DOI: 10.19884/j.1672-5220.202312003

Attention-Guided Sparse Adversarial Attacks with Gradient Dropout

ZHAO Hongzhi[1,2], HAO Lingguang[1,2], HAO Kuangrong[1,2*], WEI Bing[1,2], LIU Xiaoyan[1,2]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitized Textile and Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China

Abstract: Deep neural networks are extremely vulnerable to externalities from intentionally generated adversarial examples which are achieved by overlaying tiny noise on the clean images. However, most existing transfer-based attack methods are chosen to add perturbations on each pixel of the original image with the same weight, resulting in redundant noise in the adversarial examples, which makes them easier to be detected. Given this deliberation, a novel attention-guided sparse adversarial attack strategy with gradient dropout that can be readily incorporated with existing gradient-based methods is introduced to minimize the intensity and the scale of perturbations and ensure the effectiveness of adversarial examples at the same time. Specifically, in the gradient dropout phase, some relatively unimportant gradient information is randomly discarded to limit the intensity of the perturbation. In the attention-guided phase, the influence of each pixel on the model output is evaluated by using a soft mask-refined attention mechanism, and the perturbation of those pixels with smaller influence is limited to restrict the scale of the perturbation. After conducting thorough experiments on the NeurIPS 2017 adversarial dataset and the ILSVRC 2012 validation dataset, the proposed strategy holds the potential to significantly diminish the superfluous noise present in adversarial examples, all while keeping their attack efficacy intact. For instance, in attacks on adversarially trained models, upon the integration of the strategy, the average level of noise injected into images experiences a decline of 8.32%. However, the average attack success rate decreases by only 0.34%. Furthermore, the competence is possessed to substantially elevate the attack success rate by merely introducing a slight degree of perturbation.

Key words: deep neural network; adversarial attack; sparse adversarial attack; adversarial transferability; adversarial example
CLC number: TP183 **Document code**: A
Article ID: 1672-5220(2024)05-0545-12

Open Science Identity (OSID)

0 Introduction

Recent years have witnessed widespread research on adversarial examples, shedding light on the shortcomings of models whilst improving their robustness. Given the susceptibility exhibited by deep neural networks (DNNs) in computer vision tasks like semantic segmentation[1-2], object detection[3-5] and image classification[6-8], this becomes critically important. Such vulnerability is often exploited through incorporating human-imperceptible perturbations into the clean image, leading to misclassifications or incorrect segmentation with malicious intent.

Typically, adversarial attacks can be divided into white-box attacks and black-box attacks[9]. A white-box attack involves having access to the target model's output, structure and parameters. In contrast, attacks on black-box models only allow access to the predictions of the target models. Additionally, black-box attacks can be further classified as query-based attacks[10-11] and transfer-based attacks[12-13]. Query-based attacks require a large amount of queries to the target model, and transfer-based attacks exploit the transferability of adversarial examples. Black-box attacks have gained attention due to limitations of the former two techniques. Transfer-based attacks are promising in the field of adversarial attacks because they do not necessarily require the knowledge of the target model's architecture and parameters like white-box attacks, and are not restricted by limited access times like query-based attacks.

In fact, transfer-based attacks encounter a bottleneck that can be attributed to the inclination towards overfitting to surrogate models[12,14]. Recent research endeavors to boost the effectiveness of transfer-based attacks by adjusting gradients[15-17]. Additional approaches[18-20] entail employing several models to enhance the adversarial transferability. Furthermore, some methods[13,16,21] seek to augment adversarial transferability through transformations.

Received date: 2023-12-07
Foundation items: Fundamental Research Funds for the Central Universities, China (No. 2232021A-10); Shanghai Sailing Program, China (No. 22YF1401300); Natural Science Foundation of Shanghai, China (No. 20ZR1400400); Shanghai Pujiang Program, China (No. 22PJ1423400)
* Correspondence should be addressed to HAO Kuangrong, email: krhao@dhu.edu.cn
Citation: ZHAO H Z, HAO L G, HAO K R, et al. Attention-guided sparse adversarial attacks with gradient dropout[J]. *Journal of Donghua University (English Edition)*, 2024, 41(5): 545-556.

Despite achieving significant success, these methods add perturbations to images at a global scale. This implies that every pixel of the input image is altered to craft adversarial examples. Though the perturbations are restricted to a very low level, almost imperceptible to the human eyes, it is believed that globally adding perturbations is not an optimal choice. Existing research indicates that different pixels contribute different to the outputs of DNNs[22-25]. Therefore, it is reasonable to assume and experimentally verify that perturbing different regions of the input image may produce different effects (see Section 3.5 for details). To this end, better ways to enhance the imperceptibility of adversarial examples while maintaining their effectiveness should be explored.

In this paper, a novel attention-guided sparse adversarial attack strategy with gradient dropout (AG&GD) is proposed. It can be easily integrated with any gradient-based method to reduce the intensity and the scale of adversarial perturbations. Specifically, during the gradient dropout (GD) phase, the gradient information corresponding to the pixels that have a relatively minor impact on the output of models in a single iteration is identified and randomly discarded, with the aim of reducing the intensity of perturbations. After completing all iterations, the scale of adversarial perturbations that have minimal impact on causing model misclassification is limited by using a soft mask-refined attention mechanism. Overall, the adversarial perturbation is limited from both the intensity and the scale, only adding perturbation to the foreground objects which are critical for the model classification, rather than overlaying it globally.

In essence, the contributions are as follows.

1) The limitations of existing gradient-based attack methods are analyzed and the existence of redundant perturbations is experimentally verified, which forms the foundation of the AG&GD strategy.

2) Based on these findings, the AG&GD strategy is presented, which enables the addition of adversarial perturbations to the original image in a more optimal manner.

3) Numerous experiments indicates that the AG&GD strategy has the potential to significantly reduce superfluous noise in adversarial examples while maintaining their attack capability.

4) Furthermore, a technique that maximizes attack efficiency by unleashing the maximum perturbation ϵ is evaluated, and a novel avenue to enhance the generalizability of adversarial examples is offered.

1 Related Work

1.1 Adversarial attack

Szegedy et al.[26] documented the susceptibility of DNNs and initially presented the box-constrained approach for creating adversarial examples. Afterwards, Goodfellow et al.[14] developed the first gradient-based technique known as the fast gradient sign method (FGSM) which generated highly transferable adversarial examples craftily using one gradient direction step. However, due to the fact that only one step was performed, FGSM was incapable of accurately fitting the decision boundary of the target model. To resolve this issue, Kurakin et al.[27] improved FGSM by executing several rounds of iterations. Based on these previous works, Dong et al.[15] introduced a momentum parameter to evade weak local maxima. Lin et al.[16] implemented the Nesterov accelerated gradient to adjust the prior accumulation of gradient to facilitate prudent prediction. Wang et al.[17] achieved gradient update stabilization by assimilating the gradient variance information from the last iteration. The proposed strategy can be combined with the aforementioned gradient-based adversarial attack methods to retain their original attack effectiveness while simultaneously constraining the magnitude and the range of perturbations.

