

# High-frequency Component Helps Explain the Generalization of Convolutional Neural Networks

Haohan Wang, Xindi Wu, Zeyi Huang, Eric P. Xing  
 School of Computer Science  
 Carnegie Mellon University

{haohanw, epxing}@cs.cmu.edu, {xindiw, zeyih}@andrew.cmu.edu

## Abstract

We investigate the relationship between the frequency spectrum of image data and the generalization behavior of convolutional neural networks (CNN). We first notice CNN’s ability in capturing the **high-frequency components** of images. These high-frequency components are almost imperceptible to a human. Thus the observation leads to multiple hypotheses that are related to the generalization behaviors of CNN, including a potential explanation for adversarial examples, a discussion of CNN’s trade-off between robustness and accuracy, and some evidence in understanding training heuristics.

## 1. Introduction

Deep learning has achieved many recent advances in predictive modeling in various tasks, but the community has nonetheless become alarmed by the unintuitive generalization behaviors of neural networks, such as the capacity in memorizing label shuffled data [65] and the vulnerability towards adversarial examples [54, 21]

To explain the generalization behaviors of neural networks, many theoretical breakthroughs have been made progressively, including studying the properties of stochastic gradient descent [31], different complexity measures [46], generalization gaps [50], and many more from different model or algorithm perspectives [30, 43, 7, 51].

In this paper, inspired by previous understandings that convolutional neural networks (CNN) can learn from confounding signals [59] and superficial signals [29, 19, 58], we investigate the generalization behaviors of CNN from a data perspective. Together with [27], we suggest that the unintuitive generalization behaviors of CNN as a direct outcome of the perceptual disparity between human and models (as argued by Figure 1): *CNN can view the data at a much higher granularity than the human can.*

However, different from [27], we provide an interpreta-

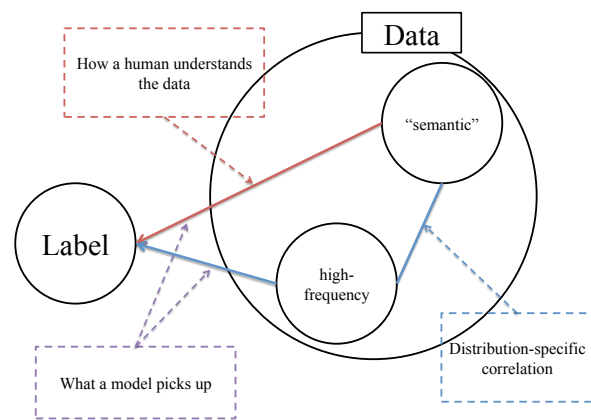


Figure 1. The central hypothesis of our paper: within a data collection, there are correlations between the high-frequency components and the “semantic” component of the images. As a result, the model will perceive both high-frequency components as well as the “semantic” ones, leading to generalization behaviors counter-intuitive to human (e.g., adversarial examples).

tion of this high granularity of the model’s perception: *CNN can exploit the high-frequency image components that are not perceivable to human.*

For example, Figure 2 shows the prediction results of eight testing samples from CIFAR10 data set, together with the prediction results of the high and low-frequency component counterparts. For these examples, the prediction outcomes are almost entirely determined by the high-frequency components of the image, which are barely perceivable to human. On the other hand, the low-frequency components, which almost look identical to the original image to human, are predicted to something distinctly different by the model.

Motivated by the above empirical observations, we further investigate the generalization behaviors of CNN and attempt to explain such behaviors via differential responses to the *image frequency spectrum* of the inputs (Remark 1). Our main contributions are summarized as follows:

