Rethinking Adversarial Examples Exploiting Frequency-Based Analysis

Sicong Han[1], Chenhao Lin[1(✉)], Chao Shen[1(✉)], and Qian Wang[2]

[1] Xi'an Jiaotong University, Xi'an, China
siconghan@stu.xjtu.edu.cn, linchenhao@xjtu.edu.cn,
chaoshen@mail.xjtu.edu.cn
[2] Wuhan University, Wuhan, China
qianwang@whu.edu.cn

Abstract. Deep neural networks (DNNs) have been recently found vulnerable to adversarial examples. Several previous works attempt to relate the low-frequency or high-frequency parts of adversarial inputs with the robustness of models. However, these studies lack comprehensive experiments and thorough analyses and even yield contradictory results. This work comprehensively explores the connection between the robustness of models and properties of adversarial perturbations in the frequency domain using six classic attack methods and three representative datasets. We visualize the distribution of successful adversarial perturbations using Discrete Fourier Transform and test the effectiveness of different frequency bands of perturbations on reducing the accuracy of classifiers through a proposed quantitative analysis. Experimental results show that the characteristics of successful adversarial perturbations in the frequency domain can vary from dataset to dataset, while their intensities are greater in the effective frequency bands. We analyze the obtained phenomena by combining principles of attacks and properties of datasets and offer a complete view of adversarial examples from the frequency domain perspective, which helps to explain the contradictory parts of previous works and provides insights for future research.

Keywords: Adversarial examples · Model robustness · Frequency analysis

1 Introduction

With the widespread deployment of deep learning systems in various fields, the robustness of deep learning models has become of paramount importance. Deep learning models are highly vulnerable to adversarial examples [19], which may lead to serious security breaches and irreparable financial loss as they have been integrated into various safety-critical systems, e.g., self-driving cars.

Adversarial examples, in the context of the image classification, look similar to the original images while they have the ability to fool the model into producing incorrect outputs with high confidence [5]. Numerous adversarial attack

D. Gao et al. (Eds.): ICICS 2021, LNCS 12919, pp. 73–89, 2021.
https://doi.org/10.1007/978-3-030-88052-1_5

methods [1,13] and defense methods [10,25] have been proposed to improve the effectiveness of attacks and enhance models' robustness, respectively. Meanwhile, the studies on understanding the intrinsic nature of adversarial examples have attracted increasing attention with various theories, including Linearity hypothesis [5], boundary tilting [20] and curvature of decision boundaries [12].

Recently, some studies attempt to understand the adversarial examples from the perspective of the frequency domain. It is argued in [21] that the generalization of convolutional neural networks (CNNs) can be related to how they process the high-frequency information of images. Besides, the performance of the attacks can be promoted by finding adversarial perturbations in specific frequency bands [4,6,16]. However, a small number of attacks and a single dataset are usually adopted in previous works, resulting in weak generalization and some contradictory parts among proposed theories.

To address the limitations of existing works, we comprehensively analyze and evaluate the adversarial examples in the frequency domain. Two categories of six attack methods in total and three commonly used datasets are adopted in this paper. We first visualize the distribution of frequency components of successful adversarial perturbations, then we decompose them into signals of two frequencies to explore which one plays a leading role in reducing the accuracy of the model. We further propose an evaluation method to quantitatively compare the differences between the distribution of successful perturbations and unsuccessful ones within the effective frequency bands to figure out what leads to the different effects of both perturbations. The bridge between the robustness of models and the properties of adversarial perturbations in the frequency domain on different datasets is built in this paper. Based on the observations and analyses, we draw several important findings and provide insights for future research:

- Unlike previous works [22,23], in which they claim that successful adversarial perturbations for naturally trained models concentrate more on the high-frequency domain, our results suggest that the distribution of successful adversarial perturbations can vary from dataset to dataset, so as the effectiveness of different frequency bands of successful perturbations on misleading models. As a result, it is not accurate to solely relate high-frequency components of images with the target features of adversarial attacks. Besides, filtering the fixed frequency bands of information of images cannot provide universal defense.
- Previous studies [4,6] illustrate that constraining adversarial directions in different frequency bands on ImageNet can obtain progress in improving the effectiveness of attacks. However, the principles behind those results are not provided systematically. Our findings and analyses show that both low- and high-frequency components of successful adversarial perturbations on ImageNet have a noticeable effect on reducing the model's accuracy. Therefore, performing adversarial attacks in specific frequency bands can be a reasonable way to promote the performance of attacks on this dataset, explaining the rationality of previous works.

- Furthermore, we find the effective frequency bands for various attacks under changeable constraints, within which the successful adversarial perturbations can reduce the accuracy of the model rapidly. Through the proposed quantitative comparison between the intensities of the distribution of both successful and unsuccessful perturbations, we conclude that within the effective frequency bands, the distribution of successful perturbations has larger intensity upon most occasions, thus misleading the model.
- We associate the characteristics of successful adversarial examples in the frequency domain with the space occupied by the objects that belong to the ground truth classes, which is a noticeable difference among the three datasets. By exploiting the rules we found, it is possible to establish the defensive measure from the perspective of the frequency domain by adopting the properties of the datasets or usage scenarios as the prior information, but providing high-performance universal defense is still a challenge.

