:

$$I'_i = \frac{1}{2}(\tanh(w_i) + 1) \tag{3}$$

w is an alternative variable of I'. The purpose of formula (3) is to transform the optimization problem of I' into the optimization of w. The advantage of replacing I' with w is that no matter what the optimized value of W is, the mapped value range of I' can be guaranteed to be [0,1], thus avoiding the error caused by using truncation function.

2. 3 PerC-C&W algorithm

The global adversarial algorithm can also be transplanted to the study of the attack of large color images. For color images, most algorithms directly make scrambling attacks in RGB color space. According to Zhao[11], the sensitivity of human eyes to the change of R, G and B color is different. For some sensitive color gamut, slight modification will make people perceive obvious noise. However, for non-sensitive color gamut, even the enhanced disturbing noise intensity is not easy to be perceived. [11] put forward the improvement of color image adversarial example: (1) to realize the adversarial-disturbance processing of images from the aspect of CIELAB color space; (2) to add the limit term of CIEDE2000[12] color distance to the loss function, so as to obtain low noise adversarial examples which cannot be perceived by human eyes. PerC-C&W algorithm is an improved strategy based on C&W, and its loss function is:

$$loss = \min_w \text{minimize} \left\| VE_{00}(I, I') \right\|_2 + \lambda \cdot f(I') \tag{4}$$

Where, $VE_{00}(I, I')$ is CIEDE2000 color distance.

PerC-C&W algorithm determines λ by binary search, which is inefficient. [11] proposed an improved algorithm based on PGD: PerC-AL, which updates disturbing noise by alternately calculating class loss and gradient decline of perceived chromatic aberration, and realizes the optimization of counter samples.

Combined with human color perception factors, PerC-C&W and PerC-AL can still obtain good and imperceptible image visual effects even if the disturbing noise intensity is increased. Therefore, within the same color distance, the two algorithms can obtain higher robustness than the conventional algorithms, and the examples can obtain better immunity in anti-defense processing experiments such as JPEG conversion and color compression.

2. 4 Grad-Cam algorithm

Grad-Cam algorithm is a method for visual interpretation of classification status in convolutional neural networks [6], which can display a certain type of key attention area in images by means of thermodynamic chart.

This method has been successfully applied to the interpretation of some unconventional classification results [13, 14]. Grad-Cam algorithm first uses the last convolution layer of neural network to obtain feature activation graph, and the specific method is as follows:

(1) Derive the convolution layer of the last layer with the classification output result to obtain $\frac{\partial y^c}{\partial A^k_{i,j}}$, where y^c is the classification result and $A^k_{i,j}$ is the value at the position (i, j) in the kth feature map.

(2) Calculate the weight α^c_k by formula (5), where z is the size of the feature map.

$$\alpha^c_k = \frac{1}{z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i,j}} \tag{5}$$

(3) Obtain the final visualization result of Grad-Cam algorithm by formula (6), and the calculated weight information is the important area in the feature map that determines the classification result. The Relu function is used to eliminate the influence of negative values, and only the influence of positive values in the feature graph on the classification results is considered.

$$L^c_{\text{Grad-Cam}} = \text{Relu}\left(\sum_k \alpha^c_k A^k\right) \tag{6}$$

In Fig. 2(a), the recognition result in InceptionV3 is “Eskimo dog” of category 248, with a confidence of 42.12%. Grad-Cam supports testing the recognition of input graph with any category, and obtains the corresponding interpretation thermodynamic map. The explanation results of “Eskimo dog” to “Eskimo dog” are (b) and (c), and the L2 normal value of the thermodynamic map is 10.34. If “tabby cat” of category 281 is input, the interpretation results are (d) and (e), and L2 is 4.18. If “desk” of category 526 is input, Grad-Cam can obtain the interpretation thermodynamic maps (f) and (g) even though this category does not exist. Its L2 is as small as 1.70. It can be seen that Grad-Cam can explain the location of category generation, and its numerical value can reflect the existence probability of category.