1.2 Adversarial defense

Numerous defensive techniques have been proposed to counter adversarial examples. Goodfellow et al.[14] incorporated these examples into the training process to produce a more resilient model. Additionally, a considerable amount of research has been conducted on adversarial training[28-31]. Tramèr et al.[32] suggested an ensemble adversarial training approach to further improve the robustness of models by supplementing the training data with adversarial examples from multiple models. The defense models generated by the aforementioned technique are collectively referred to as adversarially trained models. Guo et al.[33] countered perturbations caused by adversarial attacks via input transformations. Xie et al.[34] lessened the efficacy of adversarial examples by unpredictably resizing and padding input images. Cohen et al.[35] accomplished a certifiably robust ImageNet classifier by employing randomized smoothing. Naseer et al.[36] proposed a self-supervised adversarial training mechanism to purify adversarial examples. In this paper, methods for adversarial robustness are evaluated by adversarially trained models.

1.3 Sparse adversarial attack

To enhance the imperceptibility of adversarial examples, several approaches have been proposed, including the Curls & Whey attack by Shi et al.[37]. They used gradient ascent and descent to obtain diverse adversarial examples and minimize noise. Liu et al.[38] developed a fixed-radius approach to craft examples with minimum-norm. Dong et al.[25] discovered in the experiment that adding perturbations on foreground objects could lead to more effective attacks, and this phenomenon suggested the presence of redundant perturbations. Therefore, they put forwards a superpixel-guided attentional adversarial attack method that modified only the pixels in salient regions by identifying the attention map of inputs. Moreover, Huang et al.[39] proposed a gradient-mask and attention-Whey method that reduced redundant noise by employing an attention mechanism

and a gradient optimization. The proposed strategy is based on enhanced attention mechanism and GD technique, which can be integrated with existing gradient-based attacks to reduce perturbations and maintain the attacking ability of examples at the same time.

2 Methodology

2.1 Motivation

While existing techniques for crafting adversarial examples have proven to be highly effective, they modify images on a global scale, affecting every pixel in the input image. Even though these modifications are minimal and typically go unnoticed by the human eyes, applying them across the entire image is arguably not the most efficient strategy.

Studies have shown that the importance of individual pixels varies in their impact on the outputs of DNNs[22-25]. Experimental evidence shows that perturbing different areas of an image can lead to varied outcomes (see Section 3.5 for details). With this insight, the research is directed towards finding more refined approaches that not only enhance the imperceptibility of adversarial examples but also retain their intended effectiveness.

2.2 Problem description

Adversarial attacks refer to the process of adding an imperceptible perturbation $\boldsymbol{\delta}$ to an image $\boldsymbol{x}$ with a known ground-truth label y^{true} in order to deceive a classifier into producing an incorrect output. This is accomplished by finding a value of $\boldsymbol{\delta}$ that satisfies the constraint $\|\boldsymbol{\delta}\|_p \leqslant \epsilon$, where $\|\cdot\|_p$ denotes the L_p norm of the perturbation. $p=\infty$ is set to align with the previous work in Ref. [15].

To achieve this, the classifier is represented as $f(\boldsymbol{x};\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the model's parameter, and $J(\boldsymbol{x}^{\text{adv}}, y^{\text{true}};\boldsymbol{\theta})$ is the loss function. Adversarial attacks can be formulated as a constrained optimization problem where the target is to maximize the loss function subject to the constraint that the perturbation is within the allowed range:

$$\arg\max_{x^{\text{adv}}} J(\boldsymbol{x}^{\text{adv}}, y^{\text{true}};\theta),\ s.t.\ \|\delta\|_p \leqslant \epsilon, \tag{1}$$

where $\boldsymbol{x}^{\text{adv}}$ is the resulting adversarial example, and is obtained by adding the perturbation to the original image $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \boldsymbol{\delta}$. For gradient-based attacks, crafting adversarial examples with momentum[15] can be simplified as

$$\boldsymbol{g}_t = \mu \boldsymbol{g}_{t-1} + \frac{\nabla_x J(\boldsymbol{x}_t^{\text{adv}}, y^{\text{true}};\boldsymbol{\theta})}{\|\nabla_x J(\boldsymbol{x}_t^{\text{adv}}, y^{\text{true}};\boldsymbol{\theta})\|_1},\ t=1,2,\cdots,T, \tag{2}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}} = \text{Clip}_{\boldsymbol{x}}^{\epsilon}\{\boldsymbol{x}_t^{\text{adv}} + \alpha \cdot \text{sign}(\boldsymbol{g}_t)\}, \tag{3}$$

where μ denotes the decay factor; $\boldsymbol{g}_t$ is the gradient at the tth iteration; $\alpha = \epsilon/T$, and T represents the number of iterations; $\text{sign}(\cdot)$ denotes the sign function; $\text{Clip}_x^{\epsilon}\{\cdot\}$ is utilized to constrain the perturbation.

2.3 AG&GD strategy

The AG&GD strategy effectively controls redundant adversarial perturbations in both the intensity and the scale, ensuring that perturbations are only added to foreground objects with great impact on the misclassification of the model, rather than low priority background regions.

The pipeline of the AG&GD strategy is presented in Fig. 1. It primarily comprises of two steps: limiting the magnitude of perturbation by GD; restricting the range of perturbation based on attention-guided (AG) mechanism.

