Understanding robustness and generalization of artificial neural networks through Fourier masks.

Nikos Karantzazas$^{1,3+}$, Emma Besier$^{1,3+}$, Josue Ortega Caro$^{1,3}$, Xaq Pitkow$^{1,2,3}$, Andreas S. Tolias$^{1,2,3}$, Ankit B. Patel$^{1,2,3}$ and Fabio Anselmi$^{1,3,*}$

$^1$Department of Neuroscience, Baylor College of Medicine, Houston, 77030, USA.
$^2$Department of Electrical and Computer Engineering, Rice University, Houston, 77005, USA.
$^3$Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, 77030, USA.

† co-first authors
∗ corresponding authors

Abstract

Despite the enormous success of artificial neural networks (ANNs) in many disciplines, the characterization of their computations and the origin of key properties such as generalization and robustness remain open questions. Recent literature suggests that robust networks with good generalization properties tend to be biased towards processing low frequencies in images. To explore the frequency bias hypothesis further, we develop an algorithm that allows us to learn modulatory masks highlighting the essential input frequencies needed for preserving a trained network’s performance. We achieve this by imposing invariance in the loss with respect to such modulations in the input frequencies. We first use our method to test the low-frequency preference hypothesis of adversarially trained or data-augmented networks. Our results suggest that adversarially robust networks indeed exhibit a low-frequency bias but we find this bias is also dependent on directions in frequency space. However, this is not necessarily true for other types of data augmentation. Our results also indicate that the essential frequencies in question are effectively the ones used to achieve generalization in the first place. Surprisingly, images seen through these modulatory masks are not recognizable and resemble texture-like patterns.

Keywords: Neural Networks, Fourier analysis, Symmetry, Robustness, Generalization, Data augmentation.

1 Introduction

Artificial neural networks (ANNs) have achieved impressive performance in a variety of tasks, e.g., object recognition, function approximation, natural language processing, etc. [1]. However, their computational capacity remains rather opaque. In particular, the operations performed by ANNs are profoundly constrained by the choice of architecture, initialization, optimization techniques, etc., and such constraints have a significant impact on key properties such
as generalization power and robustness. Studying adversarial robustness has been a very active area of research, since it is closely related to how trustworthy and reliable neural networks can be [2]. One of the most explored directions has been the analysis of adversarial perturbations from a frequency standpoint. For example, the work of [3] establishes a relationship between the frequency domain of different noises (e.g., Adversarial examples and Common corruptions) and model performance. In particular, they show that deep neural networks are more sensitive to high frequency adversarial attacks or common corruptions such as random noise, contrast change, and blurring. Additionally, adversarial perturbations of commonly trained models tend to be higher frequency than their adversarially trained counterparts. Furthermore, [4] found that high frequency features are necessary for good generalization performance while the work of [5] shows that performance improvements in white-box and black-box transfer settings can be achieved only when low frequency components are preserved.

These results have led to various methodologies that help us understand artificial neural networks through a frequency lens. One such method is Neural Anisotropic Directions (NADs) [6, 7]. NADs are input directions for which a network is able to linearly classify data. Furthermore, [8] introduced a method to compute a neural network’s sensitivity to input directions in the Fourier domain. Moreover, [9] show that robust deep learning object recognition models rely on low frequency information in natural images. Finally, [10] divides the image frequency spectrum into disjoint disks and provides evidence that mid or high-level frequencies are important for ANN classification.

In this work we introduce a simple and easy-to-use method to learn the input frequency features that a network deems essential in order to achieve its classification performance. We visualize the relevant frequencies by learning a modulatory mask on the Fourier transform of the input data that defines a modulation-invariant loss function obtained via a simple optimization algorithm (Section 2.1). We compare such masks with their adversarially trained or data augmented counterparts (Section 3). In the case of adversarial training, the comparison is done at two levels of analysis. At a global level, we learn a mask for the entire test set. Our goal is to find the frequencies that allow for robust generalization. At a single image level, we explore the frequencies responsible for adversarial success/failure. Those comparisons allow us to test the hypothesis that adversarially trained models have a bias towards low frequency features and assess if the same holds for other types of data augmentation.