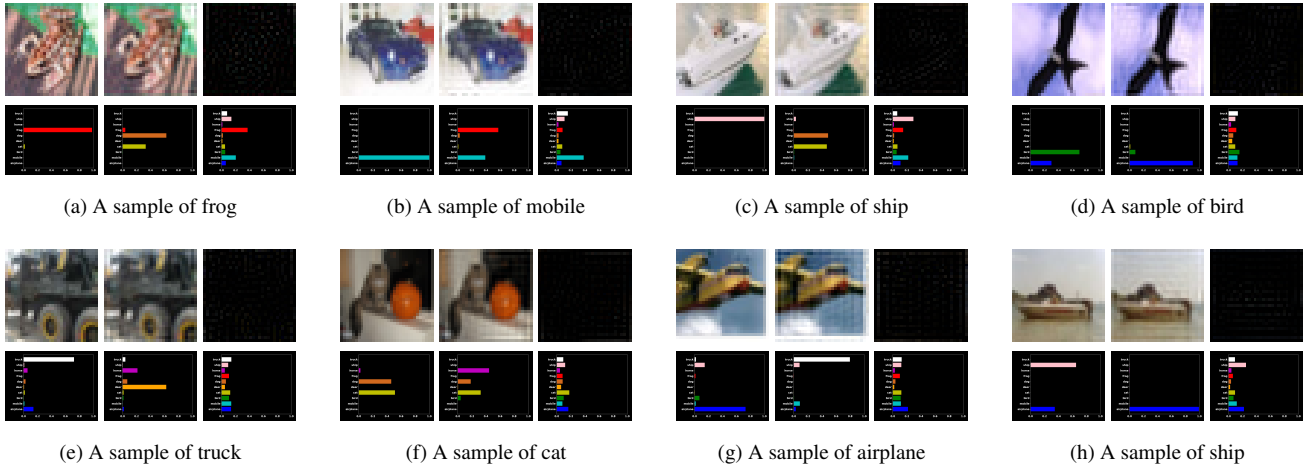


Figure 2. Eight testing samples selected from CIFAR10 that help explain that CNN can capture the high-frequency image: the model (ResNet18) correctly predicts the original image (1<sup>st</sup> column in each panel) and the high-frequency reconstructed image (3<sup>rd</sup> column in each panel), but incorrectly predict the low-frequency reconstructed image (2<sup>nd</sup> column in each panel). The prediction confidences are also shown. The frequency components are split with  $r = 12$ . Details of the experiment will be introduced later.

- We reveal the existing trade-off between CNN’s accuracy and robustness by offering examples of how CNN exploits the high-frequency components of images to trade robustness for accuracy (Corollary 1).
- With image frequency spectrum as a tool, we offer hypothesis to explain several generalization behaviors of CNN, especially the capacity in memorizing label-shuffled data.
- We propose defense methods that can help improving the adversarial robustness of CNN towards simple attacks without training or fine-tuning the model.

The remainder of the paper is organized as follows. In Section 2, we first introduce related discussions. In Section 3, we will present our main contributions, including a formal discussion on that CNN can exploit high-frequency components, which naturally leads to the trade-off between adversarial robustness and accuracy. Further, in Section 4-6, we set forth to investigate multiple generalization behaviors of CNN, including the paradox related to capacity of memorizing label-shuffled data (§4), the performance boost introduced by heuristics such as Mixup and BatchNorm (§5), and the adversarial vulnerability (§6). We also attempt to investigate tasks beyond image classification in Section 7. Finally, we will briefly discuss some related topics in Section 8 before we conclude the paper in Section 9.

## 2. Related Work

The remarkable success of deep learning has attracted a torrent of theoretical work devoted to explaining the generalization mystery of CNN.

For example, ever since Zhang *et al.* [65] demonstrated the effective capacity of several successful neural network architectures is large enough to memorize random labels, the community sees a prosperity of many discussions about this apparent “paradox” [61, 15, 17, 15, 11]. Arpit *et al.* [3] demonstrated that effective capacity are unlikely to explain the generalization performance of gradient-based-methods trained deep networks due to the training data largely determine memorization. Kruger *et al.* [35] empirically argues by showing largest Hessian eigenvalue increased when training on random labels in deep networks.

The concept of adversarial example [54, 21] has become another intriguing direction relating to the behavior of neural networks. Along this line, researchers invented powerful methods such as FGSM [21], PGD [42], and many others [62, 9, 53, 36, 12] to deceive the models. This is known as *attack methods*. In order to defend the model against the deception, another group of researchers proposed a wide range of methods (known as *defense methods*) [1, 38, 44, 45, 24]. These are but a few highlights among a long history of proposed attack and defense methods. One can refer to comprehensive reviews for detailed discussions [2, 10]

However, while improving robustness, these methods may see a slight drop of prediction accuracy, which leads to another thread of discussion in the trade-off between robustness and accuracy. The empirical results in [49] demonstrated that more accurate model tend to be more robust over generated adversarial examples. While [25] argued that the seemingly increased robustness are mostly due to the increased accuracy, and more accurate models (*e.g.*, VGG, ResNet) are actually less robust than AlexNet. Theoretical discussions have also been offered [56, 67], which also inspires new defense methods [67].