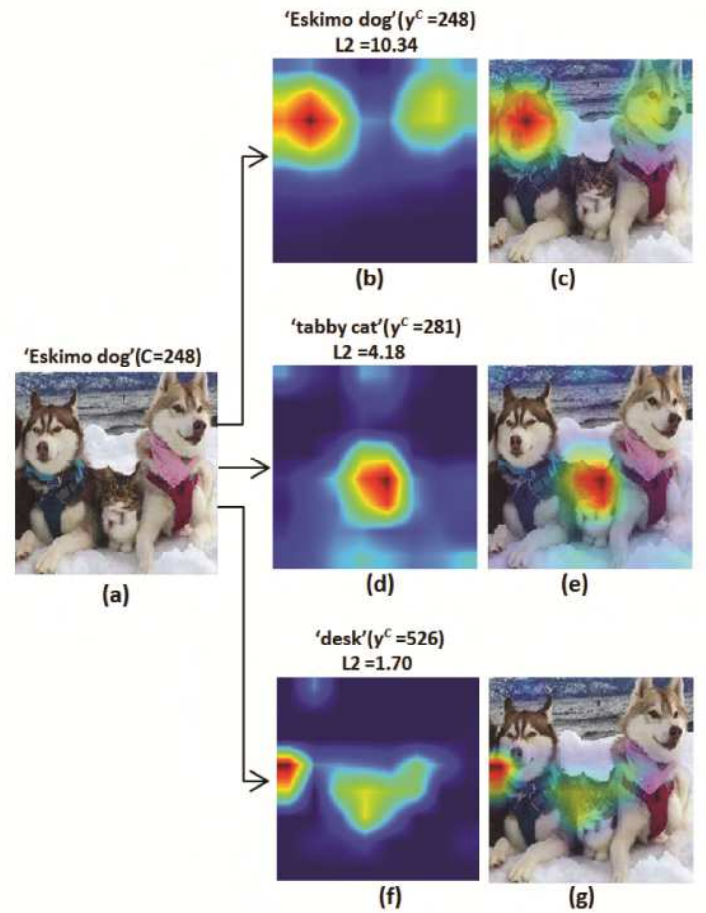

Fig. 2 Grad-Cam thermodynamic charts and L2 normal form values of different categories

3. Adversarial example algorithm for disturbance of Grad-Cam significant area

3. 1 Analysis of sensitivity of global disturbing noise

For global disturbing noise, some studies argue that there is information redundancy, and try to improve a global algorithm into a local algorithm. In [15], the disturbing signal is superimposed in the texture area of the image through Shannon information entropy in the local domain, so as to improve the imperceptibility of the attack. In [16], the image is divided into several independent blocks, and 10 adversarial examples are realized by superimposing disturbing signals on the 10 largest areas respectively. According to [17], the disturbing noise outside the target area is redundant and can be deleted, so that the disturbing signal can only be generated in the target area. These algorithms are feasible in theory, and can get better results for some images. However, the size of the attack area cannot be guaranteed. When the attack area is small, the attack of local scrambling methods may fail, or there will be obvious noise. The studies do not analyze attack success rate or attack robustness. According to experiment results, there is a great correlation between the effectiveness of adversarial attack and the global nature of disturbing noise. If the information of disturbing noise in some areas is randomly removed from the adversarial examples, the attack performance will be lost. This is because the new target category (i.e., attack category) may exist in the removed area, even the seemingly unimportant marginal area.

Figs. 3 and 4 illustrate the problem with FGSM algorithm based on MobileV2 and C&W algorithm based on InceptionV3, respectively. When disturbing signals with seemingly insignificant edges are deleted, the adversarial examples of the two algorithms do no have attack effect. Deleting local disturbing noise may lead to attack failure, because the expected recognition category may utilize the disturbing signal at that location to

realize category conversion. This reason will be further analyzed in the section of Grad-Cam visual interpretation.