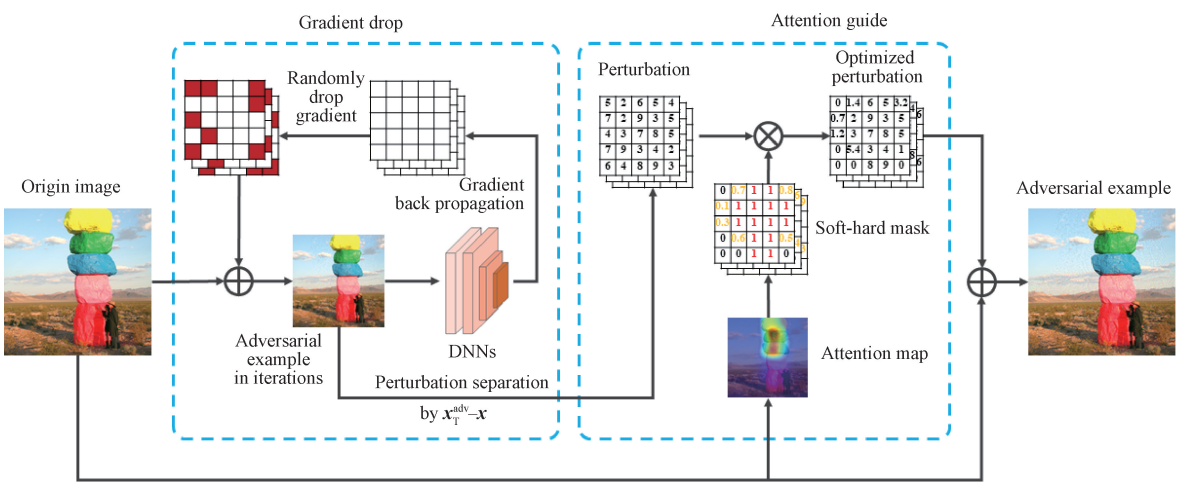

Fig. 1 Pipeline of AG&GD strategy

2.3.1 GD

According to Eqs. (2)-(3), in each iteration, the existing methods add the computed gradient $\boldsymbol{g}_t$ to the image without considering the varying impact of gradients on the prediction at different pixels. In fact, according to Ref. [39], to make adversarial examples closer to the decision boundary, it is preferable to prioritize updating the pixels with larger gradients, which leads to rapid loss growth and produce effective attacks. Therefore, to reduce redundant perturbations, a natural idea is to randomly dropout a subset of the pixels with smaller gradients for updating in a single iteration, which is

operationally similar to the dropout strategy used in training DNNs. In this way, redundant perturbations can be effectively controlled, while the pixels with smaller gradients may also be updated during T iterations. Specifically, in the tth iteration, a portion of the gradient information is randomly discarded with relatively small absolute values, such that the corresponding pixels do not receive updates, thereby reducing unnecessary perturbations. The index to remove ir is

$$ir = \text{rand}\{\min_{i=1\cdots n}(|\boldsymbol{g}_t|)\}, \tag{4}$$

where $\min_{i=1\cdots n}(\cdot)$ returns a set of n minimum absolute gradient values; $\text{rand}\{\cdot\}$ indicates the indices of a randomly selected subset of these gradient values that will be discarded.

2.3.2 Attention guide

According to Refs. [40 – 41], perturbing pixels within the highly-attended regions and altering key features that have the potential to transfer across different models can be used to craft adversarial examples, forming the basis of the AG&GD strategy. In this paper, layer class activation mapping (LayerCAM)[24] is employed to generate attention maps pertaining to the original category. However, it is worth noting that LayerCAM requires original gradient information, but in transfer-based attacks, it is entirely implausible to obtain any information from the target model aside from the predicted results. Thus, can we employ a surrogate model to produce attention maps for clean images? According to Refs. [13, 42], the attention maps of normally trained models are similar. Hence, we can utilize the attention maps generated by the surrogate model instead of the ones generated by the target model and apply them in transfer-based adversarial attacks. Consequently, the AG&GD strategy relies on the surrogate model to generate attention maps of images with respect to their ground-truth label by LayerCAM. Subsequently, we use the generated attention maps to restrict the scale of adversarial perturbations thereby achieving the objective of reducing redundant distortions.

To distinguish the foreground objects and background regions in attention maps created by LayerCAM, it is customarily practiced to employ a binary factor as a threshold[25] or retain the top 90% of pixels[43]. However, according to Ref. [44], certain surrounding areas of the foreground objects may also contribute to image classification. Therefore, it may not be the most optimal to superimpose perturbations only on the foreground objects. To this end, a progressive method utilizing a combination of soft-hard masks is adopted to progressively differentiate foreground from background, where all values of the soft-hard masks range between 0 and 1. The mask can be calculated as

$$mask_{i,j} = \text{Clip}_x^1\{4m_{i,j}(1 - m_{i,j}) + m_{i,j}^*\}, \tag{5}$$

where $\boldsymbol{m} \in \mathbf{R}^{H\times W}$ denotes the attention map; $4m_{i,j}(1 - m_{i,j})$ is used to progressively preserve the surrounding areas of foreground objects; $m_{i,j}^*$ represents the hard mask.

$$m_{i,j}^* = \begin{cases} 0, & m_{i,j} < \phi, \\ 1, & m_{i,j} < \phi, \end{cases} \tag{6}$$

where ϕ is a dynamic threshold, and varies for each input image. It can be utilized to dynamically extract the top $k\%$ pixels in the attention map. The attention guide allows us to preserve the foreground objects entirely ($mask_{i,j} = 1$) while partially retaining some surrounding regions of them ($0 < mask_{i,j} < 1$).

The AG&GD strategy can be conveniently combined with existing gradient-based adversarial attack methods. For instance, the AG MI-FGSM with GD, denoted as AG-GD-MI-FGSM, is elaborated in Algorithm 1. Similarly, the approach can also be integrated with I-FGSM[27], NI-FGSM[16] and VNI-FGSM[17], resulting in more superior AG-GD-I-FGSM, AG-GD-NI-FGSM and AG-GD-VNI-FGSM.

Algorithm 1: AG-GD-MI-FGSM

Input(s): An input image $\boldsymbol{x}$ with a ground-truth label y^{true} and a classifier $f(\boldsymbol{x};\boldsymbol{\theta})$ with a loss function $J(\boldsymbol{x}^{\text{adv}}, y^{\text{true}};\boldsymbol{\theta})$.

Input(s): The maximum perturbation ϵ, the iteration T and the decay factor μ.

Output(s): An adversarial example $\boldsymbol{x}^{\text{adv}}$.

$\alpha = \epsilon/T$, $\boldsymbol{g}_0 = 0$, $\boldsymbol{x}_1^{\text{adv}} = \boldsymbol{x}$

Based on the attention map of $\boldsymbol{x}$, derive the unique soft-hard ***mask*** by Eqs. (5)–(6)

for $t = 1 \rightarrow T$ **do**

  Update $\boldsymbol{g}_t$ with momentum by Eq. (2)

  Randomly select gradients to be discarded by Eq. (4)

  Dropout the selected gradients by $\boldsymbol{g}_t[ir] = 0$

  Update $\boldsymbol{x}_{t+1}^{\text{adv}}$ by Eq. (3)

end for

Separate and attenuate the noise by $\boldsymbol{noise} = \boldsymbol{mask} \cdot (\boldsymbol{x}_T^{\text{adv}} - \boldsymbol{x})$

return $\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \boldsymbol{noise}$

3 Experiments

Extensive experiments verify that the implementation of the AG&GD strategy on the baseline yields significantly reduced adversarial perturbations. Nevertheless, the associated adversarial examples retain their formidable attack potential. Furthermore, by adding slightly more noise, the AG&GD strategy can significantly enhance the efficiency of attacks.

3.1 Experimental setup

This section entails the exposition of experimental configuration.