### 3. High-frequency Components & CNN's Generalization

We first set up the basic notations used in this paper:  $\langle \mathbf{x}, \mathbf{y} \rangle$  denotes a data sample (the image and the corresponding label).  $f(\cdot; \theta)$  denotes a convolutional neural network whose parameters are denoted as  $\theta$ . We use  $\mathcal{H}$  to denote a human model, and as a result,  $f(\cdot; \mathcal{H})$  denotes how human will classify the data  $\cdot$ .  $l(\cdot, \cdot)$  denotes a generic loss function (e.g., cross entropy loss).  $\alpha(\cdot, \cdot)$  denotes a function evaluating prediction accuracy (for every sample, this function yields 1.0 if the sample is correctly classified, 0.0 otherwise).  $d(\cdot, \cdot)$  denotes a function evaluating the distance between two vectors.  $\mathcal{F}(\cdot)$  denotes the Fourier transform; thus,  $\mathcal{F}^{-1}(\cdot)$  denotes the inverse Fourier transform. We use  $\mathbf{z}$  to denote the frequency component of a sample. Therefore, we have  $\mathbf{z} = \mathcal{F}(\mathbf{x})$  and  $\mathbf{x} = \mathcal{F}^{-1}(\mathbf{z})$ .

Notice that Fourier transform or its inverse may introduce complex numbers. In this paper, we simply discard the imaginary part of the results of  $\mathcal{F}^{-1}(\cdot)$  to make sure the resulting image can be fed into CNN as usual.

#### 3.1. CNN Exploit High-frequency Components

We decompose the raw data  $\mathbf{x} = \{\mathbf{x}_l, \mathbf{x}_h\}$ , where  $\mathbf{x}_l$  and  $\mathbf{x}_h$  denote the low-frequency component (shortened as LFC) and high-frequency component (shortened as HFC) of  $\mathbf{x}$ . We have the following four equations:

$$\begin{aligned} \mathbf{z} &= \mathcal{F}(\mathbf{x}), & \mathbf{z}_l, \mathbf{z}_h &= t(\mathbf{z}; r), \\ \mathbf{x}_l &= \mathcal{F}^{-1}(\mathbf{z}_l), & \mathbf{x}_h &= \mathcal{F}^{-1}(\mathbf{z}_h), \end{aligned}$$

where  $t(\cdot; r)$  denotes a thresholding function that separates the low and high frequency components from  $\mathbf{z}$  according to a hyperparameter, radius  $r$ .

To define  $t(\cdot; r)$  formally, we first consider a grayscale (one channel) image of size  $n \times n$  with  $\mathcal{N}$  possible pixel values (in other words,  $\mathbf{x} \in \mathcal{N}^{n \times n}$ ), then we have  $\mathbf{z} \in \mathcal{C}^{n \times n}$ , where  $\mathcal{C}$  denotes the complex number. We use  $\mathbf{z}(i, j)$  to index the value of  $\mathbf{z}$  at position  $(i, j)$ , and we use  $c_i, c_j$  to denote the centroid. We have the equation  $\mathbf{z}_l, \mathbf{z}_h = t(\mathbf{z}; r)$  formally defined as:

$$\begin{aligned} \mathbf{z}_l(i, j) &= \begin{cases} \mathbf{z}(i, j), & \text{if } d((i, j), (c_i, c_j)) \leq r \\ 0, & \text{otherwise} \end{cases}, \\ \mathbf{z}_h(i, j) &= \begin{cases} 0, & \text{if } d((i, j), (c_i, c_j)) \leq r \\ \mathbf{z}(i, j), & \text{otherwise} \end{cases} \end{aligned}$$

We consider  $d(\cdot, \cdot)$  in  $t(\cdot; r)$  as the Euclidean distance in this paper. If  $\mathbf{x}$  has more than one channel, then the procedure operates on every channel of pixels independently.