2 Related Work

Various viewpoints have been introduced to understand the adversarial examples and explore how they mislead the DNNs. Szegedy *et al.* [19] argued that those adversarial examples were rarely observed in the test set due to the extremely low probability, while they were actually dense and thus could be found in every test case. Goodfellow *et al.* [5] illustrated that the infinitesimal changes to the input of a simple linear model could accumulate to one large change to the output due to the sufficient dimensionality. Moosavi-Dezfooli *et al.* [11] associated the robustness of DNNs to the curvature of decision boundaries.

The counter-intuitive phenomenon brought by adversarial examples arouses the discussion about the gap in the visual information processing between humans and machines. Intuitively, human visual sensitivity for the different frequency components of the images can be various. As a result, analyzing adversarial examples from the frequency perspective provides a possible way to explore the robustness of DNNs [22]. Wang *et al.* [21] found that the CNN could make correct predictions using only high-frequency counterparts of images, which were not perceivable to humans. They attempted to exploit the high-frequency components of images to explain the trade-off between the accuracy and the robustness of CNNs. Yin *et al.* [23] demonstrated that adversarial perturbations generated towards a naturally trained model concentrated on the high-frequency domain, while after adversarial training, those perturbations became more low-frequency.

In addition to exploring the distribution of target features of adversarial attacks, several works focused on improving the strength of adversarial examples by finding perturbations in different frequency bands. Guo *et al.* [6] illustrated that adversarial directions might occur in high density in the low-frequency subspace of images. Therefore, finding adversarial perturbations in the low-frequency domain could result in improving the query efficiency of attacks. Besides, low-frequency

perturbations were showed to be highly effective against defended models [16]. On the other hand, performing universal attacks on middle and high frequency bands could balance the fooling rates and perceptiveness [4].

In the literature related to analyzing adversarial examples in the frequency domain, proposed views align well with the local empirical observations. However, these views can be contradictory to each other. Finding adversarial perturbations in low-frequency [6] and high-frequency bands [4], as mentioned above, are examples that are not consistent well with each other. Besides, a single dataset and a small number of attack methods with limited norms of constraints are used in previous works, which results in weak consistency of existing illustrations. Therefore, a comprehensive analysis of adversarial examples in the frequency domain on multiple datasets and diverse attack methods with changeable constraints are urgently required to address the limitations of existing works.

3 Methodology

We first define the basic notations used in this paper. $\mathcal{F} : x \in \mathbb{R}^{d\times d} \rightarrow z \in \mathbb{R}^k$ is defined as a neural network which takes x as an input and outputs logits z, where k is the number of classes. We denote the ℓ_p norm as $\|\cdot\|_p$, and $p \in \{2, \infty\}$ is considered in this paper. The image x that can be correctly classified as its ground truth label y will be attacked to generate corresponding adversarial example x^{adv}. Let $v = f(x)$ denote the Discrete Fourier Transform (DFT) of x and $f(\cdot)^{-1}$ represent the inverse DFT (IDFT), where $v \in \mathbb{C}^{d\times d}$ and $v(i, j)$ represents the value of v at position (i, j). Low-frequency components are shifted to the center when we visualize the frequency spectra of adversarial perturbations.

3.1 Adversarial Attack Methods

We adopt untargeted white- and black-box attacks in this paper, due to targeted attacks may make biases on obtained perturbations when targets are specified. The principles of six used attacks are illustrated as follows:

FGSM Attack. Fast Gradient Sign Method (FGSM) [5] focuses on efficiently generating adversarial examples. By using the gradients of the loss function $\mathcal{L}(x, y; \mathcal{F})$, pixels of the original example are modified to increase the loss in a single step. Formally, FGSM Attack can be expressed as:

$$x^{adv} = \text{clip}\{x + \epsilon \cdot sign(\nabla_x \mathcal{L}(x, y; \mathcal{F}))\}, \tag{1}$$

where ϵ is the constraint that ensures the ℓ_∞ perturbation is small enough to be undetectable, and the clip function forces x^{adv} to be a legitimate image.

BIM Attack. Basic Iterative Method (BIM) [8] is an iterative variant of FGSM, which follows the update rule:

$$x_0^{adv} = x,\ x_{n+1}^{adv} = \text{project}\{x_n^{adv} + \alpha \cdot sign(\nabla_x \mathcal{L}(x_n^{adv}, y; \mathcal{F}))\}, \tag{2}$$

where α is the step size, the number of iterations n is set to be $\min(\epsilon + 4, 1.25\epsilon)$, and the project function keeps x_{n+1}^{adv} residing in both ℓ_∞ ϵ-neighbourhood of the original image x and the image value range.

PGD Attack. Compared with BIM, Projected Gradient Descent (PGD) [10] has more iterations and performs random starts as the initialization to improve the diversity. It is a strong first-order adversary that can be expressed as:

$$x_0^{adv} = \text{clip}(x + \mathcal{S}),\ x_{n+1}^{adv} = \text{project}\{x_n^{adv} + \alpha \cdot sign(\nabla_x \mathcal{L}(x_n^{adv}, y; \mathcal{F}))\}, \tag{3}$$

where $\mathcal{S} \in \mathbb{R}^{d \times d}$ is the random vector which is chosen from the uniform distribution of $[-\epsilon, \epsilon]$, the clip function makes x_0^{adv} stay in the image value range, and the project function keeps the generated adversarial example within in both ℓ_∞ ϵ-neighbourhood of the original image and the image value range.