Dataset. The experimentation is predicated on the NeurIPS 2017 adversarial dataset and the ILSVRC 2012 validation dataset, encompassing 1 000 images across 1 000 categories. Antecedent to being supplied to the model, all images are resized to 3 pixel×299 pixel×299 pixel($C\times H\times W$) during input preparation.

Models. Five undefended models and four defended models were cherry-picked to serve as the targets. For undefended models, the selected models comprise Inception-v3 (Inc-v3)[45], ResNet101 (Res101)[7], VGG19[6], DenseNet201 (Dense201)[46] and DPN98[47]. Moving on to the defended models, four adversarially trained models, namely adv-Inception-v3 (Inc-v3_{adv})[14], ens3-adv-Inception-v3 (Inc-v3_{ens3}), ens4-adv-Inception-v3 (Inc-v3_{ens4}) and ens-adv-InceptionResNet-v2 (Inc-Res-v2_{ens})[32], were assessed. Moreover, the impact of adversarial examples on two real-world image classification systems, provided by the Alibaba Cloud and the Baidu AI Cloud, were investigated.

Baselines. To demonstrate the significant improvement, the AG & GD strategy is combined with six gradient-based methods, namely: I-FGSM[27], MI-FGSM[15], NI-FGSM[16], VNI-FGSM[17], SMI-FGSM[48] and RI-FGSM[20].

Attack details. To concur with the prior research[15], $\epsilon = 16$ and $T = 10$. As in MI-FGSM, μ is set to 1.0. Additionally, for VNI-FGSM, the upper bound β and the number of sampled examples N are set to 1.5 and 20, respectively. In GD of the strategy, we randomly discard gradients that has the smallest 25% of absolute values, while in the AG phase, we set k to 75. To ensure the reliability of the findings, three homogeneous experiments were done, and the mean values were obtained. Figure 2 illustrates a comparison of different attack methods, and it is evident that AG-GD-MI-FGSM significantly reduces the adversarial perturbations. For the purpose of enhanced observation, the perturbations have been amplified by a factor of 8.

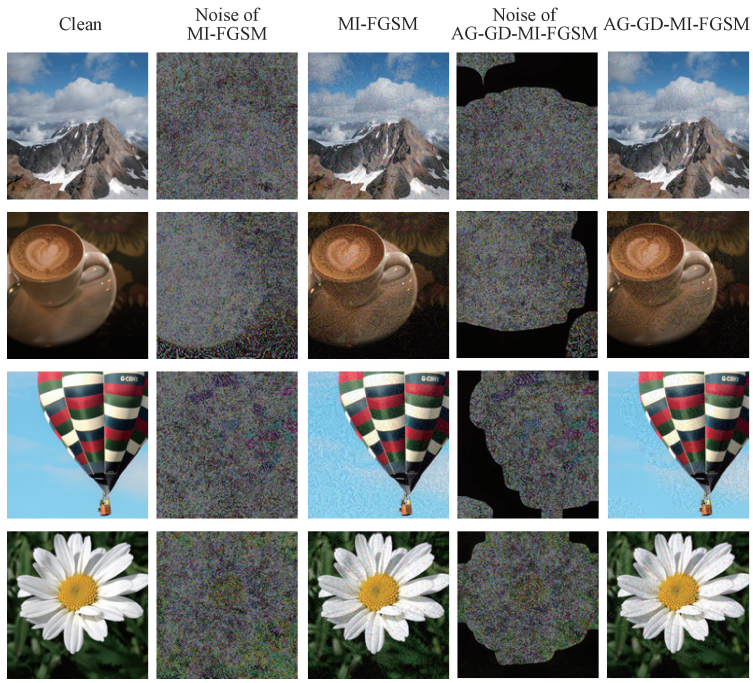

Fig. 2 Crafted adversarial perturbations and examples based on Inc-v3 with different attack methods

3.2 Attacks on undefended models

To validate the effectiveness of the AG&GD strategy, adversarial examples were initially generated using Inc-v3 as the surrogate model. Subsequently, the magnitude of perturbation was evaluated by three metrics: the L_2 distance (L_2-dist), the normalized learned perceptual image patch similarity (LPIPS^*)[49-50] and the normalized Fréchet inception distance (FID^*)[50-51]. While the efficacy of examples is evaluated by five normally trained models, namely Inc-v3, Res101, VGG19, Dense201 and DPN98. It is noteworthy that LPIPS^* can calculate the texture perception similarity and

FID* is useful to evaluate the fidelity. Relevant experimental outcomes are shown in Table 1 and Table 2.

Table 1 Attacks on undefended models: crafted adversarial examples on Inc-v3 based on NeurIPS 2017 adversarial dataset

Method	L_2-dist	LPIPS*	FID*	Efficacy/%				
				Inc-v3	Res101	VGG19	Dense201	DPN98
I-FGSM	10.19	**1.00**	0.90	**100.00**	**14.00**	**34.50**	**19.00**	**14.40**
AG-GD-I-FGSM	**8.45**	**1.00**	**0.91**	**100.00**	12.60	33.10	16.80	13.70
MI-FGSM	23.49	0.86	0.78	**100.00**	**30.20**	**57.90**	**41.13**	**33.07**
AG-GD-MI-FGSM	**19.88**	**0.95**	**0.81**	**100.00**	27.63	56.60	37.90	31.37
NI-FGSM	23.56	0.86	0.75	**100.00**	**34.50**	**63.20**	**46.73**	**36.87**
AG-GD-NI-FGSM	**19.96**	**0.94**	**0.78**	**100.00**	31.83	61.30	42.93	34.50
VNI-FGSM	22.87	0.88	0.71	**100.00**	**52.37**	**73.77**	**66.57**	**57.53**
AG-GD-VNI-FGSM	**19.56**	**0.96**	**0.74**	**100.00**	48.63	71.17	62.43	53.57
SMI-FGSM	26.04	0.84	0.71	**99.80**	**56.30**	**78.00**	**73.40**	**65.80**
AG-GD-SMI-FGSM	**21.93**	**0.92**	**0.74**	99.70	51.70	75.60	70.70	63.20
RI-FGSM	23.94	0.86	0.66	**97.90**	**65.70**	**84.90**	**83.20**	**76.60**
AG-GD-RI-FGSM	**20.32**	**0.93**	**0.70**	97.50	61.80	82.90	78.60	73.00

Table 2 Attacks on undefended models: crafted adversarial examples on Inc-v3 based on ILSVRC 2012 validation dataset

