A Fourier Perspective of Feature Extraction and Adversarial Robustness

Liangqi Zhang$^1$, Yihao Luo$^2$, Haibo Shen$^1$ and Tianjiang Wang$^1$*

$^1$Huazhong University of Science and Technology
$^2$Yichang Testing Technique R&D Institute

{zhangliangqi, luoyihao, shenhaibo, tjwang}@hust.edu.cn

Abstract

Adversarial robustness and interpretability are longstanding challenges of computer vision. Deep neural networks are vulnerable to adversarial perturbations that are incomprehensible and imperceptible to humans. However, the opaqueness of networks prevents one from theoretically addressing adversarial robustness. As a human-comprehensible approach, the frequency perspective has been adopted in recent works to investigate the properties of neural networks and adversarial examples. In this paper, we investigate the frequency properties of feature extraction and analyze the stability of different frequency features when attacking different frequencies. Therefore, we propose an attack method, $\mathcal{F}$-PGD, based on the projected gradient descent to attack the specified frequency bands. Utilizing this method, we find many intriguing properties of neural networks and adversarial perturbations. We experimentally show that contrary to the low-frequency bias of neural networks, the effective features of the same class are distributed across all frequency bands. Meanwhile, the high-frequency features often dominate when the neural networks make conflicting decisions on different frequency features. Furthermore, the attack experiments show that the low-frequency features are more robust to the attacks on different frequencies, but the interference to the high frequencies makes the network unable to make the right decision. These properties indicate that the decision-making process of neural networks tends to use as few low-frequency features as possible and cannot integrate features of different frequencies.

1 Introduction

Although deep neural networks (DNNs) have shown considerable promises in various visual applications, these neural networks are actually quite brittle. In particular, the models trained using standard methods are vulnerable to adversarial perturbations [Szegedy et al., 2014; Goodfellow et al., 2015]. However, the lack of interpretability of neural networks leads to the inability to address these drawbacks theoretically, which makes us eager to understand their feature extraction and decision-making processes. Although many specialized methods have been proposed to find the decision basis of neural networks in recent years, such as feature visualization [Erhan et al., 2009; Yosinski et al., 2015] and attribution methods [Zeiler and Fergus, 2014; Simonyan et al., 2014; Springenberg et al., 2015; Smilkov et al., 2017], there is still a huge gap in understanding the feature extraction and decision process.

Besides, data augmentation [Cubuk et al., 2019; Zhong et al., 2020] can improve the robustness of neural networks in many aspects, but these methods are not defensive against ad-
versarial attack methods. In practice, one can often generate imperceptible perturbations of the input images and cause the model to make highly-confident but erroneous predictions. The vulnerability of models trained using standard methods to adversarial perturbations makes it clear that the paradigm of adversarially robust learning differs from the classic learning setting. Consequently, various adversarial defense methods have been proposed, among which adversarial training has proven to be the most effective means of adversarial defense. However, adversarial training consumes considerable computational resources and cannot fully address adversarial robustness.

Many works have interpreted the adversarial examples from aspects such as networks themselves or datasets to understand them deeply. Goodfellow et al. [2015] argue that adversarial examples are due to the high-dimensional linear nature of neural networks. Ilyas et al. [2019] thought that adversarial examples are non-robust features, the models can still understand them deeply. Goodfellow from aspects such as networks themselves or datasets to unrobustness. However, adversarial training consumes considerable computational resources and cannot fully address adversarial robustness.

2 Related Work

2.1 Robustness

Robustness is a long-standing and challenging goal of computer vision. Although data augmentation can improve the robustness of neural networks in many aspects, it cannot defend against adversarial examples. Recent works have proven that adversarial training is an effective method of defending against adversarial attacks.

Adversarial Robustness. Szegedy et al. [2014] discover neural networks are vulnerable to adversarial examples, various adversarial attack algorithms [Goodfellow et al., 2015; Carlini and Wagner, 2017; Madry et al., 2018; Guo et al., 2019b; Su et al., 2019; Croce and Hein, 2020] have been proposed to investigate the vulnerability of machine learning models. Adversarial perturbations are almost imperceptible changes to the input that cause neural networks to make erroneous predictions. Goodfellow et al. [2015] proposed the Fast Gradient Sign Method (FGSM) to generate perturbations with a single gradient step. It is an attack for an \( l_\infty \)-bounded adversary and computes an adversarial example as

\[
x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y; \theta)),
\]

where \( \text{sign} \) operation makes perturbations meet the \( l_\infty \)-norm bound as soon as possible.