DeepFool Attack. The aim of DeepFool Attack [13] is to compute a minimal ℓ_2 perturbation for the target image. This method starts from the original image and calculates the vector, leading the image to step over the decision boundary of the approximated polyhedron iteratively.

CW Attack. The untargeted version of Carlini & Wagner (CW) Attack [2] uses a new loss function to maximize the distance between the ground truth class y and the most-likely class outside of y, which can be expressed as:

$$\mathcal{L}_{CW}(x, y; \mathcal{F}) = \max(\mathcal{F}(x)_{(y)} - \max_{i \neq y} \mathcal{F}(x)_{(i)}, -\kappa), \tag{4}$$

where $\mathcal{F}(x)_{(t)}$ represents the output logit of class t and κ encourages the solver to find the adversarial example that decreases the original class's prediction probability with high confidence. The ℓ_2 perturbation δ is optimized as follows:

$$x^{adv} = x + \arg\min_\delta \{c \cdot \mathcal{L}_{CW}(x + \delta, y; \mathcal{F}) + \|\delta\|_2^2\}, \tag{5}$$

where c is a constant found by the binary search and an Adam optimizer can be used to effectively solve this optimization problem.

Boundary Attack. Boundary Attack [1] starts with the image that is already adversarial, which can be achieved by sampling each pixel of the initial image from a uniform distribution. Furthermore, a random walk is performed to keep the adversarial image in the adversarial region and decrease the distance towards the clean image simultaneously. In this way, the minimal ℓ_2 perturbation is found.

3.2 Frequency-Based Analysis Methods

Distribution of Adversarial Perturbations. Inspired by [23], we visualize the distribution of successful adversarial perturbations in the frequency domain to understand adversarial examples by adopting

$$v_{dis}^{sec} = \sum \frac{\left| f(x_{sec}^{adv} - x) \right|}{\| x_{sec}^{adv} - x \|} \Big/ num_{sec}, \tag{6}$$

where num_{sec} is the number of adversarial examples than can mislead the target model successfully. For the 3-channel input image, DFT and the norm calculation will be performed in each channel separately.

Normalization is used to visualize the distribution of adversarial perturbations produced by different attack methods under various constraints. Since different attack algorithms result in different successful adversarial examples on the same dataset for the target model, the norms and numbers of perturbations are considered in Eq. (6) to avoid the biases among the obtained distribution brought by the differences in quantities and contents of different successful adversarial perturbations.

Effectiveness of Different Frequency Bands of Adversarial Perturbations. To explore how different frequency bands of successful adversarial perturbations influence the prediction results of models, we adopt

$$v_l = Mask_{low}^r(f(x_{sec}^{adv} - x)), \tag{7}$$

$$v_h = Mask_{high}^r(f(x_{sec}^{adv} - x)), \tag{8}$$

to preserve low-frequency components v_l and high-frequency components v_h of the transformed perturbation respectively.

To be specific, $Mask_{low}^r$ and $Mask_{high}^r$ can be seen as the low-pass filter and high-pass filter respectively to preserve the corresponding parts of the transformed perturbations. Let (c_n, c_m) denote the centroid, and $Mask^r$ operation is formally defined as:

$$\begin{aligned} Mask_{low}^r(v(i,j)) &= \begin{cases} v(i,j), & \text{if} d((i,j),(c_n,c_m)) \le r \\ 0, & \text{otherwise} \end{cases}, \\ Mask_{high}^r(v(i,j)) &= \begin{cases} 0, & \text{if} d((i,j),(c_n,c_m)) \le r \\ v(i,j), & \text{otherwise} \end{cases}, \end{aligned} \tag{9}$$

where $d(\cdot)$ quantifies the distance between two positions, which is set as Euclidean distance, and r is the predefined radius. Furthermore, successful adversarial examples are reconstructed by performing IDFT on certain frequency bands of perturbations and adding them to original images. Inputting those reconstructed images to the model can help with figuring out which frequency band of perturbations mainly results in reducing the accuracy. The process mentioned above can be expressed as:

$$\mathcal{F}(x + f^{-1}(Mask^r(f(x_{sec}^{adv} - x)))). \tag{10}$$

It should be noted that the DFT, mask operation and IDFT of 3-channel adversarial perturbations is performed in each channel respectively, and then three 1-channel filtered perturbations are connected to be added to the original images.

Quantitative Analysis on Intensity. After verifying which frequency band of perturbations makes main contributions to the false predictions of the network, i.e. low-frequency band or high-frequency band, we attempt to figure out the differences between the distribution of successful and unsuccessful adversarial perturbations generated by various attacks within their concrete effective frequency bands on three datasets.