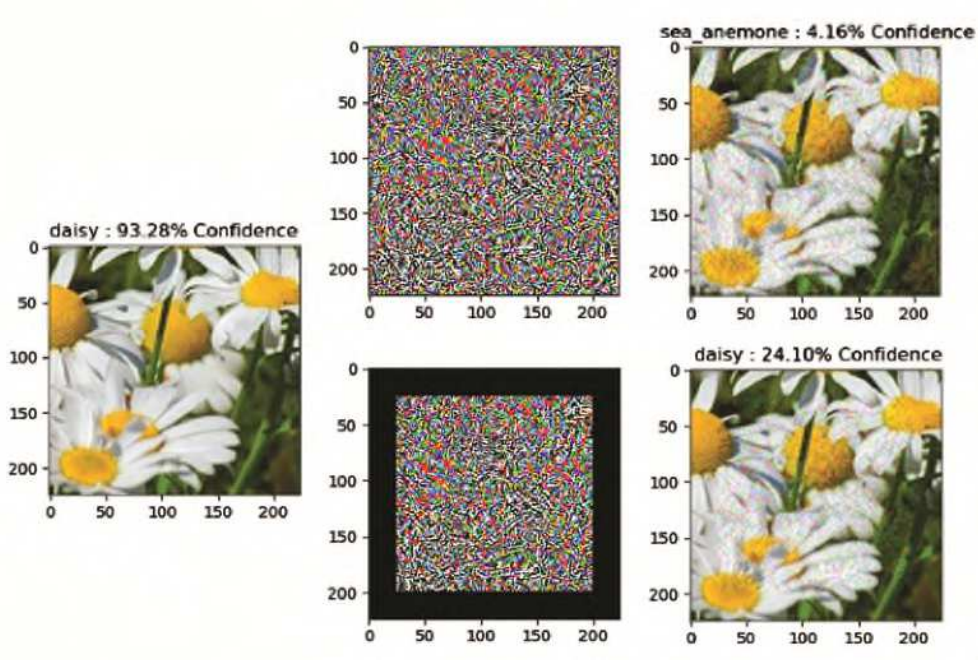

Fig. 3 Comparison of global and non-global FGSM attack effects

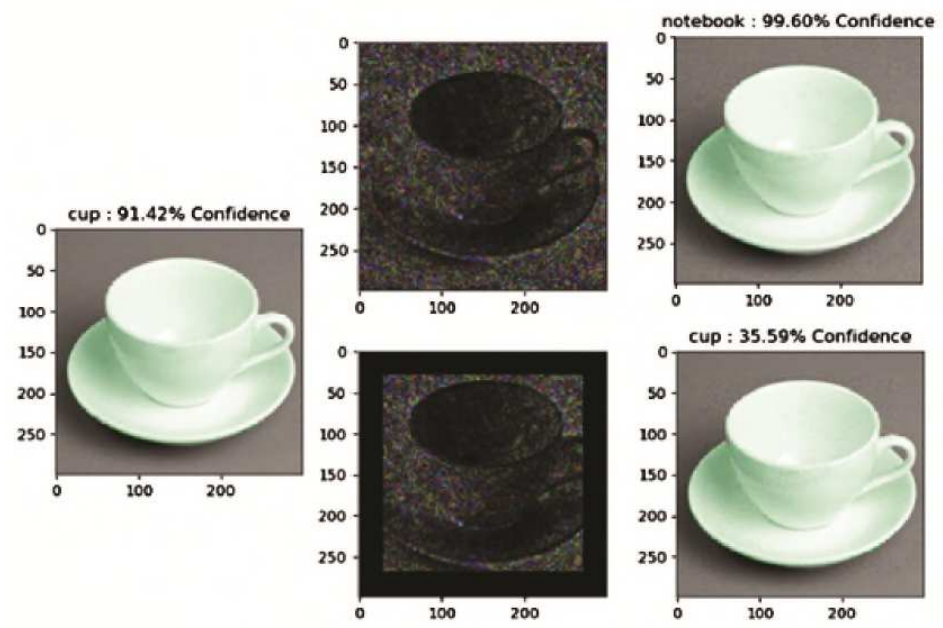

Fig. 4 Comparison of global and non-global C&W attack effects

3. 2 Grad-Cam visual interpretation of adversarial examples

In order to study the modification process of the original image and identify it as the reason for target category, the Grad-Cam analysis of the original category and the target category is carried out for each generated image in the iterative generation process of the adversarial example, and the L2 normal form value is calculated. The left side of Fig. 5 is a “goldfish” picture with a confidence of 99.85%, and the right side is a C&W algorithm adversarial example based on InceptionV3. The recognition type is “desk” with a confidence of 99.41%. The heat thermodynamic charts of the original category and the target category in the adversarial example generation process are in the middle of the two figures, which are displayed according to the generation order. The upper thermodynamic chart represents the explanation diagram of the “goldfish” category in each generation diagram, and the lower one is the explanation chart of the “goldfish” category. With the increase of

the attack intensity, the location and scope of "goldfish" category (y^c=1) are almost unchanged, but the L2 normal form value decreases gradually. On the contrary, the change of "desk" category (y^c=526) is obviously active, its location and scope continuously change and expand, and the L2 normal form value also increases. When the maximum confidence among all categories is obtained, it is recognized as the first category by the model. In Fig. 5, the red vertical dashed line is the watershed between "goldfish" category and "desk" category. After this position, the L2 normal form value of "desk" category is larger than that of "goldfish" category, and the recognition confidence of the model for "desk" is also higher than that of "goldfish" category, that is, the initial successful attack state is reached.