In the case of adversarial augmentation, our results confirm the low frequency bias hypothesis. However, they also highlight that the important frequency redistribution due to the augmentation is highly anisotropic. In the case of common data augmentations instead, our results show how the frequency reorganization depends on the type of augmentation, e.g., rotation- or scale-augmented models exhibit mid-high and low frequency biases, respectively.

The single-image mask analysis reveals that only a few, class-specific frequencies are crucial to determine a network’s decision. Moreover, those frequencies are effectively the ones used to achieve its performance. In fact, mask-filtered images do not alter performance at all. However, surprisingly, they are not recognizable. They are characterized by texture-like patterns. This is in line with previous work by [11], which provided evidence that Convolutional Neural Networks (CNNs) are biased towards textures rather than shapes in object recognition. Our method differs from all previous ones in that we explicitly learn the frequencies defining the features a model is sensitive to.
2 Methods

2.1 Approach

Artificial neural networks and their associated task-dependent losses define highly non-linear functions of their input. In terms of the frequency content found in a signal, the effect of the application of a non-linear function can be understood by considering the following simple one-dimensional example. Suppose \( f(t) = \cos(w_1 t) + \cos(w_2 t) \) is a sound wave and let \( \sigma(t) = t^2 \). Then

\[
(\sigma \circ f)(t) = \frac{1}{2} [2 + \cos(2w_1 t) + \cos(2w_2 t) + 2 \cos((w_1 + w_2) t) + 2 \cos((w_1 - w_2) t)].
\]

We see that one of the effects of \( \sigma \) on \( f \) is to generate the new frequency components \( w_1 - w_2, w_1 + w_2, 2w_1, 2w_2 \). The first two are due to a phenomenon called intermodulation, the last are due to what is called harmonic distortion. Harmonic distortion has been studied in the context of neural networks with different activation functions by [12], where an empirical demonstration and theoretical arguments are given to support the claim that the presence of non-linear elements mainly causes a spread in the frequency content of the loss function. Their reasoning is the following: let \( \phi : \mathbb{R} \rightarrow \mathbb{R} \) be a non-linear function and \( T \phi \) denote its Taylor expansion around the origin. For \( x \in \mathbb{R}^d \), using the convolution theorem yields

\[
FT \phi(x) = F \sum_n a_n x \odot \cdots \odot x = \sum_n a_n \hat{x} \ast \cdots \ast \hat{x},
\]

where \( \phi \) is acting pointwise on the components of \( x \), \( Fx = \hat{x} \), and the RHS is a weighted sum of self-convolutions. [12] show that repeated convolutions broaden the frequency spectrum by adding higher frequency components corresponding to large coefficients \( a_n \), an effect they call “blue shift”. A visual illustration of the blue-shift effect is shown in Figure 1, where we considered a one dimensional sinusoidal stimulus \( s \) filtered by softplus, tanh, ReLU, and hardtanh non-linearities. Additional to the blue shift effect (harmonic distortion), we also see the impact of intermodulation.

Figure 1: Non-linear distortions in the frequency domain due to the application of (B) softplus, (C) tanh, (D) ReLU and (E) hardtanh non-linear activations on \( s(\cdot) = \sin(\cdot) \) (A).

Let us now consider a more complex non-linear function such as a trained neural network. In this case, the non-linear distortion induced by the network will be manifested in its representation space and therefore in its decision making.