At first, we illustrate the concept of the effective frequency band. As mentioned before, the clean images are added to the low- or high-frequency parts of perturbations preserved by the predefined radius r to form the reconstructed images. The area specified by the pair of radii (r_h, r_l) that leads to the accuracy of the network against the reconstructed images drop rapidly from a high value acc_{r_h} (e.g., 90%) to a relatively low value acc_{r_l} (e.g., 50%) is referred as the effective frequency band. It should be noted that when the low-frequency band of information has the main effect on reducing the accuracy, $r_l > r_h$, while $r_l < r_h$ when high-frequency components mainly lead to the decrease of the accuracy.

The distribution of successful perturbations v_{dis}^{sec} and the distribution of unsuccessful ones v_{dis}^{unsec} are obtained using the same way described in Eq. (6) respectively. Then the values reside within the effective frequency band (r_h, r_l) are preserved discretely in k areas. When the low-frequency components play a leading role in reducing the accuracy, the preservation can be expressed as:

$$\begin{aligned} v_{dis}^{sec,k} &= Mask_{low}^{r_l-k+1}(v_{dis}^{sec}) - Mask_{low}^{r_l-k}(v_{dis}^{sec}), \\ v_{dis}^{unsec,k} &= Mask_{low}^{r_l-k+1}(v_{dis}^{unsec}) - Mask_{low}^{r_l-k}(v_{dis}^{unsec}), \end{aligned} \tag{11}$$

where $k = 1, 2, \cdots, r_l - r_h$. If high-frequency components of perturbations mainly result in the false predictions, the preservation can be expressed as:

$$\begin{aligned} v_{dis}^{sec,k} &= Mask_{high}^{r_l+k-1}(v_{dis}^{sec}) - Mask_{high}^{r_l+k}(v_{dis}^{sec}), \\ v_{dis}^{unsec,k} &= Mask_{high}^{r_l+k-1}(v_{dis}^{unsec}) - Mask_{high}^{r_l+k}(v_{dis}^{unsec}), \end{aligned} \tag{12}$$

where $k = 1, 2, \cdots, r_h - r_l$.

We propose an evaluation method to compare the intensities of the distribution of both successful and unsuccessful adversarial perturbations within the effective frequency band. We first calculate the proportion of the pixels of the successful perturbation distribution that have higher values in every discrete area, and then allocate coefficients to these items according to the resulting decrease of the accuracy. The evaluation method is formally defined as:

$$score = \sum_{k=1}^{|r_h - r_l|} \frac{p_k}{q_k} \times \frac{\Delta acc_k}{acc_{r_h} - acc_{r_l}}, \tag{13}$$

where p_k is the number of positions where $v_{dis}^{sec,k}(i,j) > v_{dis}^{unsec,k}(i,j)$ in the k-th area, q_k is the number of pixels in the k-th area, and Δacc_k is the decrease of accuracy brought by the k-th area of perturbations. A higher score indicates successful perturbations within effective frequency band have higher intensities.

4 Results and Analyses

4.1 Experimental Setup

Datasets. The MNIST database [9] contains a training set of 60000 examples and a test set of 10000 examples, which are all 28×28 grey-scale images with handwritten digits of numbers 0–9. There are 50000 training images and 10000 test images on CIFAR-10 [7], which are all 32×32 RGB images in 10 classes. ILSVRC2012 [15] is a large dataset which chooses RGB images in 1000 classes from ImageNet [3] dataset. A small subset of the validation set on ILSVRC2012 will be used in this paper, which is briefly referred to as ImageNet, and each image is resized to $299 \times 299 \times 3$ to be input to the network.

Models. For the MNIST classification task, we use two convolutional layers followed by a fully connected hidden layer. Each convolutional layer is followed by a 2×2 max-pooling layer. WideResNet [24] is adopted for the CIFAR-10 classification task. The architectures, selected hyper-parameters and training approaches of both models are identical to [10]. We achieve 99.22% accuracy on MNIST, and 93.81% accuracy on CIFAR-10. For the complex dataset ImageNet, we use the pretrained Inception-v3 network [18] provided by Keras and it achieves 77.9 % top-1 accuracy and 93.7% top-5 accuracy.

Attacks. Six mainstream attack methods, which can be divided into two categories, are applied in this paper. The first one attempts to increase the loss to mislead the model, *i.e.* FGSM, BIM, and PGD attacks. Perturbations generated by this strategy are constrained in ℓ_∞-norm. The second one pursues minimal perturbations, *i.e.* DeepFool, CW and Boundary attacks, which are ℓ_2-norm and are calculated by using Foolbox [14]. Pixel values of images on MNIST and CIFAR-10 are resized to $[0, 1]$, while the ones on ImageNet are resized to $[-1, 1]$ according to the request of using the pre-trained model. Specific constraints chosen for different attack methods and the numbers of used images on each dataset are shown in Table 1.

Table 1. Constraints and numbers of used images of attacks on each dataset.