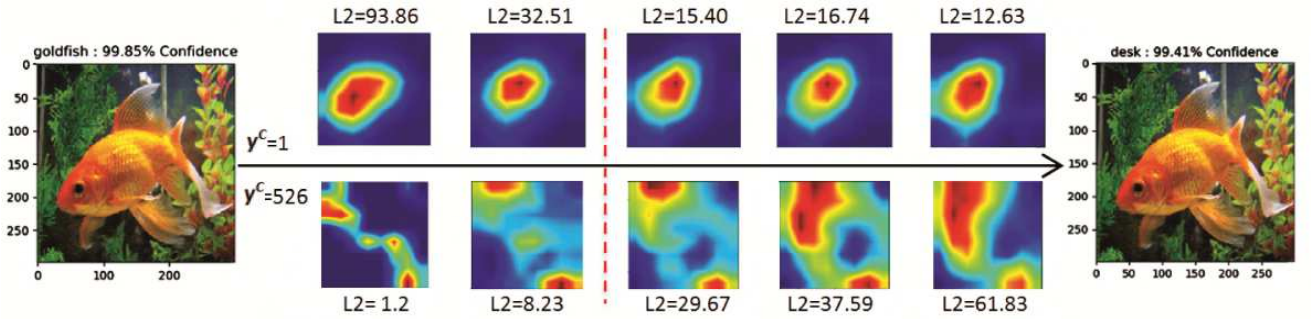

Fig. 5 Grad-Cam interpretation of category recognition transfer in the process of counter sample generation

Grad-Cam explains the main reasons for the generation of targeted adversarial examples through visualization: the recognition strength of the original category on the network model is weakened continuously, and the recognition strength of the target category is enhanced continuously. All modifications to the original image in this process are imperceptible to human visual system, but they are perceived by the network model, thus influencing the classification results. In addition, it is worth mentioning that in the adversarial example, the target category may be generated anywhere in the image, especially at the edge position. For example, the "desk" category mainly exists at the upper left and lower right edge positions. Therefore, randomly eliminating the disturbing noise in some positions in the global disturbing attack may lead to the failure of attack, because the noise may be the signal source of the target category.

3. 3 Algorithm ideas and implementation steps

Based on the reasons for category change of adversarial examples in the above analysis, this study proposed a non-global disturbance adversarial example idea to strengthen the salient areas of images. Based on the C&W algorithm, two Grad-Cam interpretation maps of the original category and the target category are merged in each optimization process. The signals are sorted in descending order, and the first 3/4 image positions are selected for disturbance attack. The generated image is taken as the processing object of the next iteration until the generated image reaches the target category and the confidence is not less than 50%. In the setting of the attack area scale, too large scale makes the attack effect close to the global attack and lose its meaning. However, too small scale leads to local obvious noise in significant areas of images. By carrying out a large number of experiments, it is found that choosing 3/4 scale can not only ensure the attack success rate, but also reduce the total noise and improve the attack concealment performance, thus ensuring a suitable attack area in each iteration. The purpose of the above improvement is to use the initial strong signal of the original category to influence the generation position of the target category on the image, so that the coincidence is higher, and the disturbing data volume of the non-significant area of the original category can be decreased. Since the total confidence of all categories is 1, 50% confidence can ensure that the target category becomes the largest of all categories while not producing too many iterations.

Experiments will be carrie out to verify that, under the same conditions, the proposed method has higher attack robustness in the conventional image countermeasure and defense processing because the attack signal in the significant area is strengthened.