The more powerful adversary is the multi-step variant, which is essentially Projected Gradient Descent (PGD) [Carlini and Wagner, 2017] on the negative loss function

\[
x^{t+1} = \text{clip}_{x^t, \epsilon}(x^t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(x, y; \theta))),
\]

where

\[
\text{clip}_{x, \epsilon}(\cdot) = \min(\max(\cdot, x - \epsilon), x + \epsilon).
\]

PGD is a very stable and widely used adversarial attack method, and our subsequent analyses are based on PGD-style attacks.

Adversarial Training. The adversarial training method has proven to be an effective method of defending against adversarial attacks. It was first proposed by [Goodfellow et al., 2015], which is the most successful approach for building robust models so far for defending adversarial examples. Adversarial training can be formulated as solving a robust optimization problem [Shaham et al., 2015]

\[
\min_{\theta} \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \max_{\delta} \mathcal{L}(x + \delta, y; \theta) \right],
\]

where \( \mathcal{L}(\cdot, \cdot; \cdot) \) is the chosen loss function and \( \theta \) denotes the parameters of the neural network; the data pair \( (x, y) \) is sample from the data distribution \( \mathcal{D} \) and \( \delta \) denotes the corresponding adversarial perturbation. The inner maximization is approximated by adversarial examples generated by various adversarial attack methods.
2.2 Frequency Perspective

According to the convolution theorem, convolutional neural networks (CNNs) have the natural ability to separate different frequency information. And various experiments [Geirhos et al., 2019; Brendel and Bethge, 2019] have demonstrated that neural networks have different sensitivities to various frequency components of the input image, standard CNNs make their predictions rely on the local textures rather than long-range dependencies encoded in the shape of objects. Some works [Rahaman et al., 2019; Xu et al., 2019] find empirical evidence of a spectral bias that lower frequencies are learned first and then the higher frequencies are captured slowly. Zhang et al. [2023] further argue that frequency bias is also data-dependent. Different classes or samples have different frequency biases, and the bias is related to the scale of the image classification target.

The frequency perspective has also been employed in the analysis of adversarial examples. For a long time, it was thought that adversarial perturbations were mostly high-frequency perturbations, but recent work [Maiya et al., 2021] has demonstrated that adversarial perturbations are neither in high-frequency nor in low-frequency components but are dataset dependent. Other experiments [Guo et al., 2019a; Sharma et al., 2019; Tancik et al., 2020; Long et al., 2022] have found that training methods that improve the robustness of the network, such as data augmentation or adversarial training, make the neural networks prefer lower frequencies. And this also leads to the opinion that the robustness of low-frequency features is higher than that of high-frequency features.

3 Preliminaries

Let us consider a standard classification task with an underlying data distribution \( D \) over pairs of examples \( x \in X \) and corresponding labels \( y \in Y \). We also assume that we are given a suitable loss function \( \mathcal{L}(x, y; \theta) \), for instance the cross-entropy loss for a neural network. As usual, \( \theta \) is the set of model parameters. The goal is to find model parameters \( \theta \) that minimize the risk

\[
\min_{\theta} \mathbb{E}_{(x,y) \sim D} [\mathcal{L}(x, y; \theta)]. \tag{5}
\]

3.1 Statistics and Features

Before moving on to the Fourier perspective, let us consider what information a neural network needs to extract. According to the Information Bottleneck (IB) [Tishby et al., 2000] neural networks extract relevant information that an input random variable \( X \) contains about an output random variable \( Y \), the relevant information is defined as the mutual information \( I(X; Y) \). In statistical terms, the relevant information of \( X \) with respect to \( Y \), denoted by \( T \), is a minimal sufficient statistics of \( X \) with respect \( Y \). In this case, \( Y \) implicitly determines the relevant and irrelevant features in \( Y \). Since exact minimal sufficient statistics only exist for special distributions, the information bottleneck method [Tishby et al., 2000] relaxed this optimization problem by allowing the map to be stochastic, defined as an encoder \( P(T|X) \), and capture as much as possible of \( I(X, Y) \), not necessarily all of it. Due to the limited capacity of neural networks, \( P(T|X) \) will always compress the information, the Equation 5 will be implicitly equivalent to the information bottleneck problem. And it is formulated by the following optimization problem with the Markov chain: \( Y \rightarrow X \rightarrow T \):

\[
\min \{ I(X; T) - \beta I(T; Y) \} \tag{6}
\]

where the hyperparameter \( \beta \) controls the information loss ratio.