As mentioned above, one of the purposes of this work is to propose an algorithm to identify the essential input frequencies in a trained ANN’s decisions. To this end, let us consider an image dataset \( \mathcal{X} = \{(x_i, y_i)\}_{i=1}^N \), where \( x_i \in \mathbb{R}^{d \times d} \) denotes the \( i \)-th input image and \( y_i \in \mathbb{Z}_C \) its associated label (\( C \) denotes the number of classes). We split \( \mathcal{X} \) into a training set \( \mathcal{X}_T \) and a validation set \( \mathcal{X}_V \). We obtain the masks via the following optimization algorithm: we first pre-train a network \( \Phi \) on \( \mathcal{X}_T \) with the objective of solving a classification task. We subsequently freeze the
weights of $\Phi$ and attach a pre-processing layer whose weights are the entries $m_{ij}$ of a mask matrix $M_\Phi \in \mathbb{R}^{d \times d}$. This layer acts as follows: for every $x \in \mathcal{X}_V$ we modulate its Fourier transform $\mathcal{F}x$ by computing the product $M_\Phi \circ \mathcal{F}x$, where $\circ$ indicates the Hadamard product. We next compute the inverse Fourier transform $\tilde{x} = \mathcal{F}^{-1}(M_\Phi \circ \mathcal{F}x)$, which is then fed into the network (see Figure 2). Finally, we learn the mask $M_\Phi$ by solving the optimization problem

$$M_\Phi(\lambda, p) = \arg\min_{M_\Phi} \sum_{x \in \mathcal{X}_V} e^{[\mathcal{L}(\Phi(\tilde{x}), y) - \mathcal{L}(\Phi(x), y)]^2} + \lambda \|M_\Phi\|_p, \quad \lambda \in \mathbb{R}^+, \quad (2)$$

where $\Phi$ denotes the pre-trained network, $\lambda \|M_\Phi\|_p$ is a regularization term penalizing the $p$-norm of the learned mask, and $\mathcal{L}$ is the loss function associated with the classification task. The first term in Equation (2) enforces an invariance in the loss with respect to the transformation $x \mapsto \tilde{x}$ induced by the mask. The latter is key because we are expecting the desired frequencies to be revealed when there is no change in the loss $\mathcal{L}$ and maximal change in the $p$-norm of the mask $M_\Phi$. In other words, the mask is determined by a symmetry operation in the Fourier space of the input with minimal $p$-norm. A solution to Equation (2) is a mask $M_\Phi$ addressing the question: which frequencies are essential in this trained ANN’s decision making? Such masks, obtained for various data augmentation choices reveal the frequencies associated with each particular choice.

At this point, we note that the mask is learned on the validation set $\mathcal{X}_V$ and not on the training set $\mathcal{X}_T$. This is because we are interested in exploring the minimal set of frequencies preserving generalization power of $\Phi$. Moreover, we tested the stability of our mask generation algorithm across different runs. This is crucial since it attests to the reliability of the qualitative and quantitative analysis. We also note that masks can be obtained for a single image, simply considering a single $x \in \mathcal{X}_V$ in Equation (2) instead of the full validation set.

2.2 Dataset and simulations

Our data consisted of 6,644 image/label pairs from 5 classes of ImageNet [13]. 4,710 of those pairs belong to our training set $\mathcal{X}_T$ and the remaining 1,934 pairs belong to our validation set $\mathcal{X}_V$. For simplicity, we choose grayscale versions of our dataset images, though our method can be applied for any number of input channels. Our images were centered with respect to the mean and standard deviation of $\mathcal{X}_T$.

We initially trained a VGG11 [14] baseline model on $\mathcal{X}_T$ using the Pytorch framework. For each subsequent training run we varied the type of data augmentation used for pre-processing (adversarial examples, random scales, random translations, random rotations).
Each of the 5 networks in total was trained using the Adam optimizer ([15]) and a maximum learning rate of $10^{-3}$. The learning rate of each learnable parameter group was scheduled according to the one-cycle learning rate policy with a minimum value of 0 ([16]). We found that this set of hyperparameter choices allowed us to achieve stable training for all our models. We trained each model for a maximum of 50 epochs and eventually evaluated our models on the validation set $X_V$. We finally saved the weight-state of each model that achieved the minimum Cross Entropy loss within the chosen interval of epochs. For each of our pre-trained networks, we train its corresponding Fourier mask according to the algorithmic process presented in section 2.1. We use $\ell_1$-regularization on the norm of the mask to ensure sparsity. In the next section we present masks for every data augmentation scheme we chose as well as their respective differences. For a given set of masks, we center the mask differences around the origin. This helps with the interpretation of the masks without altering the geometry of the particular set.