Dataset	Attack method	Constraint	Number of images
MNIST	FGSM, BIM, PGD	0.1, 0.2, 0.3	10000
	DeepFool, CW, Boundary	2	10000
CIFAR-10	FGSM, BIM, PGD	2/255, 4/255, 8/255	10000
	DeepFool, CW, Boundary	0.5	1000
ImageNet	FGSM, BIM, PGD	2/255, 4/255, 8/255	1000
	DeepFool, CW, Boundary	3	500

4.2 Analysis on MNIST

Distribution of Adversarial Perturbations. As illustrated in Fig. 1(a) and (b), we visualize the distribution of adversarial perturbations generated by various attacks under different constraints on MNIST in the frequency domain, where the red represents higher intensity, while the blue means lower intensity. It can be seen that adversarial perturbations generated by FGSM, BIM, and PGD attacks all concentrate on the low-frequency domain. With the increase of ℓ_∞-constraints, BIM, and PGD attacks concentrate more on a low-frequency domain, which is implied by the extension of the deep red area in the centers of the frequency spectra in Fig. 1(a). Nevertheless, the distribution of perturbations generated by the FGSM attack changes very little, which may be because FGSM is a one-step attack and the positions of attacked pixels hardly change.

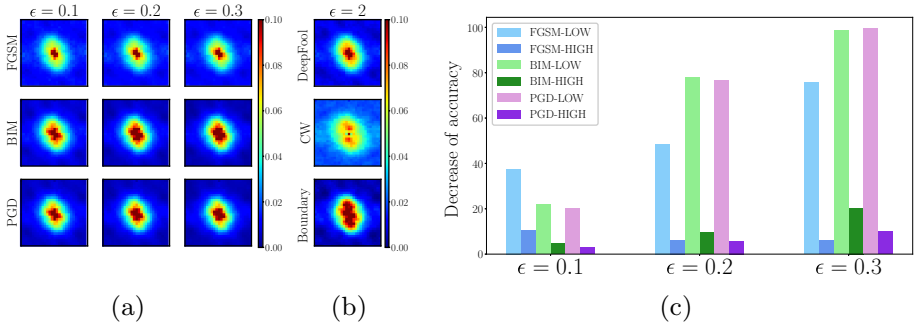

Fig. 1. The distribution of successful adversarial perturbations generated by FGSM, BIM, and PGD attacks on MNIST is depicted in the leftmost image and the distribution of the ones generated by DeepFool, CW and Boundary attacks is shown in the middle. The rightmost image shows the decrease of accuracy brought by two frequency bands of perturbations produced by FGSM, BIM, and PGD attacks on MNIST.

When it comes to DeepFool, CW, and Boundary attacks, we do not change the constraints because they belong to the strategy that finds the minimal perturbations. The ℓ_2-constraint is set to guarantee the imperceptibility. It is shown

in Fig. 1(b) that all the perturbations are low-frequency as well. Compared with the other two attacks, the CW attack generates relatively more high-frequency perturbations, because it tends to find the perturbations that modify the outlines of digits instead of the images' backgrounds.

Decrease of Accuracy. Low frequency and high frequency are two relative concepts, and it is hard to specify a constant radius to separate images on each dataset. As a result, we change the radius from zero to the maximum value to separate the low- and high-frequency components of perturbations, and add the filtered perturbations to the original images to obtain the decrease of accuracy brought by those perturbations. Then we choose a proper radius according to the size of the image to show experimental results. On MNIST, results obtained when $r = 8$ are depicted in Fig. 1(c). It is shown that low-frequency components mainly contribute to the decrease of accuracy while high-frequency components have a subtle influence on launching successful attacks on MNIST, which is consistent with distribution of adversarial perturbations shown in Fig. 1. The degree of decrease of accuracy is proportional to the intensity of the attack.

The decrease of accuracy results from low- and high-frequency components of DeepFool perturbations are 37.31% and 7.77% respectively. We can know that low-frequency components still play a leading role in reducing the accuracy. An exciting phenomenon emerges when we separate perturbations produced by CW and Boundary attacks: neither low-frequency components nor high-frequency components reduce the accuracy. In other words, adversarial perturbations of both attacks are out of operation after the separation in the frequency domain. This phenomenon also exists on the rest two datasets. It may be attributed to that both attacks do not build the direct connections with the outputs of the network w.r.t. to the inputs, and thus do not exploit the frequency information learned by the network during training. Therefore, the obtained perturbations cannot reflect the network's sensitivity to the specific frequency bands of perturbations.

Intensity Analysis. The effective frequency bands found for various attacks under different constraints and calculated scores are shown in Table 2. There is a lack of scores of BIM and PGD attacks under 0.3 and $8/255$ l_∞-norm constraints, because fooling rates of both attacks under that conditions can be 100%.

On MNIST, acc_{r_h} is set to be 95%, and acc_{r_l} is set to be 65%. Table 2 illustrates that the pixels of successful perturbation distribution have higher intensities within effective frequency bands, which can be seen as a reason that those perturbations are equipped with the ability to mislead the model. While for PGD attack constrained in 0.1 l_∞-norm constraint, the score is less than 0.5, and it may be attributed to the area divided by the r_l does not strictly belong to low-frequency area. Unsuccessful perturbations can attack the area other than the low-frequency area harder, i.e., the image's background.

Table 2. The effective frequency bands and results of intensity comparison on three datasets.