The specific operation steps of the algorithm are as follows:

Step 1: obtain the original category C_{origin}=Model ($\boldsymbol{I}$) of the original image $\boldsymbol{I}$, and input the target category C_{target} to make $\boldsymbol{I}'=\boldsymbol{I}$;

Step 2: the Grad-Cam method is used to implement category interpretation of C_{origin} and C_{target}, as shown in formulas (7) and (8). Combine the data of the two output images in formula (9):

$$\boldsymbol{Map1} = \mathrm{Grad\_Cam}(\boldsymbol{I}', C_{origin}) \tag{7}$$

$$\boldsymbol{Map2} = \mathrm{Grad\_Cam}(\boldsymbol{I}', C_{target}) \tag{8}$$

$$\boldsymbol{Map} = \boldsymbol{Map1} + \boldsymbol{Map2} \tag{9}$$

Step 3: Sort the values on the *Map* in descending order to obtain the threshold value *T* at the 3/4 position in the total data;

Step 4: Implement C&W optimization for the adversarial examples by using formulas (1), (2) and (3). The optimization process is only valid for pixels with *Map* value greater than *T,* that is, the pixels with the *Map* value less than *T* are not optimized;

Step 5: Determine whether $\boldsymbol{I}'$ achieves the attack goal: Model($\boldsymbol{I}'$)== C_{target} and Confidence($\boldsymbol{I}'$)≥50%, output the adversarial example and end the algorithm, otherwise, jump to step 2 and continue to execute the next iteration.

4. Experiment and result analysis

4. 1 Visual analysis of attack position

To verify the influence of non-global attack on the location of target category and the improvement effect of the proposed algorithm compared with C&W algorithm, the experiment first compares the adversarial examples and disturbing noise maps of the two algorithms, and realizes the Grad-Cam localization interpretation of the target category on the adversarial examples. The attacked network model is InceptionV3 trained in ImageNet, and both algorithms adopt Adam optimization method with learning rate r of 0.01. Two experimental cases are presented in Fig. 6. By comparing the disturbing noise images in the middle column, it can be seen that the proposed algorithm reduces the disturbing noise in some edge areas, and makes the anti-disturbing signals more concentrated in the salient areas of the original category. As can be seen from the thermodynamic positioning of Grad-Cam in the third column, the detection position of the target category has also changed, and it also tends to the obvious areas of the original category.

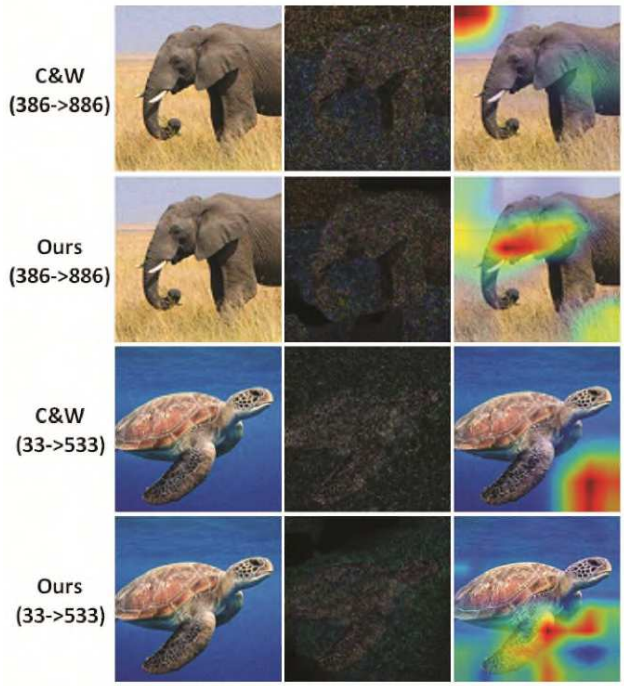

Fig. 6 C&W and the algorithm in this paper are compared

The obvious region of an image represents the most important content information and it is often located in the middle part of the image. Compared with the unimportant background area, the obvious area information of the image is more likely to be preserved in conventional image processing. Therefore, the targeted attack in the obvious region of the image can also effectively improve the robustness of confrontation.

4. 2 Analysis of sample quality

PerC-C&W[xx] and PerC-AL[xx] are the improved algorithms of C&W and PGD after combining CIEDE2000[12] color distance and color difference perception, and their robustness and defense are improved. In order to analyze the quality of samples, six targeted attack algorithms, namely PerC-C&W, PerC-AL algorithm, L-BFGS, PGD, C&W and the proposed algorithm are tested in the same environment. The network model is InceptionV3, and the terminal condition is that all attacks reach the target category with 50% confidence. Fig. 7 reflects the adversarial example graph (original category 386, target category 886), disturbing noise graph and noise MSE value of each algorithm. It can be seen from the graph that the proposed algorithm achieves the best confrontation quality in the visual effect of the adversarial example graph and the comparison of MSE values.