3 Results

Adversarial training can be seen as a type of data augmentation where the inputs are augmented with adversarial examples [2] to increase robustness to adversarial attacks. Here we test the commonly accepted hypothesis that adversarially trained models need low frequency features for robustness. We do so by comparing the Fourier mask learned for a vanilla network $\Phi_\lambda$ with that of an adversarially trained network $\Phi_\lambda$ when the learning occurs over the whole validation set. Specifically, we compare a naturally trained VGG11 with an adversarially trained one using the torchattacks library [17] and a Projected Gradient Descent attack (PGD). [18] has shown the frequency structures of adversarial attacks are similar across different adversarial attacks. Therefore, although the set of potential choices one can explore is vast, in this work we focus on PGD for simplicity. Besides the mask difference we also compute the radial and angular energy of each mask by considering radial and angular partitions of the frequency domain (Figure 4 (A), (B)). We then test if the same low-frequency preference hypothesis holds true in the case of common data augmentations. To gain some intuition, let us consider a simple one layer network $(A), (B))$. We then test if the same low-frequency preference hypothesis holds true in the case of adversarial examples [2] to increase robustness to adversarial attacks. Here we test the other words, the frequency content of the loss, as well as how it is modified by different data augmentations, will impact the frequency content observed in the mask.

Let us consider a simple one dimensional example ($d = 1$) and the translation operator. In this case the loss $\mathcal{L}$ is invariant to translations of the weights, i.e.,

$$\mathcal{L}(T_t(w)) = \mathcal{L}(w), \quad \forall t \in \mathbb{R},$$

where the second equality holds because $(w, g_\theta x_i) = (g_\theta^* w, x_i)$ and $g_\theta^*$ denotes the adjoint. We note that in this context the loss function is invariant to $G$ transformations of the weights, i.e., $\mathcal{L}(g_\theta w) = \mathcal{L}(w)$ for any $g_\theta \in G$ (the proof of this statement relies on simple properties of group transformations, see [19]). Here we explore the impact such an invariance of the loss function has on the learned Fourier masks. The reasoning is as follows: updating the weights of an ANN is achieved through gradient descent, i.e., $\Delta w_t = -a \nabla_w \mathcal{L}(w_t)$, where $w_t$ denotes the weights of the network at iteration $t$ and $a \in \mathbb{R}^+$ is the learning rate. The frequency content of the gradient of the loss at iteration $t$ affects the frequency content of the weights. In turn, the latter determine the input frequencies the network is analyzing and thus will determine the mask. In other words, the frequency content of the loss, as well as how it is modified by different data augmentations, will impact the frequency content observed in the mask.
where $T_t : \mathbb{R} \to \mathbb{R}$ is the translation operator defined as $T_t(\cdot) = \cdot - t$. For $x_i \in \mathcal{X}$ and $t \in \mathbb{R}$, let $q_i(\cdot) := \ell(\sigma(T_t(\cdot)x_i); y_i)$. Then the Fourier transform of $\mathcal{L}$ yields

$$
\mathcal{F}(\mathcal{L})(\gamma) = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{+\infty} \mathcal{F}(q_i)(\gamma)e^{-2\pi i \gamma t} dt = \frac{1}{N} \sum_{i=1}^{N} \delta(\gamma)\mathcal{F}(q_i)(\gamma) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{F}(q_i)(0)
$$

where we used the translation property of the Fourier transform and $\delta$ denotes the Dirac delta. This simple example illustrates the effect of the translation operator on the loss $\mathcal{L}$, i.e., a shift towards low frequencies (in this case a full shift of all frequencies to the DC component, the only non-zero component in the above equation). Note that an augmentation with all possible translations is not realistic. However, even a finite range of translations in the interval $t \in [-a, a]$, for a sufficiently large $a$, will produce a similar effect. Indeed, we have

$$
\mathcal{F}(\mathcal{L})(\gamma) = \frac{1}{N} \sum_{i=1}^{N} \int_{-\infty}^{+\infty} \mathcal{F}(q_i)(\gamma)e^{-2\pi i \gamma t} dt = \frac{2a}{N} \sum_{i=1}^{N} \text{sinc}(2\pi \gamma a)\mathcal{F}(q_i)(\gamma)
$$

where $\chi$ denotes the characteristic function. Thus, the impact of averaging over an interval of translations on $\mathcal{L}$ is to dampen its frequencies with a sinc function profile, i.e., a frequency re-weighting with a bias for low frequencies. However, we stress that the above argument is developed with a 1-layer network in mind. The effect of data-augmentation with respect to random translations viewed through a deep network is expected to be more intricate.