Method	L_2-dist	LPIPS*	FID*	Efficacy/%				
				Inc-v3	Res101	VGG19	Dense201	DPN98
I-FGSM	10.17	**1.00**	0.90	**100.00**	**8.80**	**23.80**	**19.00**	**10.50**
AG-GD-I-FGSM	**8.40**	**1.00**	**0.91**	**100.00**	**8.80**	**23.40**	**16.50**	**9.70**
MI-FGSM	**23.39**	**0.86**	**0.80**	**100.00**	**26.00**	**48.70**	**40.30**	**28.00**
AG-GD-MI-FGSM	**19.77**	**0.93**	**0.81**	**100.00**	25.30	47.80	37.30	25.80
NI-FGSM	23.43	0.85	0.76	**100.00**	**32.50**	**56.90**	**46.40**	**33.40**
AG-GD-NI-FGSM	**19.85**	**0.93**	**0.78**	**100.00**	30.10	55.20	43.60	30.90
VNI-FGSM	23.00	0.87	0.72	**100.00**	**48.60**	**69.90**	**66.20**	**57.30**
AG-GD-VNI-FGSM	**19.63**	**0.94**	**0.74**	**100.00**	47.00	68.30	61.00	53.50
SMI-FGSM	25.65	0.84	0.73	**99.70**	**52.80**	**74.00**	**72.90**	**62.70**
AG-GD-SMI-FGSM	**21.54**	**0.92**	**0.74**	**99.70**	49.80	73.90	68.90	59.00
RI-FGSM	23.81	0.85	0.70	**97.60**	**64.90**	**83.10**	**81.60**	**74.90**
AG-GD-RI-FGSM	**20.20**	**0.92**	**0.71**	96.70	61.90	81.40	79.00	71.80

The primary row pertains to the criteria for evaluating the magnitude of adversarial perturbations, along with the five target models subjected to attack. After incorporating the AG&GD strategy, the level of adversarial perturbation has substantially reduced, and the effectiveness of the generated adversarial examples has slightly weakened. For instance, in Table 1, compared to the traditional MI-FGSM, AG-GD-MI-FGSM could reduce the image distortion by 9.01% but only decrease the attack success rate (ASR) by 1.76% in average. Moreover, in Table 2, these metrics are 7.83% and 1.64%, respectively, showcasing the superior generalizability of the AG&GD strategy, unaffected by variations between training and testing datasets. The excellent outcome can be attributed to the AG&GD strategy. On the one hand, the magnitude of perturbations has been restricted, with the perturbations that have relatively minor impact on model misclassification being randomly discarded. On the other hand, the range of perturbations has also been limited by utilizing attention mechanism, with perturbations being confined to specific regions encompassing both the foreground objects and their surrounding areas.

Moreover, by slightly increasing ϵ in the AG-GD-MI-FGSM, the generated adversarial examples are comparable to those constructed by MI-FGSM in distortion with respect to the original images, while demonstrating a significant increase in the ASR. The relevant experiments and analyses are presented in Section 3.4.

3.3 Attacks on defended models

3.3.1 Attacks on adversarially trained models

According to Refs. [13, 43], the robustness of adversarially trained models against adversarial examples primarily stems from the feature where such models generate predictions based on discriminative regions that differ from those of normally trained models. Therefore,

the attack performance was assessed under the adversarially trained models. The relevant experimental results are shown in Table 3 and Table 4. It can be observed that the AG&GD strategy is capable of effectively reducing adversarial perturbations while maintaining excellent attack capability towards adversarial examples. Specifically, in Table 3, upon the integration of the strategy, the average level of noise injected into images experienced a decline of 8.32%. However, the average ASR against adversarially trained models only decreases by 0.34%, thereby providing further validation for the effectiveness of the AG&GD strategy. Furthermore, as depicted in Table 4, these metrics are respectively 7.72% and 0.00%. It is indicated that the robust generalizability of the AG&GD strategy is unaffected by disparities between training and test datasets, and when confronting adversarially trained models, the strategy can significantly mitigate extrinsic adversarial perturbations whilst fully maintaining the efficacy of adversarial examples.

Similarly, a slight increase in ϵ can significantly enhance the efficacy of the generated adversarial examples, while still keeping their distortion within manageable limits. More details are in Section 3.4.

Table 3 Attacks on adversarially trained models: crafted adversarial examples on Inc-v3 based on NeurIPS 2017 adversarial dataset

Method	L_2-dist	LPIPS *	FID *	Efficacy/%			
				Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Inc-Res-v2$_{ens}$
I-FGSM	10.19	**1.00**	0.90	**10.80**	**10.20**	**11.10**	**4.70**
AG-GD-I-FGSM	**8.45**	**1.00**	**0.91**	10.10	7.30	9.70	4.10
MI-FGSM	23.49	0.86	0.78	**23.93**	**17.37**	**17.03**	**9.00**
AG-GD-MI-FGSM	**19.88**	**0.95**	**0.81**	23.00	15.80	15.70	8.90
NI-FGSM	23.56	0.86	0.75	24.17	**17.30**	**16.20**	8.17
AG-GD-NI-FGSM	**19.96**	**0.94**	**0.78**	**24.43**	15.27	16.00	**8.23**
VNI-FGSM	22.87	0.88	0.71	43.03	**37.60**	37.13	**22.30**
AG-GD-VNI-FGSM	**19.56**	**0.96**	**0.74**	**43.33**	36.27	**37.33**	22.27
SMI-FGSM	26.04	0.84	0.71	45.00	44.00	42.10	25.20
AG-GD-SMI-FGSM	**21.93**	**0.92**	**0.74**	**46.60**	**44.50**	**42.20**	**26.00**
RI-FGSM	23.94	0.86	0.66	54.40	49.10	48.10	**28.30**
AG-GD-RI-FGSM	**20.32**	**0.93**	**0.70**	**55.30**	**49.60**	**48.20**	28.00

Table 4 Attacks on adversarially trained models: crafted adversarial examples on Inc-v3 based on ILSVRC 2012 validation dataset

Method	L_2-dist	LPIPS *	FID *	Efficacy/%			
				Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Inc-Res-v2$_{ens}$
I-FGSM	10.17	**1.00**	0.90	6.00	**4.60**	4.70	**2.90**
AG-GD-I-FGSM	**8.40**	**1.00**	**0.91**	**6.20**	4.00	**5.20**	2.80
MI-FGSM	23.39	0.86	0.80	17.70	**12.90**	**12.30**	6.20
AG-GD-MI-FGSM	**19.77**	**0.93**	**0.81**	**18.50**	12.80	12.00	6.30
NI-FGSM	23.43	0.85	0.76	**20.20**	**12.60**	**13.60**	**7.00**
AG-GD-NI-FGSM	**19.85**	**0.93**	**0.78**	19.50	11.70	12.60	6.30
VNI-FGSM	23.00	0.87	0.72	39.50	**34.90**	**33.00**	**17.50**
AG-GD-VNI-FGSM	**19.63**	**0.94**	**0.74**	**39.70**	32.50	32.80	16.90
SMI-FGS	25.65	0.84	0.73	39.50	39.90	**36.90**	19.90
AG-GD-SMI-FGSM	**21.54**	**0.92**	**0.74**	**40.60**	**42.30**	36.70	21.90
RI-FGS	23.81	0.85	0.70	52.70	**50.00**	46.00	**28.00**
AG-GD-RI-FGSM	**20.20**	**0.92**	**0.71**	**54.40**	48.50	**46.40**	27.90