Dataset	MNIST				CIFAR-10				ImageNet			
Attack	Constraint	r_h	r_l	Score	Constraint	r_h	r_l	Score	Constraint	r_h	r_l	Score
FGSM	0.1	3	7	0.69	2/255	19	15	0.72	2/255	13	31	0.48
	0.2	4	7	0.72	4/255	19	15	0.72	4/255	11	25	0.61
	0.3	3	6	0.54	8/255	20	16	0.80	8/255	9	21	0.60
BIM	0.1	5	10	0.50	2/255	18	15	0.70	2/255	18	40	0.41
	0.2	4	6	0.77	4/255	19	16	0.56	4/255	17	35	0.48
PGD	0.1	5	10	0.36	2/255	18	15	0.65	2/255	19	41	0.45
	0.2	4	6	0.52	4/255	19	16	0.53	4/255	19	38	0.49
DeepFool	2	4	7	0.54	0.5	20	16	0.89	3	48	105	0.57

4.3 Analysis on CIFAR-10

Distribution of Adversarial Perturbations. The properties of the distribution of successful adversarial perturbations on CIFAR-10 are pretty different from those on MNIST. As shown in Fig. 2(a), adversarial perturbations generated by FGSM, BIM, and PGD attacks mainly concentrate on high-frequency domains. Compared with the distribution of perturbations produced by FGSM, which still hardly changes when the predefined constraints enlarge, target features attacked by BIM and PGD attacks gradually concentrate on the low-frequency domain, leading to both center and margin of the frequency spectra being high-value. This may be attributed to that both attacks find perturbations iteratively, leading to the reduction of differences between adjacent pixels, and the generated perturbations contain more low-frequency information. Besides, the intensity of the central area of perturbation distribution produced by the

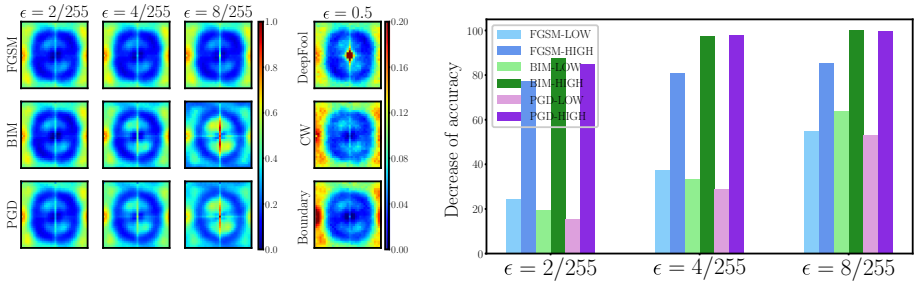

Fig. 2. The distribution of successful adversarial perturbations generated by FGSM, BIM, and PGD attacks on CIFAR-10 is depicted in the leftmost image and the distribution of the ones generated by DeepFool, CW and Boundary attacks is shown in the middle. The rightmost image shows the decrease of accuracy brought by two frequency bands of perturbations produced by FGSM, BIM, and PGD attacks on CIFAR-10.

PGD is lower than the one produced by the BIM, which results from the noises introduced by random initialization at the beginning of the PGD attack.

Adversarial perturbations generated by DeepFool, CW, and Boundary attacks also exhibit high-frequency characteristics, which are shown in Fig. 2(b). Compared with CW and Boundary attacks which mainly focus on attacking high-frequency components, the DeepFool attack generates perturbations on extremely low-frequency domains. The ℓ_2-norm of distortions caused by DeepFool attacks are at least 3 times bigger than that caused by CW and Boundary, which leaves that its perturbations contain more low-frequency information.

Decrease of Accuracy. Here we show the results when $r = 10$ is used to filter the low- and high-frequency components of adversarial perturbations on CIFAR-10. It can be seen from Fig. 2(c) that high-frequency parts of adversarial perturbations have a superior effect on misleading the model. With the increase of ℓ_∞-constraints of FGSM, BIM, and PGD attacks, the gap of effectiveness on deceiving the model between low- and high-frequency components generated by each attack is narrowed. When $\epsilon = 8/255$, both frequency bands of perturbations of each attack can reduce the accuracy by more that 50%.

Low- and high-frequency components of perturbations generated by the DeepFool attack reduce the accuracy by 18.47% and 80.33%, respectively. While it can be seen from Fig. 2(b) that the central spectrum of perturbations generated by DeepFool has an extremely high value, the effectiveness of the most low-frequency components is minimal, which may because models are hardly sensitive to the additive perturbations in lowest frequencies [23]. The experimental results also imply this phenomenon that the accuracy remains above 99% until radius $r > 3$ in most cases when the model is attacked by low-frequency parts of adversarial perturbations added to the original images. The decrease of accuracy brought by high-frequency components of perturbations of CW attack is 5.85%, and the low-frequency components of which are out of operation. As for Boundary attacks, the low- and high-frequency parts of perturbation produced by these attacks are not effective anymore after separation, which is the same as the experimental result on MNIST. We also adopt the VGG16 model [17] to conduct the same experiments to verify that whether the properties of adversarial perturbations in the frequency domain are independent of model architectures. Obtained results exhibit similar characteristics in both aspects and confirm the generalization of illustrated characteristics furthermore.

Intensity Analysis. On CIFAR-10, acc_{r_h} and acc_{r_l} are set to be 90% and 50% respectively. Evaluation results in Table 2 show that unsuccessful perturbation distribution exhibits extremely lower intensity compared with the successful perturbation distribution in some cases. Successful perturbation distribution gets a score of over 0.5 in every situation, which is assumed by us to be an explanation for the failure of unsuccessful perturbations on cheating the model.