### 3.1 Masks generated for the whole dataset

We generated masks over $\mathcal{X}$ for networks trained to be robust to adversarial examples, random scales, translations, and rotations. The masks in Figure 3 (Top) and their differences reveal how distinct frequency biases depend on the type of data augmentation. We also note how model performance is not altered by the introduction of the mask layer Figure 3 (Bottom). In the case of adversarial augmentation there exists a net bias towards low frequencies as shown by the

![Figure 3: Top: Learned masks for (A) vanilla network ($M_N$), (B) Adversarially trained ($M_A$) (C) scale-invariant network ($M_S$), (D) translation-invariant network ($M_T$), (E) rotation-invariant network ($M_R$) and their differences. Bottom: Model performance with and without the mask layer.](image-url)
difference between the masks generated by the vanilla and adversarial trained network in Figure 3 (B-A). This is further confirmed by the radial energy difference in Figure 4 (A1), while the angular energy difference in Figure 4 (B1) shows that the redistribution of the frequencies occurs anisotropically.

Figure 4: (A) Radial Partitions of the Fourier domain (B) Angular partitions of the Fourier domain; Energy difference in radial (A1, A2, A3, A4) and angular (B1, B2, B3, B4) directions, for the augmentations in B, C, D, E of Figure 3 (∥M_i∥_2 − ∥M_N∥_2, i = adversarial (A), scale (S), translation (T) and rotation (R) augmentations).
In the case of common augmentations our results exhibit contrasting effects in the Fourier masks. While the redistribution of the mask frequencies seems to be directionally-dependent (Figure 4B1, B2, B3), only robustness to scales endows the net with a bias towards low frequencies (Figure 4(A2)). For translations the mask implies a less clear effect (Figure 4(A3)), where a mixed behavior is present for mid and low frequencies. Interestingly, in the case of rotational robustness, Figure 4(A4) shows a high frequency bias.

3.2 Masks generated for single images

To further investigate the nature of adversarial robustness and how it is related to a network’s generalization properties in the frequency domain we generated Fourier masks $M_{N,x}$ for each single image $x$ in the validation set $X_V$. Figure 5(Top) shows such masks randomly sampled for images in all 5 data classes calculated for the vanilla network $\Phi_N$.

![Figure 5](image)

Figure 5: (Top) Randomly sampled single image masks divided by class. The colormap is the same as that in Figure 3. (Bottom) (A) randomly sampled images; (B) filtered by their complementary masks $M'_{N,x}$; (C) the same images filtered by their associated Fourier masks $M_{N,x}$.

It is worth noting that the masks are very sparse, i.e., very few frequencies are essential for preserving the prediction of the pretrained network. Additionally, for every mask $M_{N,x}$, we also consider its complementary mask $M'_{N,x}$ defined as

$$M'_{N,x}(i,j) = \begin{cases} 1, & M_{N,x}(i,j) < 10^{-8} \\ 0, & \text{otherwise.} \end{cases}$$
Filtering an image with its complementary mask \( M'_{N,x} \) does not compromise our ability to recognize the filtered image (Figure 5(B, Bottom)). On the contrary, filtering with the mask \( M_{N,x} \) renders the image unrecognizable (Figure 5(C, Bottom)). Filtered images resemble texture-like patterns. Interestingly, a recent work by [20] shows how ImageNet-trained CNNs are strongly biased towards recognizing textures rather than shapes. Surprisingly, the opposite phenomenon is observed in the way the network perceives the changes induced by the respective masks. Performance drops drastically (\(~45\%\) decrease) for images filtered by complementary masks \( M'_{N,x} \) (Figure 6(Bottom)). Additionally, filtering adversarial examples using masks generated from original images reverses the effect of the attack, restoring the original performance of the network.