3.3.2 Attacks on real-world application programming interfaces (APIs)

In order to assess the practicality of the AG&GD strategy, the APIs provided by real-world image classification systems were attacked to select the prediction with the highest confidence score as the label. The APIs were furnished by the Baidu AI Cloud and the Alibaba Cloud. Experiments were conducted using adversarial examples crafted based on Inc-v3. The results are shown in Fig. 3. It is noteworthy that the adversarial examples utilized in attacks are identical to the ones employed in Section 3.2. Hence, the magnitude of their perturbations can be referenced from Table 1. Consistent with previous findings, introducing the AG&GD strategy

to baselines results in a slight decrease in the ASR, which indicates that certain perturbations excluded from the attention map may still have a subtle impact on image classification tasks. Notwithstanding, the AG&GD strategy still retains the ability to effectively control noise levels while simultaneously preserving notable attack efficacy.

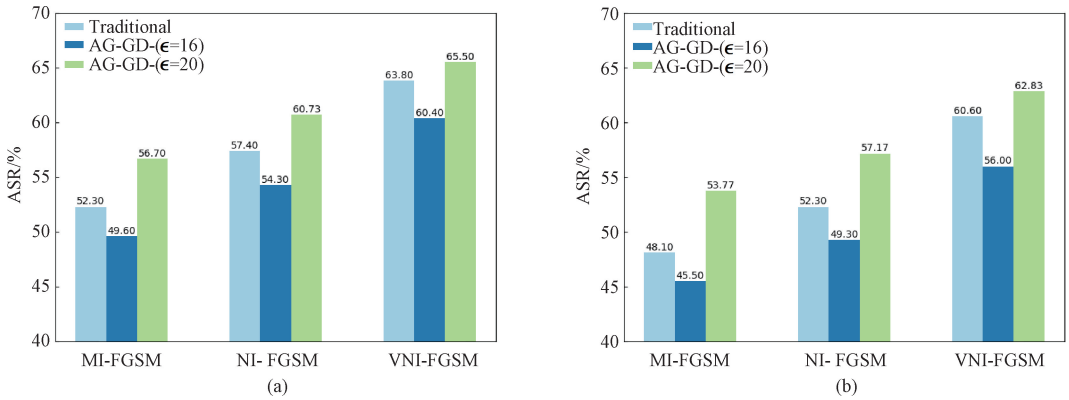

Fig. 3 Results of attacks on real-world APIs: (a) based on Baidu AI Cloud; (b) based on Alibaba Cloud

3.4 Unleashing ϵ: maximizing the benefits

Based on previous findings, the implementation of the AG&GD strategy leads to a slight decrease in the efficacy of adversarial examples. This can be attributed to the strategy's emphasis on preserving perturbations on foreground objects while discarding relatively insignificant ones in background regions. Such an operation can significantly reduce the overall perturbations while maintaining the attack efficacy. However, those discarded perturbations may still have some effect on model misclassification.

To maximize benefits, ϵ increases from 16 to 20, which makes the overall perturbation level of generated adversarial examples similar to that produced by traditional methods. In other words, perturbations in background regions are restricted as much as possible while the upper limit of perturbations increases, thus making the perturbations overlaid on foreground objects more effective. Despite this, the overall level of perturbation in adversarial examples remains comparable to that produced by the original method. Relevant experimental results based on Inc-v3 are presented in Table 5 and Table 6, as well as Fig. 3.

For both the undefended models in Table 5 and the adversarially trained models in Table 6, compared to the traditional VNI-FGSM, the AG-GD-VNI-FGSM increases the average ASR by 4.37% and 6.77%, respectively, with only a marginal increase of 1.48% in the average perturbation. Additionally, in Fig. 3, the AG&GD strategy improves the attack efficacy on average by 3.70%.

Table 5 Attacks on undefended models: crafted adversarial examples on Inc-v3 with ϵ of 20

Method	L_2-dist	LPIPS*	FID*	Efficacy/%				
				Inc-v3	Res101	VGG19	Dense201	DPN98
AG-GD-I	10.46	1.00	0.88	100.00	14.50	37.30	20.50	16.70
AG-GD-MI	24.72	0.87	0.76	100.00	34.40	63.77	46.53	36.50
AG-GD-NI	24.79	0.86	0.73	100.00	40.60	68.67	52.20	41.00
AG-GD-VNI	24.09	0.89	0.71	100.00	58.77	78.10	71.27	63.93
AG-GD-SMI	27.20	0.85	0.69	99.80	59.00	80.30	75.50	68.70
AG-GD-RI	25.21	0.86	0.66	98.50	69.20	86.10	84.30	78.70

Table 6 Attacks on adversarially trained models: crafted adversarial examples on Inc-v3 with ϵ of 20

Method	L_2-dist	LPIPS*	FID*	Efficacy/%			
				Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Inc-Res-v2$_{ens}$
AG-GD-I	10.46	1.00	0.88	11.30	8.20	9.60	5.20
AG-GD-MI	24.72	0.87	0.76	28.20	18.43	18.80	11.17
AG-GD-NI	24.79	0.86	0.73	30.23	18.73	18.90	11.13
AG-GD-VNI	24.09	0.89	0.71	52.17	42.73	43.23	29.00
AG-GD-SMI	27.20	0.85	0.69	51.00	48.30	47.00	30.10
AG-GD-RI	25.21	0.86	0.66	60.20	55.10	54.50	33.80

Overall, the AG&GD strategy enhances the efficiency of the attack with a slight increase in perturbation, presenting a novel perspective for boosting the robustness of adversarial examples and exploring the vulnerability of models.

3.5 Discussion

Might perturbing different regions of the input image produce different effects? Different pixels of images contribute differently to the outputs of DNNs[22-25]. Therefore, it is reasonable to assume that perturbations applied to different regions of input images may have different effects. To validate this hypothesis, the Inc-v3 was employed as the surrogate model to create a variety of adversarial examples for attacking target models, as indicated in Table 7. The findings reveal that adding noise in foreground regions is considerably more effective than in background regions.