**Remark 1.** With an assumption (referred to as A1) that presumes “only  $\mathbf{x}_l$  is perceivable to human, but both  $\mathbf{x}_l$  and  $\mathbf{x}_h$

are perceivable to a CNN,” we have:

$$\mathbf{y} := f(\mathbf{x}; \mathcal{H}) = f(\mathbf{x}_l; \mathcal{H}),$$

but when a CNN is trained with

$$\arg \min_{\theta} l(f(\mathbf{x}; \theta), \mathbf{y}),$$

which is equivalent to

$$\arg \min_{\theta} l(f(\{\mathbf{x}_l, \mathbf{x}_h\}; \theta), \mathbf{y}),$$

CNN may learn to exploit  $\mathbf{x}_h$  to minimize the loss. As a result, CNN's generalization behavior appears unintuitive to a human.  $\square$

Notice that “CNN may learn to exploit  $\mathbf{x}_h$ ” differs from “CNN overfit” because  $\mathbf{x}_h$  can contain more information than sample-specific idiosyncrasy, and these more information can be generalizable across training, validation, and testing sets, but are just imperceptible to a human.

As Assumption A1 has been demonstrated to hold in some cases (e.g., in Figure 2), we believe Remark 1 can serve as one of the explanations to CNN's generalization behavior. For example, the adversarial examples [54, 21] can be generated by perturbing  $\mathbf{x}_h$ ; the capacity of CNN in reducing training error to zero over label shuffled data [65] can be seen as a result of exploiting  $\mathbf{x}_h$  and overfitting sample-specific idiosyncrasy. We will discuss more in the following sections.

#### 3.2. Trade-off between Robustness and Accuracy

We continue with Remark 1 and discuss CNN's trade-off between robustness and accuracy given  $\theta$  from the image frequency perspective. We first formally state the accuracy of  $\theta$  as:

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y})} \alpha(f(\mathbf{x}; \theta), \mathbf{y}) \quad (1)$$

and the adversarial robustness of  $\theta$  as in e.g., [8]:

$$\mathbb{E}_{(\mathbf{x}, \mathbf{y})} \min_{\mathbf{x}': d(\mathbf{x}', \mathbf{x}) \leq \epsilon} \alpha(f(\mathbf{x}'; \theta), \mathbf{y}) \quad (2)$$

where  $\epsilon$  is the upper bound of the perturbation allowed.

With another assumption (referred to as A2): “for model  $\theta$ , there exists a sample  $\langle \mathbf{x}, \mathbf{y} \rangle$  such that:

$$f(\mathbf{x}; \theta) \neq f(\mathbf{x}_l; \theta),”$$

we can extend our main argument (Remark 1) to a formal statement:

**Corollary 1.** *With assumptions A1 and A2, there exists a sample  $\langle \mathbf{x}, \mathbf{y} \rangle$  that the model  $\theta$  cannot predict both accurately (evaluated to be 1.0 by Equation 1) and robustly (evaluated to be 1.0 by Equation 2) under any distance metric  $d(\cdot, \cdot)$  and bound  $\epsilon$  as long as  $\epsilon \geq d(\mathbf{x}, \mathbf{x}_l)$ .*

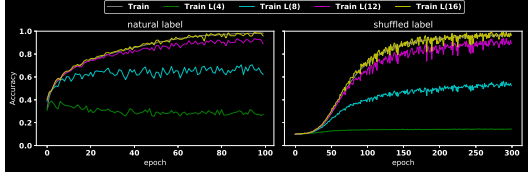


Figure 3. Training curves of the original label case (100 epoches) and shuffled label case (300 epoches), together plotted with the low-frequent counterpart of the images. All curves in this figure are from train samples.

The proof is a direct outcome of the previous discussion and thus omitted. The Assumption A2 can also be verified empirically (*e.g.*, in Figure 2), therefore we can safely state that Corollary 1 can serve as one of the explanations to the trade-off between CNN’s robustness and accuracy.

## 4. Rethinking Data before Rethinking Generalization

### 4.1. Hypothesis

Our first aim is to offer some intuitive explanations to the empirical results observed in [65]: neural networks can easily fit label-shuffled data. While we have no doubts that neural networks are capable of memorizing the data due to its capacity, the interesting question arises: “if a neural network can easily memorize the data, why it cares to learn the generalizable patterns out of the data, in contrast to directly memorizing everything to reduce the training loss?”

Within the perspective introduced in Remark 1, our hypothesis is as follows: Despite the same outcome as a minimization of the training loss, the model considers different level of features in the two situations:

- In the original label case, the model will first pick up LFC, then gradually pick up the HFC to achieve higher training accuracy.
- In the shuffled label case, as the association between LFC and the label is erased due to shuffling, the model has to memorize the images when the LFC and HFC are treated equally.