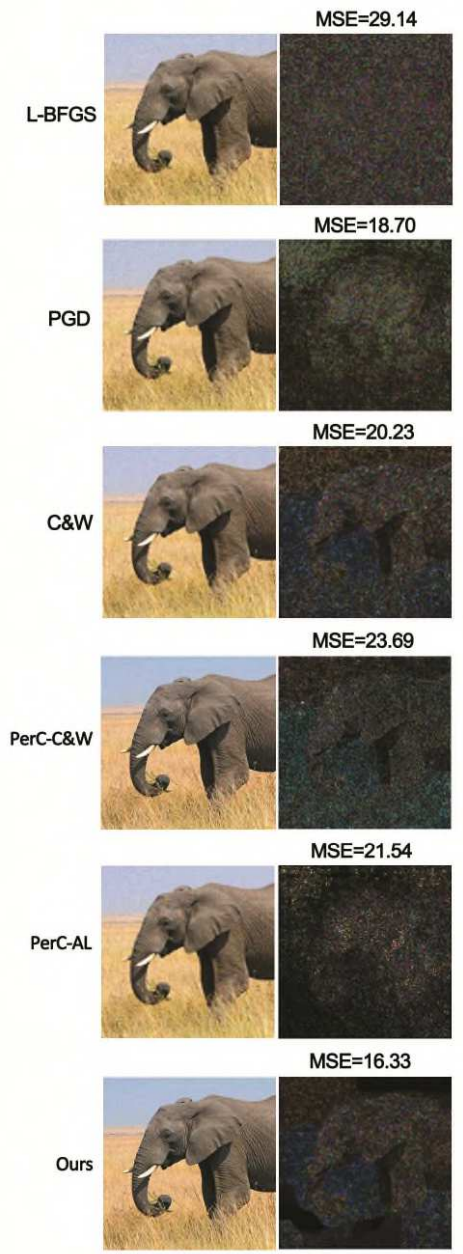

Fig. 7 MSE comparison of different algorithms for adversarial examples and disturbance graphs

Table 1 Average PSNR of different algorithms

To verify the universality of the algorithm and the effective migration of its performance, 100 images were selected in the ImageNet test set, which were correctly classified. The above six algorithms were implemented on four common network models of VGG16, Resnet50, MobileV2 and InceptionV3. For all attacked images, the target category is set as below:

$$C_{target} = \mathrm{mod}(C_{origin} + 500, 1000) \tag{10}$$

The adjacent categories in ImageNet data set generally have similar contents. For example, the first seven categories are fish, and the 8th to 18th categories are flying birds. The calculation of the target category in formula (10) limits the distance between the target category and the original category, so as to realize the category attack with higher discrimination.

Each of the above four models executes six attack algorithms, namely, 24 adversarial examples are generated. On the basis of obtaining all 2400 example graphs, the corresponding peak signal-to-noise ratios (PSNR) are calculated. The experimental results are shown in Table 1. The proposed algorithm has achieved high image quality in the four network models, and the PSNR values of Resnet50、MobileV2、InceptionV3

models are the highest.

Table 1 Average PSNR of various algorithms

Network Model	L-BFGS	PGD	C&W	PerC-C&W	PerC-AL	Ours
VGG16	33.04	**36.08**	35.59	31.72	34.18	35.76
Resnet50	32.51	34.38	34.44	33.02	33.91	**35.33**
MobileV2	33.92	35.83	35.80	32.77	34.61	**36.03**
InceptionV3	33.62	33.20	35.25	33.40	35.03	**36.47**

C&W and PGD are imperceptible attack algorithms with strong concealment[1]. Tanh optimization function introduced by C&W can automatically adjust the disturbance intensity and avoid the use of truncation function. In this way, less noise is generated compared with L-BFGS algorithm under the same condition [9]. PGD is an adversarial attack based on gradient iteration [8]. It uses gradient descending direction to guide sample generation, and limits incremental noise during iteration. Low-noise attack effect is achieved. However, both C&W and PGD are global algorithms, and there are still some redundant noises. The proposed algorithm is a non-global scrambling improvement of C&W, and the scrambling shielding of non-significant areas in the iteration process can effectively reduce the noise required for attack, which explains why the average scrambling noise of the proposed algorithm is less than that of C&W and PGD.