In practice, due to the limited model capability and dataset, it can only guarantee to extract the statistic \( T \) related to the output \( Y \) from the input data, but not to the classified objects \( O^Y \) in the images. In other words, there are no constraints to make the networks learn to distinguish between various statistics and human-comprehensible robust features (or concepts) and tends to learn to use less information for classification purposes. This is consistent with subsequent experiments showing that the networks tend to use a small amount of low-frequency information to make decisions.

3.2 Fourier-Based PGD Attack

To evaluate the effect of the adversarial attack on different frequencies, we use Fast Fourier Transform (FFT) to constrain the frequencies of adversarial perturbations. Let \( F \) and \( F^{-1} \) represent the forward Fast Fourier Transform and its corresponding inverse.

Our algorithm is based on the PGD and removes specific frequency components of the output perturbation \( \delta^t \) by applying a mask to its frequency spectrum FFT(\( \delta^t \)), and reconstruct the gradient by applying the IFFT on the masked spectrum. Specifically, the mask, \( M \in \{0, 1\}^{d \times d} \), is a two-dimensional matrix, and the mask operation is done by element-wise product \( \odot \). In our work, we consider the \( l_\infty \)-norm and the algorithm that attack \( f \in \mathcal{F} \) frequencies performs \( T \)-step attack with a small step size \( \alpha = \epsilon/T \):

\[
\delta^t = \nabla_{x_t'} \mathcal{L}(x_t', y; \theta) \tag{7}
\]

\[
\delta^t_f = F^{-1}(F(\delta^t) \odot M) \tag{8}
\]

\[
x^{t+1} = \text{clip}_{x_t, \epsilon} (x^t + \alpha \cdot \text{sign}(\delta_f^t)). \tag{9}
\]

Note that the non-linear sign and clip operators alias some passed information into other frequencies, and so the perturbations are not strictly contained in the frequency band.

---

Figure 2: Perturbations and their spectrums of Fourier-based PGD attack. \( \mathcal{F} \)-PGD always passes information into other frequencies, but \( \mathcal{F} \)-PGD-L will not.
4 Frequency Properties of Feature Extraction

To analyze the frequency properties of the feature extraction, we need to evaluate the performance of the networks in different frequency bands. Furthermore, we analyze the frequency properties of the networks over the entire dataset as well as for each class. Therefore, we transform the images to the frequency domain and then use ideal low-pass filters (LPF), high-pass filters (HPF), and band-pass filters (BPF) to remove specified frequency information, respectively. Then all the experiments are conducted on ImageNet [Deng et al., 2009].

The tested models include both CNNs and transformer-based networks, such as ResNets [He et al., 2016], MobileNetV1 [Howard et al., 2017], MobileNetV3 [Howard et al., 2019], EfficientNets [Tan and Le, 2019], ViT [Dosovitskiy et al., 2021] and Swin [Liu et al., 2021]. All the tested images are at the resolution of 224 × 224.

Since directly removing frequencies from the images would lead to inconsistent distribution between test and training data, we preprocess the data with ideal low-pass, high-pass, and band-pass filters and retrain the MobileNetV1 × 1.0 model on ImageNet from scratch, respectively.

4.1 Distribution of Features

We first test the distribution of features in different frequency bands (bandwidth of 15) using band-pass filters. As shown in Table 1 and Figure 1a, all models retrained with band-pass filters achieve more than 50% accuracy in the frequency range [0,120] (note that the information decreases rapidly when the filter radius exceeds 112). There is no essential difference in each band for the classification models, which indicates that each band is rich in effective features.

Then, we test the feature distribution of each class. As shown in Figure 4, most of the classes have similar accuracy in different frequency bands, rather than different classes achieving different accuracy in different frequency bands. It is only significantly different from the accuracy achieved in

---

Table 1: Comparison of full-frequency pre-trained and retrained models at different $r_L$. These experiments show that different frequency bands of image data have a large number of effective features. And training a network using only the low-frequency bands does not improve robustness but decreases it. Clean: The pre-trained models are tested on the full band and the retrained models are tested on the passed bands.

$$x^{t+1} = x^t + \alpha \cdot \frac{\text{card}(g_j^t)}{||g_j^t||_1} \cdot g_j^t$$

(10)
the full frequency band. We hold the opinion that these components of the same feature at different frequencies are redundant, and the neural networks do not need to extract the information from all frequency bands.

As shown in Figure 3, images with a frequency bandwidth of 15 are also indistinguishable to humans. This further demonstrates that the classification task does not require the models to extract robust features and can rely only on the i.i.d. statistics to achieve high accuracy.