Table 7 Attacks on undefended models: crafted adversarial examples based on diverse regions

Method	Efficacy/%				
	Inc-v3	Res101	VGG19	Dense201	DPN98
MI (foreground)	100.00	26.60	56.80	35.90	28.80
MI (background)	20.40	5.10	16.60	5.30	5.00
MI	100.00	30.20	57.90	41.13	33.07

Is each step of the strategy effective? To ascertain the effectiveness of each step of the strategy, further experiments were conducted using the Rand-GD-MI-FGSM and AG-Rand-MI-FGSM for attacking. For the steps replaced by Rand, a proportionate number of pixels was randomly selected to attack compared to the original steps. As indicated in Table 8, the approach achieved the best results on all target models in the black-box setting, thus validating the effectiveness[22-25]. The AG&GD strategy aids in constructing more transferable adversarial examples at the same level of distortion, facilitating exploration of the vulnerability of the models.

Table 8 Performance of Rand-GD-MI-FGSM and AG-Rand-MI-FGSM for attacking

Surrogate	Method	Efficacy/%				
		Inc-v3	Res101	VGG19	Dense201	DPN98
Inc-v3	Rand-GD-MI-FGSM	**100.00**	21.83	47.77	31.47	25.30
	AG-Rand-MI-FGSM	99.87	15.20	36.80	21.37	18.40
	AG-GD-MI-FGSM	**100.00**	**27.63**	**56.60**	**37.90**	**31.37**
Res101	Rand-GD-MI-FGSM	22.60	**98.80**	40.40	30.57	28.80
	AG-Rand-MI-FGSM	15.79	94.09	33.34	22.33	21.30
	AG-GD-MI-FGSM	**30.90**	97.70	**50.50**	**38.13**	**36.03**

Does redundant perturbation really exist? In Fig. 4, the performance of the AG-GD-VNI-FGSM in attacking different proportions of pixels under the black-box setting was compared. The adversarial examples were constructed based on Inc-v3 and used to attack four normally trained models, namely Res101, VGG19, Dense201 and DPN98. Statistical results indicate that the increase in the ASR becomes less significant when the percentage of attacked pixels is over 60%, and when 80% of pixels are attacked, the performance of the AG-GD-VNI-FGSM is comparable to that of the traditional VNI-FGSM. As the percentage of attacked pixels increases, the L_2-dist between adversarial examples and original images also increases, and the upwards trend is consistently significant. These findings indicate that as the proportion of attacked pixels increases, the impact of newly added perturbations on model misclassification decreases. Therefore, in a sense, redundant perturbation does exist, and the AG&GD strategy can effectively limit the intensity and the scale of adversarial perturbations, ensuring that they are primarily superimposed on foreground objects that models are more attentive to.

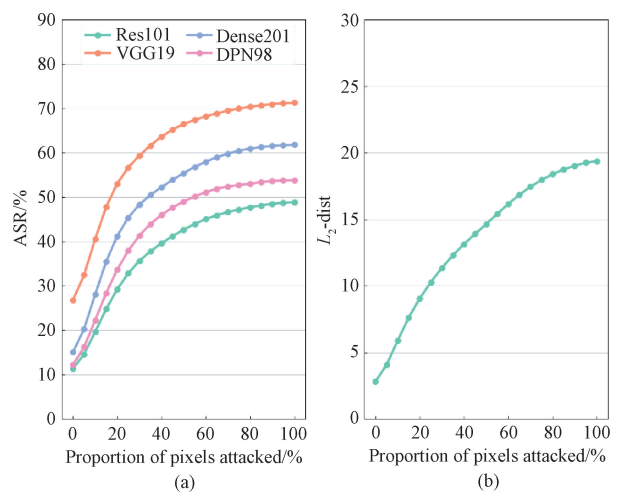

Fig. 4 Performance of AG-GD-VNI-FGSM in attacking different proportions of pixels under black-box setting: (a) ASR; (b) L_2-dist

4 Conclusions and Future Work

In this paper, the limitations of the existing adversarial attack methods were analyzed and it was pointed out that these methods added perturbations with the same weight to every pixel of the image, resulting in unnecessary noise to adversarial examples. To address this issue, the AG&GD strategy was proposed, which ensured that adversarial perturbations were mainly added to foreground objects. The experimental results showed that the AG&GD strategy could significantly reduce redundant noise in adversarial examples, while still maintaining their attack effectiveness. In attacks on undefended models and adversarially trained models, integrating the strategy led to an 8.32% decrease in the average noise level of the injected images, while the average ASR decreased by only 2.27% and 0.34%, respectively. Moreover, the efficacy of attacks can be significantly enhanced by incorporating a minor degree of disruption.

In future work, the adaptability of the AG&GD strategy across various deep learning models would be enhanced and its efficiency would be further refined by exploring advanced attention mechanisms and gradient manipulation techniques. Additionally, integrating the AG&GD strategy with other adversarial attack frameworks to broaden its applicability and investigating defensive measures against such sophisticated attacks will be crucial. This dual approach will not only advance the field of adversarial machine learning by developing more robust and stealthy attack methods but also contribute to the creation of more secure artificial intelligence systems capable of resisting these advanced threats.

References

[1] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(4): 640-651.

[2] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 834-848.

[3] GIRSHICK R. Fast R-CNN [C]//2015 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2015: 1440-1448.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 779-788.

[5] LI C Z, WEI K H, ZHAO Y B, et al. Improvement of high-speed detection algorithm for nonwoven material defects based on machine vision [J]. *Journal of Donghua University (English Edition)*, 2024, 41(4): 416-427.

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2023-12-01]. https://arxiv.org/pdf/1409.1556.

[7] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 770-778.

[8] LUO X, XIA D M, TAO R, et al. Fabric image retrieval based on fine-grained features [J]. *Journal of Donghua University (English Edition)*, 2024, 41(2): 115-129.

[9] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial attacks and defences competition [M]//The NIPS'17 Competition: Building Intelligent Systems. Cham: Springer International Publishing, 2018: 195-231.

[10] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information [C]//35th International Conference on Machine Learning (ICML). New York: ACM, 2018: 2137-2146.

[11] VIDNEROVÁ P, NERUDA R. Vulnerability of classifiers to evolutionary generated adversarial examples [J]. *Neural Networks*, 2020, 127: 168-181.

[12] ZHOU W, HOU X, CHEN Y J, et al. Transferable adversarial perturbations [M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 471-486.

[13] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 4307-4316.

[14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples [EB/OL]. (2015-03-20) [2023-12-01]. https://arxiv.org/pdf/1412.6572.

[15] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 9185-9193.

[16] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks [EB/OL]. (2020-02-03) [2023-12-01]. https://arxiv.org/pdf/1908.06281.