4.4 Analysis on ImageNet

Distribution of Adversarial Perturbations. Figure 3(a) shows that the distribution of successful adversarial perturbations generated by FGSM, BIM, and PGD attacks mainly concentrates on low-frequency domains, similar to that on MNIST. However, compared with the BIM attack that constantly attacks low-frequency components, the PGD attack gradually increases the intensity of the attack on high-frequency components. In our experiments, besides more iteration steps, PGD attack introduces the random initialization at the beginning of the attack. To figure out which one is the main factor that leads to the difference mentioned above, we force both attacks to have the same numbers of iterations under changeable constraints and obtain similar results, which implies that the random initialization can be the reason for changing the distribution of adversarial perturbations generated by the PGD attack.

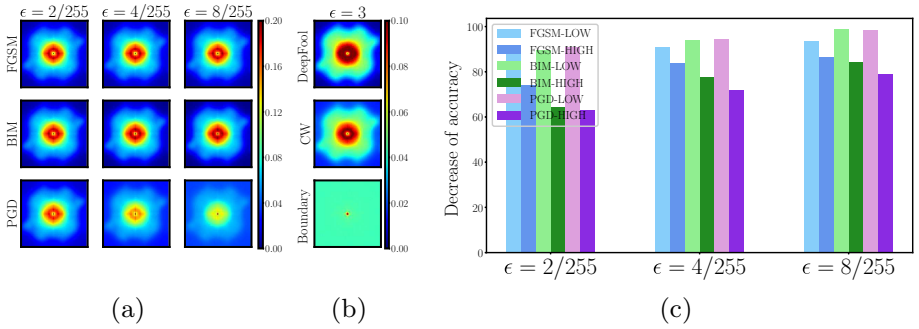

Fig. 3. The distribution of successful adversarial perturbations generated by FGSM, BIM, and PGD attacks on ImageNet is depicted in the leftmost image and the distribution of the ones generated by DeepFool, CW and Boundary attacks is shown in the middle. The rightmost image shows the decrease of accuracy brought by two frequency bands of perturbations produced by FGSM, BIM, and PGD attacks on ImageNet.

It is shown in Fig. 3(b) that DeepFool and CW attacks exhibit the property of taking the low-frequency components as the target features. In contrast, the adversarial perturbations produced by Boundary attack are uniformly distributed across the frequencies. After comparing the attacked and clean images, we find that the Boundary attack modifies attacked pixels in each image to a similar extent. However, the CW and DeepFool attacks mainly target at attacking concrete objects of each image.

Decrease of Accuracy. Because the image's resolution is large, we do not change the radius continuously but choose several radii values to record the decrease of accuracy instead. The experimental results obtained when $r = 80$ are depicted in Fig. 3(c). Low-frequency components are the main factors that

reduce the accuracy of the model, while high-frequency components have an obvious effect on misleading the model as well, which may be attributed to that the accuracy of model on original images is not high enough, leaving the model vulnerable to both frequencies of perturbations. The results explain the fact that both of the works [4,6] promote the performance of attacks while finding the adversarial examples in different frequency bands. With the increase of constraints, the gap between the effectiveness of low- and high-frequency components is narrowed, which is similar to the trend exhibiting on CIFAR-10.

Low-frequency components of perturbations produced by DeepFool reduce the accuracy of the model by 36.50%. However, the high-frequency components of perturbations result in a 13.87% decrease of accuracy, which is consistent with the distribution of perturbations shown in Fig. 3(b). As to CW and Boundary attacks, the perturbations are still out of operation after the separation.

Intensity Analysis. On ImageNet, acc_{r_h} and acc_{r_l} are set to be 90% and 50%, respectively. Unlike the experimental results on the other two datasets that the intensities of successful perturbation distribution are visibly greater, a large proportion of obtained scores are around 0.5. From Fig. 2(c) we can understand that low-frequency components mainly cause the decrease of accuracy, but the advantages of which are not that obvious. As a result, both successful and unsuccessful perturbation distribution can have competitive intensities within the effective frequency bands, which are restricted in the low-frequency bands on ImageNet.

4.5 Discussion

For the successful adversarial perturbations produced by the attack strategy that increases the loss to mislead the model, with the improvement of the constraints of attacks, the distribution of perturbations generated by FGSM has no apparent changes. Adversarial perturbations generated by BIM and PGD attacks gradually concentrate on the low-frequency domain during this procedure, while the random initialization introduced by PGD attack may change this trend as the proportion of the background in the image enlarges. For the successful adversarial perturbations generated by the attack strategy that pursues the minimal perturbations, because of the vast differences among the principles of attacks, they do not have unified laws on three datasets while maintaining the same concentration area with the ones generated by the first attack strategy.

We find that the high-frequency components of successful adversarial perturbations have a minimal effect on MNIST. While on CIFAR-10, high-frequency components play a leading role in fooling models. Both low- and high-frequency components of successful perturbations on ImageNet can obviously affect models, and low-frequency ones have superior performance. Although the characteristics of successful adversarial perturbations vary from dataset to dataset, the evaluation results illustrate that they have larger attack intensities within different effective frequency bands compared with the unsuccessful ones in most situations, which can be seen as the factor that misleads the model successfully.