Figure 6: (Top): (A) Correctly classified image \( x \in X \) (Blue); (B) Adversarial Image \( x_A \) (Orange); (C) \( M_{N,x} \) - learned mask for vanilla network; (D) \( M_{N,x} \) - filtered image \( x \); (E) \( M_{A,x_A} \) - learned mask for adversarially trained network; (F) \( M_{A,x_A} \) - filtered adversarial image \( x_A \); (G) \( M_{N,x_A} \) - learned mask for vanilla network using the adversarial image \( x_A \); (H) \( M_{N,x_A} \) - filtered adversarial image \( x_A \). The colormap is the same as that in Figure 3. (Bottom): Model performance on dataset \( X \) filtered by single-image vanilla masks \( M_{N,X} \); filtered by their complementary masks \( M'_{N,X} \); performance on the adversarial examples \( X_A \); and on the adversarial examples \( x_A \) filtered by single-image vanilla masks \( M_{N,x} \). We generate masks only for test images that the pretrained model classifies correctly.

We also computed single image masks for clean images \( x \) and adversarial images \( x_A \) using either the naturally pretrained network \( \Phi_N \) or the adversarially pretrained network \( \Phi_A \). We compared the masks \( M_{N,X} \) with \( M_{A,X_A} \) to highlight the most important frequencies used by an adversarially trained network to make its predictions robust. We then compared the masks \( M_{N,X} \) with \( M_{N,X_A} \) to assess which frequencies are responsible for a network to make a wrong
prediction on the adversarial image. Figure 6 shows examples of such masks calculated for five randomly chosen validation images from each class. We note that the masks $M_{A,X_A}$ have more energy concentrated in lower frequencies compared to those in $M_{N,X}$, further confirming a low frequency bias in adversarially trained networks. These observations are quantitatively substantiated by the plots in Fig. 7 where we computed the percentage of perturbed images for which the per-band energy (radial or angular) of their corresponding masks $M_{A,X_A}$ exceeds that of the masks $M_{N,X}$ generated from the non-perturbed examples. Fig. 7 confirms that lower frequencies are preferred for a robust representation.

Figure 7: Comparison of $X$-trained masks ($M_{N,X}$) obtained for the vanilla network and $X_A$-trained masks ($M_{A,X_A}$) obtained for an adversarially trained network: the bars illustrate the percentage of masks for which the per-band energy in $M_{A,X_A}$ exceeds that of $M_{N,X}$.

We then performed a manifold analysis of the learned masks. We found that the masks are linearly separable and that the linear network responses cluster (Figure 8). At the same time, we see that this is not the case when the labels associated with the masks are shuffled. Therefore, the frequencies the network deems essential for prediction are class-specific since the results suggest that linear separability of the masks is due to their geometry and not the representation power of the linear classifier.

Figure 8: Clustering of the output of a linear classifier learned to separate single image Fourier masks. $M_{N,X}$ indicates the set of single image masks computed for the correctly classified test images in $X$.
4 Discussion and Conclusions

In this work we proposed a simple yet powerful approach to visualize the essential frequencies a trained network is using to solve a task. Our strategy consists of learning a frequency modulatory mask characterized by two critical properties:

- it defines a symmetry in the Cross Entropy loss, i.e., it does not alter the pretrained model’s predictions.
- it has minimal $\ell_p$-norm, which for $p = 1$ guarantees the preservation of performance while promoting sparsity in the mask.

Using our method we tested the common hypothesis that adversarially trained networks prefer low frequency features to achieve robustness. We also tested if this hypothesis holds true for common data augmentations such as translations, scales, and rotations.

In the case of adversarial augmentation, our results confirm the low frequency bias hypothesis. However, they also highlight that the frequency redistribution due to the augmentation is highly anisotropic. In the case of common data augmentations instead, our results show how the frequency reorganization depends on the type of augmentation.