[17] WANG X S, HE K. Enhancing the transferability of adversarial attacks through variance tuning [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New

York: IEEE, 2021: 1924-1933.
[18] LIU Y P, CHEN X Y, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[EB/OL]. (2017-02-07)[2023-12-01]. https://arxiv.org/pdf/1611.02770.
[19] HAO L G, HAO K R, WEI B, et al. Boosting the transferability of adversarial examples via stochastic serial attack[J]. *Neural Networks*, 2022, 150: 58-67.
[20] ZHAO H Z, HAO L G, HAO K R, et al. Remix: towards the transferability of adversarial examples[J]. *Neural Networks*, 2023, 163: 367-378.
[21] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 2725-2734.
[22] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 2921-2929.
[23] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. *International Journal of Computer Vision*, 2020, 128(2): 336-359.
[24] JIANG P T, ZHANG C B, HOU Q B, et al. LayerCAM: exploring hierarchical class activation maps for localization[J]. *IEEE Transactions on Image Processing*, 2021, 30: 5875-5888.
[25] DONG X Y, HAN J F, CHEN D D, et al. Robust superpixel-guided attentional adversarial attack[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 12892-12901.
[26] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. (2014-02-19)[2023-12-01]. https://arxiv.org/pdf/1312.6199.
[27] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2017-02-11)[2023-12-01]. https://arxiv.org/pdf/1607.02533.
[28] SONG C B, HE K, LIN J D, et al. Robust local features for improving the generalization of adversarial training[EB/OL]. (2020-02-02)[2023-12-01]. https://arxiv.org/pdf/1909.10147.
[29] ZHANG S F, HUANG K Z, ZHU J K, et al. Manifold adversarial training for supervised and semi-supervised learning[J]. *Neural Networks*, 2021, 140: 282-293.
[30] CHEN S H, SHEN H J, WANG R, et al. Towards improving fast adversarial training in multi-exit network[J]. *Neural Networks*, 2022, 150: 1-11.
[31] LAMB A, VERMA V, KAWAGUCHI K, et al. Interpolated adversarial training: achieving robust neural networks without sacrificing too much accuracy[J]. *Neural Networks*, 2022, 154: 218-233.
[32] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[EB/OL]. (2020-04-26)[2023-12-01]. https://arxiv.org/pdf/1705.07204.
[33] GUO C, RANA M, CISSÉ M, et al. Countering adversarial images using input transformations[EB/OL]. (2018-01-25)[2023-12-01]. https://arxiv.org/pdf/1711.00117.
[34] XIE C H, WANG J Y, ZHANG Z S, et al. Mitigating adversarial effects through randomization[EB/OL]. (2018-02-28)[2023-12-01]. https://arxiv.org/pdf/1711.01991.
[35] COHEN J M, ROSENFELD E, KOLTER J Z. Certified adversarial robustness via randomized smoothing[C]//36th International Conference on Machine Learning (ICML). New York: ACM, 2019: 1310-1320.
[36] NASEER M, KHAN S, HAYAT M, et al. A self-supervised approach for adversarial robustness[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 259-268.
[37] SHI Y C, WANG S Y, HAN Y H. Curls & Whey: boosting black-box adversarial attacks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2019: 6512-6520.
[38] LIU F C, ZHANG C, ZHANG H Y. Towards transferable adversarial perturbations with minimum norm[C]//38th International Conference on Machine Learning (ICML). New York: ACM, 2021: 1-9.
[39] HUANG Q F, LIAN Z C, LI Q M. Attention based adversarial attacks with low perturbations[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). New York: IEEE, 2022: 1-6.
[40] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[EB/OL]. (2019-08-12)[2023-12-01]. https://arxiv.org/pdf/1905.02175v4.
[41] LIN C H, HAN S C, ZHU J L, et al. Sensitive region-aware black-box adversarial attacks[J]. *Information Sciences*, 2023, 637: 118929.
[42] TSIPRAS D, SANTURKAR S, ENGSTROM L, et al. Robustness may be at odds with accuracy[EB/OL]. (2019-09-09)[2023-12-01]. https://arxiv.org/pdf/1805.12152.
[43] LI C, YAO W, WANG H D, et al. Adaptive

momentum variance for attention-guided sparse adversarial attacks [J]. *Pattern Recognition*, 2023, 133: 108979.
[44] YANG R J, GUO Y F, WANG R K, et al. Exploring the impact of adding adversarial perturbation onto different image regions [C]// 2022 IEEE International Symposium on Circuits and Systems (ISCAS). New York: IEEE, 2022: 2363-2367.
[45] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 2818-2826.
[46] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 2261-2269.
[47] CHEN Y P, LI J N, XIAO H X, et al. Dual path networks [EB/OL]. (2017-08-01) [2023-12-01]. https://arxiv.org/pdf/1707.01629V2.
[48] WANG G Q, YAN H Q, WEI X X. Enhancing transferability of adversarial examples with spatial momentum [M]//Pattern Recognition and Computer Vision. Cham: Springer International Publishing, 2022: 593-604.
[49] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2018: 586-595.
[50] HE Z Y, DUAN Y X, ZHANG W, et al. Boosting adversarial attacks with transformed gradient [J]. *Computers & Security*, 2022, 118: 102720.
[51] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6629-6640.

基于梯度丢弃和注意力引导的稀疏对抗攻击

赵鸿志[1,2], 郝灵广[1,2], 郝矿荣[1,2*], 隗 兵[1,2], 刘肖燕[1,2]
1. 东华大学 信息科学与技术学院, 上海 201620
2. 东华大学 数字化纺织服装技术教育部工程研究中心, 上海 201620

摘 要: 深度神经网络极易受到外部有意生成的对抗样本的影响, 这些对抗样本是通过在干净图像上叠加微小的噪声来实现的。然而, 大多数现有的基于转移的攻击方法选择在原始图像的每个像素上以相同的权重添加扰动, 导致对抗样本出现冗余噪声, 使其更容易被检测系统识别。鉴于此, 该文引入了一种新颖的由注意力引导的稀疏对抗攻击策略, 该策略结合了梯度丢弃技术, 可以与现有的基于梯度的算法结合使用, 从而最小化扰动的强度和规模, 同时确保对抗样本的有效性。具体而言, 在梯度丢弃阶段, 策略随机丢弃一些相对不重要的梯度信息, 以限制扰动的强度; 在注意力引导阶段, 通过使用软掩码优化的注意力机制评估每个像素对模型输出的影响, 并限制对输出影响较小的像素的扰动, 以控制扰动的规模。在 NeurIPS 2017 对抗数据集和 ILSVRC 2012 验证数据集上的大量实验证明了该策略可以显著减少对抗样本中的冗余噪声, 同时保持算法的攻击效果。例如, 在对于对抗训练模型的攻击中, 将对抗攻击算法引入该策略后, 注入图像的平均噪声水平下降了 8.32%, 而平均攻击成功率仅下降了 0.34%。此外, 只需引入少量扰动, 该策略便能显著提高攻击成功率。

关键词: 深度神经网络; 对抗攻击; 稀疏对抗攻击; 对抗转移性; 对抗样本