Since the distribution of successful adversarial perturbations varies from dataset to dataset, we assume that it can be associated with how much space the object occupies in the images. Attacking the objects in the images makes pixel values of corresponding parts of perturbations change dramatically due to the complex outlines and texture of objects, which means high-frequency information will remain in the perturbations. The backgrounds in the images are relatively smooth, and attacks tend to change backgrounds to a small extent to remain imperceptible, which leaves low-frequency information.

Figure 4 illustrates three pairs of clean images and successfully attacked images and visualizes the DFT of their perturbations. It is shown that, the number is placed in the center of the image on MNIST and takes up less than a quarter of space of the image. Consequently, there is more low-frequency information in the perturbation. On CIFAR-10, the object takes up almost all the space in the image, resulting in much high-frequency information in the perturbation. While on ImageNet, most of successfully attacked images have more information of backgrounds rather than objects, thus leaving perturbations concentrating on the low-frequency domain. Such visualized analyses also validate our hypothesis.

Fig. 4. The DFT of perturbations on MNIST, CIFAR-10 and ImageNet. In each 3 × 1-image part, the first one is the clean image, the second one is the adversarial image, and the last one is the DFT of corresponding adversarial perturbation.

Consequently, taking the properties of datasets or some specific usage scenarios where the sizes of objects are relatively constant may enhance the robustness of deep learning models from the perspective of the frequency domain. However, the universal defense that achieves high accuracy on both adversarial and clean images remains challenging.

5 Conclusion and Further Work

In this paper, six classic attack methods and three commonly used datasets are adopted to comprehensively analyze and evaluate the adversarial examples in the frequency domain. We explore the effectiveness of different frequency bands of perturbations through a quantitative analysis. Our significant findings successfully explain the contradictory parts of previous works. Evaluation results show that compared with the distribution of unsuccessful adversarial perturbations, the distribution of successful ones exhibits higher intensity within the effective frequency bands, providing an explanation for launching attacks successfully.

Besides addressing the limitations of existing theories, we obtain a better understanding of adversarial examples from the frequency domain perspective and provide an idea on enhancing the robustness of models by considering the frequency properties of datasets in advance. Further work is required to conduct analyses on adversarially trained models from the frequency domain perspective and build efficient and effective defense by exploiting the frequency properties.

Acknowledgement. This research is supported by the National Key Research and Development Program of China (2020AAA0107702), the National Natural Science Foundation of China (62006181, 61822309, 61703301, 61822309, 61773310, U20A20177, U1736205) and the Shaanxi Province Key Industry Innovation Program (2021ZD LGY01-02).

References

1. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248 (2017)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. arXiv preprint arXiv:1608.04644 (2016)
3. Deng, J., Dong, W., Socher, R., et al.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on CVPR, pp. 248–255. IEEE (2009). https://doi.org/10.1109/CVPR.2009.5206848
4. Deng, Y., Karam, L.J.: Frequency-tuned universal adversarial attacks. arXiv preprint arXiv:2003.05549 (2020)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
6. Guo, C., Frank, J.S., Weinberger, K.Q.: Low frequency adversarial perturbation. arXiv preprint arXiv:1809.08758 (2018)
7. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
8. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
9. LeCun, Y.: The MNIST database of handwritten digits (1998)
10. Madry, A., Makelov, A., Schmidt, L., et al.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
11. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., et al.: Analysis of universal adversarial perturbations. arXiv preprint arXiv:1705.09554 (2019)
12. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., et al.: Universal adversarial perturbations. arXiv preprint arXiv:1610.08401 (2016)
13. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. arXiv preprint arXiv:1511.04599 (2015)
14. Rauber, J., Zimmermann, R., Bethge, M., et al.: Foolbox native: fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. J. Open Source Softw. **5**(53), 2607 (2020). https://doi.org/10.21105/joss.02607 https://doi.org/10.21105/joss.02607
15. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet large scale visual recognition challenge. arXiv preprint arXiv:1409.0575 (2014)

16. Sharma, Y., Ding, G.W., Brubaker, M.: On the effectiveness of low frequency perturbations. arXiv preprint arXiv:1903.00073 (2019)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: Rethinking the inception architecture for computer vision. arXiv preprint arXiv:1512.00567 (2015)
19. Szegedy, C., Zaremba, W., Sutskever, I., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
20. Tanay, T., Griffin, L.: A boundary tilting persepective on the phenomenon of adversarial examples. arXiv preprint arXiv:1608.07690 (2016)
21. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High frequency component helps explain the generalization of convolutional neural networks. arXiv preprint arXiv:1905.13545 (2019)
22. Wang, Z., Yang, Y., Shrivastava, A., et al.: Towards frequency-based explanation for robust CNN. arXiv preprint arXiv:2005.03141 (2020)
23. Yin, D., Lopes, R.G., Shlens, J., et al.: A Fourier perspective on model robustness in computer vision. arXiv preprint arXiv:1906.08988 (2019)
24. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
25. Zhang, H., Yu, Y., Jiao, J., et al.: Theoretically principled trade-off between robustness and accuracy. arXiv preprint arXiv:1901.08573 (2019)