4. 3 Analysis of adversarial robustness

Convolutional neural network is the most commonly used image classifier. Slight geometric transformation of images, such as translation and rotation, can cause drastic changes in feature layers of CNN. For network models trained in large data sets, the geometric transformation of images will not affect the classification decision. However, for adversarial examples, disturbing signals are calculated and generated on CNN feature layers of the original image. After geometric transformation, the relevance between the original disturbing signals and the current feature layer may be weakened or lost. Some studies[18,19] have verified that JPEG compression, blurring and bit depth compression have powerful anti-attack functions, and the highest anti-attack rate is up to 80%-90%.

To test the robustness of the proposed algorithm, five kinds of attacks and defenses are carried out for six target adversarial examples based on the InceptionV3 model: right shift for 10 pixels, right turn 5°, JPEG compression with quality 70, 4-bit/channel color compression and 3*3 median filtering. The categories of the processed examples are predicted to examine whether the attack effect is maintained, that is, whether they are still identified as the specified categories. Fig. 9 shows the attack success rate after different attack-defense treatments (100 pictures). All image processing methods can play the role of attack defense and original category storage. Relatively speaking, the attack robustness of the proposed algorithm is superior to that of the other five algorithms. For 4-bit/channel color compression, the attack success rate of the algorithm is up to 68%. The main reason why the algorithm has high robustness lies in that it disturbs the non-global superposition characteristics of noise. Because the attack information is more concentrated in the significant content area, the average disturbance intensity in the disturbance area is slightly higher than that of the global algorithm, which makes it more resistant to information compression processing such as JPEG conversion, color compression and fuzzy filtering; Conventional translation and rotation usually make the image lose some boundary information. There are relatively few disturbing signals distributed at the edge of the image, and it is more immune to such defense processing. The implementation environment of the above algorithm is Tensorflow2.0, and the hardware device is NVIDIA RTX3000 (6G).

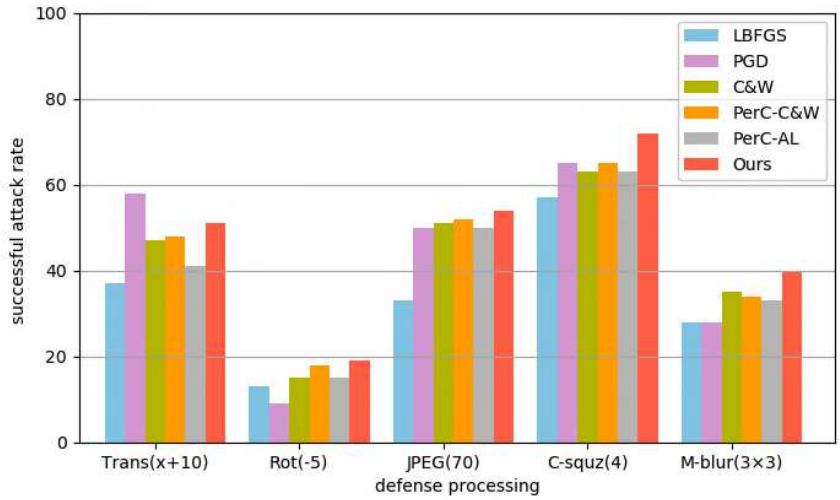

Fig. 8 Adversarial performance of examples after different adversarial defense processing

5. Conclusions

This study proposed a non-global disturbance targeted adversarial example algorithm, which has stronger concealment and better visual quality compared global scrambling method, and effectively improves the robustness of adversarial examples. Based on the C&W adversarial-attack algorithm, the proposed algorithm combines the visualization function of category explanation of Grad-Cam algorithm for region selection, and the disturbing signal is more concentrated in the obvious area of the original category of the image. Because the obvious area of the original category represents more original class image information, its position tends to the middle part of the image and usually has more texture information compared with the background region. Therefore, it is less noticeable to superimpose more adversarial noise in the obvious area of the original category. Moreover, for conventional geometric transformations such as clipping, translation and rotation, the image content in the central salient area is less likely to be deleted compared with that in the background area at the image edge. Similarly, the superimposed disturbing signal will not be eliminated. Experiments show that the adversarial example strategy of enhancement and disturbance in obvious area can not only make the adversarial example more concealed, but also effectively improve the attack robustness. Experiment results verify the feasibility and improved performance of the algorithm from different aspects.