In the case of adversarial training we also run a single image analysis to detect the frequencies useful for adversarial robustness and those responsible for adversarial weakness. Here too, masks learned on adversarially trained networks concentrate more towards lower frequencies compared to those learned on vanilla networks. Furthermore, the analysis showed that only a sparse, class-specific set of frequencies is needed to classify an image. Surprisingly, mask-filtered images in this case are not recognizable and resemble texture-like patterns, supporting the idea that ANNs use fundamentally different classification strategies from humans to achieve robust generalization [11].

To our knowledge the use of a learned mask to characterize a network’s crucial property such as robust generalization has not been proposed before. The interpretation of the masks provides us with a detailed geometrical description of directional and radial biases in the frequency domain as well as with quantifiable differences between various training schemes.

Our analysis can be extended to other architectural or optimization specifics, e.g., explicit regularizations, different optimizers/initializations, etc. The same mask approach can be employed to modulate the phase and modulus in the Fourier transform of the data. Our method effectively opens up many directions in the investigation of a network’s implicit frequency bias. Future research directions will also include a natural generalization of our approach where the image features are learned, rather than fixed to be of the Fourier type.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

N.K. and F.A. conceived the conceptualized framework. N.K. and E.B. trained and analyzed models. N.K. and F.A. wrote the first draft. A.P., J.O.C., A.S.T. and X.P. provided feedback along the way. All authors revised, edited and provided comments on the final manuscript. A.S.T., A.P. and X.P. provided funding.
Availability of data and materials

Code is available at https://github.com/nkarantzas/FourierMasks.git

Acknowledgments

N.K., E.B., J.O.C., A.P., A.S.T., X.P. and F.A. acknowledges funding support from IARPA MICRONS, DARPA L2M, and NSF Neuronex. We also thank Shell Xu Hu, Kandan Ramakrishnan, and Zhe Li for helpful discussions.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. Nature, 521, 436-444, 2015.
[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.
[3] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. Advances in Neural Information Processing Systems, 32:13276–13286, 2019.
[4] Zifan Wang, Yilin Yang, Ankit Shrivastava, Varun Rawal, and Zihao Ding. Towards frequency-based explanation for robust cnn, 2020.
[5] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the Effectiveness of Low Frequency Perturbations. In IJCAI, 2019.
[6] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Hold me tight! influence of discriminative features on deep network boundaries. arXiv preprint arXiv:2002.06349, 2020.
[7] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Neural anisotropy directions. arXiv preprint arXiv:2006.09717, 2020.
[8] Yusuke Tsuzuku and Issei Sato. On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 51–60, 2019.
[9] Zhe Li, Josue Ortega Caro, Evgenia Rusak, Wieland Brendel, Matthias Bethge, Fabio Anselmi, Ankit B. Patel, Andreas S. Tolias, and Xaq Pitkow. Robust deep learning object recognition models rely on low frequency information in natural images. bioRxiv, 2022.
[10] Antonio A. Abello, Roberto Hirata, and Zhangyang Wang. Dissecting the high-frequency bias in convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 863–871, June 2021.
[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations, 2019.
[12] Mehmeti-Göpel Christian, Hartmann David, and Wand Michael. Ringing relus: Harmonic distortion analysis of nonlinear feedforward networks. In International Conference on Learning Representations, 2021.
[13] Jia Deng, Wei Dong, Richard Socher, Jia Li, Li Kai, and Fei Fei Li. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.
[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[16] Leslie N. Smith. Cyclical learning rates for training neural networks. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 464–472, 2017.

[17] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. arXiv preprint arXiv:2010.01950, 2020.

[18] Josué Ortega Caro, Yilong Ju, Ryan Pyle, Sourav Dey, Wieland Brendel, Fabio Anselmi, and Ankit Patel. Local convolutions cause an implicit bias towards high frequency adversarial examples. arXiv preprint arXiv:2006.11440, 2020.

[19] Shuxiao Chen, Edgar Dobriban, and Jane Lee. A group-theoretic framework for data augmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 21321–21333. Curran Associates, Inc., 2020.

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231, 2018.