4.2 Importance of Low-Frequency Features

As shown in Figure 5, with the increase of high-frequency information, the accuracy of the models increases gradually; With the absence of low-frequency information, the accuracy of the models decreases sharply, and the models cannot work at all when some low-frequency information is missing. In addition, the higher-accuracy network performs better in the low-frequency band than other networks. This indicates that the classification models rely more on low-frequency components and are vulnerable to differences in data distribution due to low-frequency loss. And the models cannot rely on high-frequency information alone to make correct classifications.

On the one hand, the retrained models are significantly higher than the standard trained models at most frequencies, as shown in Table 1. In particular, the retrained model achieve 52% top-1 accuracy at $r_{LP} = 10$, which is much higher than the 4.6% top-1 accuracy of the standard training model. And it achieve 72.1% with only 1% accuracy lower than the full-frequency trained model at $r_{LP} = 65$, which also indicates that for the standard training models, high frequencies are less important for classification prediction.

On the other hand, the gap between the accuracy of the network retrained with high-frequency information and the standard trained network is even more prominent. These results indicate that the importance of low-frequency information does not show up in the features it contains but in the low-frequency bias. The lack of low frequencies will have a huge impact on the image structure and the distribution of the data set. For the standard trained networks, the low-frequency components are the basis for further utilization of the high-frequency components.

4.3 Importance of High-Frequency Features

As shown in Table 1 and Figure 6, high-frequency features seem complementary to low-frequency features to achieve better generalization. However, the adversarial attack results
show that high-frequency features are more vulnerable to adversarial perturbations, causing networks to fail to make correct decisions, as shown in Figure 1c and 1d.

The significant impact of perturbed high-frequency features on decision-making indicates that high-frequency features are no less important than low-frequency features. We hypothesize that the neural network will preferentially select low-frequency features as the basis for its decision, and when it cannot decide, it will gradually use higher-frequency features. When the network needs to use both low-frequency and high-frequency features, high-frequency features are more important to the decision result than low-frequency features. Furthermore, our experimental results support this hypothesis.

4.4 Low-Frequency vs. High-Frequency Features

As shown in Figure 6, in addition to some classes that can be used as decisions only depending on low-frequency features, other classes require the participation of low- and high-frequency features. In the experiment, we found that the accuracy of some classes will gradually decrease with the addition of high-frequency features, especially among some classes with the same low-frequency features. We believe that the conflict between low-frequency and high-frequency features causes this, and it is the high-frequency features that ultimately play a decisive role.

Conflicts between high-frequency and low-frequency features will appear in similar classes on the ImageNet, such as between different species of spiders. As shown in Figure 7, networks using high- and low-frequency features always produce different predictions in classes between 72 and 77. These samples are always predicted to be class 73 at low frequencies and gradually classified into the correct class with the increase of high-frequency information. This indicates that these classes have the same low-frequency features and that these features belong to class 73. However, the high-frequency features will dominate the final classification results. The high-frequency feature-dominated classification results extracted by the standard trained models are also shown in the adversarial attack, which we analyze in the next section.

5 Frequency-Based Attack

To verify the effectiveness of adversarial attacks on different frequencies, we performed frequency attacks on both the standard-trained and the retrained models on different frequencies. We believe that...
in Figure 6, many classes in the networks that can be correctly classified by ultra-low-frequency information alone. However, such stable low-frequency features could not improve the adversarial robustness of networks.

We perform $F$-PGD-L attacks on low-frequency dependent samples at different frequency bands. As shown in Figure 8, low-frequency features have some robustness against attacks of various frequencies, and the networks could make correct decisions when using only low-frequency features. Moreover, the attacks on the low-frequency bands will affect the features of all frequency bands, while attacks on the high-frequency band will hardly interfere with the low-frequency features. The effect of adversarial attacks on low-frequency bands is higher than on high-frequency bands. The disturbed high-frequency features dominate the networks to make the final wrong decision. This indicates that stable low-frequency features cannot improve the robustness, and the model cannot unify low-frequency and high-frequency features.

6 Conclusion

In this paper, we analyzed the feature extraction and adversarial robustness from the Fourier perspective. We experimentally show that low- and high-frequency features are both important for decision-making. Low-frequency features are the basis for classification decisions, but high-frequency features often dominate when the neural networks make conflicting decisions on different frequency features. The attack experiments show that the low-frequency features are more robust to the attacks on different frequencies. However, since the higher frequency features dominate the decision-making and are very vulnerable to interference from various frequency attacks, the robust low-frequency features will not improve the adversarial robustness of CNNs. Next, we will explore whether reducing the sensitivity of networks to high-frequency features and improving the use of broader frequency features will improve the robustness of the network in future research.
References