The purpose of the study of adversarial attack is not to verify the strong attack ability of adversarial examples, but to better reveal the existence state and harmfulness, so as to develop more advanced defense methods and safer application systems. Imperceptibility, real-time, and robustness of adversarial examples are the current study focuses. In the future, the attack of adversarial example may be faster, more concealed, and difficult-to-eliminate, and the study of adversarial attack will also face greater challenges. Therefore, it is suggested to improve the security of system application from the perspective of joint detection, which can be implemented into two ways: (1) Synchronous prediction of multi-channel heterogeneous networks. Because adversarial attack usually focuses on specific networks, introducing heterogeneous networks is more likely to generate different results, and synchronous prediction can meet the real-time requirements. However, higher costs will be paid. (2) Joint detection of multiple preprocessing. Defense processing is not competent for detecting different unknown attacks, but if it can realize predictive detection combined with deeper information processing technologies such as scale scaling and frequency domain conversion. Also, the detection reliability is significantly improved. It is also the future study direction to realize an efficient and reliable adversarial defense detection system.

Ethics approval

This article does not contain any research with human participants or animals performed by any of the au-

thor.

Acknowledgement

Chaozhou philosophy and Social Sciences 13th five year plan project (2019-A-05, 2020-C-17); Scientific research project of Hanshan Normal University (XS201908, XN202034)

Conflict of interest

We declare that there is no conflict of interest regarding the publication of this paper.

Contributions

Zhu Yinghui is responsible for the collection of experimental data and related materials, and Jiang Yuzhen is responsible for the writing and communication of the paper.

Informed consent

All authors agree to submit this version and claim that no part of this manuscript has been published or submitted elsewhere.

We appreciate your consideration of our manuscript, and we look forward to receive comments from the reviewers as soon as possible.

Reference

[1] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6: 14410-14430.

[2] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]// Proceedings of the 2nd International Conference on Learning Representation, Banff, AB, Canada, ICLR Press, 2014:1-10.

[3] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]// Proceedings of the 3rd International Conference on Learning Representations.Lille, France, ICLR Press, 2015:1-10.

[4] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]// Proceedings of 2016 IEEE European Symposium on Security and Privacy, Saarbrucken, Germany, DC: IEEE Computer Society, 2016: 372–387.

[5] Su J, Vargas D V, Kouichi S. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2017, 23(5): 828-841.

[6] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [C]// Proceedings of the2017 International Conference onComputer Vision. Venice, Italy, ICCV Press, 2017: 618–626.

[7] Kurakin A, Goodfellow I J, Bengio S. Adversarial machine learning at scale[C]// Proceedings of the 5th International Conference on Learning Representations. Toulon, France, ICLR Press, 2017:203-219.

[8] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[J/OL].arXiv preprint arXiv:1706.06083,2018.

[9] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]// Proceedings of the 38th IEEE Symposium on Security and Privacy. San Jose, California, DC: IEEE Computer Society, 2017: 39–57.

[10] Papernot N , Mcdaniel P , Wu X , et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[J]. IEEE Symposium on Security and Privacy, 2016, 3: 582-597.

[11] Zhao Z , Liu Z , Larson M. Towards Large Yet Imperceptible Adversarial Image Perturbations With Perceptual Color Distance[J/OL].arXiv preprint arXiv:1911.02466v1,2019.

[12] Luo M R, Cui G, Rigg B. The development of the CIE 2000 colour-difference formula: CIEDE2000[J]. Color Research and Application,2001,26(5):340–350.

[13] Heo J, Joo S, Moon T. Fooling Neural Network Interpretations via Adversarial Model Manipulation[J/OL].arXiv preprint arXiv:1902.02041,2019.

[14] Ma X, Niu Y, Gu L, et al. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems[J]. Pattern Recognition, 2020:107332.

[15] Gpfert J P, André Artelt, Wersing H, et al. Adversarial Attacks Hidden in Plain Sight[M].Konstanz :Advances in Intelligent Data Analysis XVIII, 2020: 235-247.

[16] Gpfert J P, Wersing H, Hammer B. Recovering Localized Adversarial Attacks[C]// Proceedings of the 28th International Conference on Artificial Neural Networks, Munich, Germany, ICANN, 2019: 302-311.

[17] Fischer V, Kumar M C, Metzen J H, et al. Adversarial Examples for Semantic Image Segmentation[J/OL]. arXiv preprint arXiv:1703.01101,2017

[18] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world [J/OL], arXiv preprint arXiv:1607.02533,2017

[19] Guo C, Rana M, Cisse M, et al. Countering Adversarial Images using Input Transformations[J/OL]. arXiv preprint arXiv:1711.00117,2017.