### 4.2. Experiments

We set up the experiment to test our hypothesis. We use ResNet-18 [22] for CIFAR10 dataset [33] as the base experiment. The vanilla set-up, which we will use for the rest of this paper, is to run the experiment with 100 epoches with the ADAM optimizer [32] with learning rate set to be  $10^{-4}$  and batch size set to be 100, when weights are initialized with Xavier initialization [20]. Pixels are all normalized to be  $[0, 1]$ . All these experiments are repeated in MNIST [14], FashionMNIST [63], and a subset of ImageNet [13]. These efforts are reported in the Appendix. We train two models,

Table 1. We test the generalization power of LFC and HFC by training the model with  $\mathbf{x}_l$  or  $\mathbf{x}_h$  and test on the original test set.

| LFC |            |           | HFC |            |           |
|-----|------------|-----------|-----|------------|-----------|
| $r$ | train acc. | test acc. | $r$ | train acc. | test acc. |
| 4   | 0.9668     | 0.6167    | 4   | 0.9885     | 0.2002    |
| 8   | 0.9786     | 0.7154    | 8   | 0.9768     | 0.092     |
| 12  | 0.9786     | 0.7516    | 12  | 0.9797     | 0.0997    |
| 16  | 0.9839     | 0.7714    | 16  | 0.9384     | 0.1281    |

with the natural label setup and the shuffled label setup, denote as  $M_{\text{natural}}$  and  $M_{\text{shuffle}}$ , respectively; the  $M_{\text{shuffle}}$  needs 300 epoches to reach a comparative training accuracy. To test which part of the information the model picks up, for any  $\mathbf{x}$  in the training set, we generate the low-frequency counterparts  $\mathbf{x}_l$  with  $r$  set to 4, 8, 12, 16 respectively. We test the how the training accuracy changes for these low-frequency data collections along the training process.

The results are plotted in Figure 3. The first message is the  $M_{\text{shuffle}}$  takes a longer training time than  $M_{\text{natural}}$  to reach the same training accuracy (300 epoches vs. 100 epoches), which suggests that memorizing the samples as an “unnatural” behavior in contrast to learning the generalizable patterns. By comparing the curves of the low-frequency training samples, we notice that  $M_{\text{natural}}$  learns more of the low-frequency patterns (*i.e.*, when  $r$  is 4 or 8) than  $M_{\text{shuffle}}$ . Also,  $M_{\text{shuffle}}$  barely learns any LFC when  $r = 4$ , while on the other hand, even at the first epoch,  $M_{\text{natural}}$  already learns around 40% of the correct LFC when  $r = 4$ . This disparity suggests that when  $M_{\text{natural}}$  prefers to pick up the LFC,  $M_{\text{shuffle}}$  does not have a preference between LFC vs. HFC.

If a model can exploit multiple different sets of signals, then why  $M_{\text{natural}}$  prefers to learn LFC that happens to align well with the human perceptual preference? While there are explanations suggesting neural networks’ tendency towards simpler functions [48], we conjecture that this is simply because, since the data sets are organized and annotated by human, the LFC-label association is more “generalizable” than the one of HFC: picking up LFC-label association will lead to the steepest descent of the loss surface, especially at the early stage of the training.

To test this conjecture, we repeat the experiment of  $M_{\text{natural}}$ , but instead of the original train set, we use the  $\mathbf{x}_l$  or  $\mathbf{x}_h$  (normalized to have the standard pixel scale) and test how well the model can perform on original test set. Table 1 suggests that LFC is much more “generalizable” than HFC. Thus, it is not surprising if a model first picks up LFC as it leads to the steepest descent of the loss surface.

### 4.3. A Remaining Question

Finally, we want to raise a question: The coincidental alignment between networks’ preference in LFC and human perceptual preference might be a simple result of the “survival bias” of the many technologies invented one of the other along the process of climbing the ladder of the



state-of-the-art. In other words, the almost-100-year development process of neural networks functions like a “natural selection” of technologies [60]. The survived ideas may happen to match the human preferences, otherwise, the ideas may not even be published due to the incompetence in climbing the ladder.

However, an interesting question will be how well these ladder climbing techniques align with the human visual preference. We offer to evaluate these techniques with our frequency tools.

## 5. Training Heuristics

We continue to reevaluate the heuristics that helped in climbing the ladder of state-of-the-art accuracy. We evaluate these heuristics to test the generalization performances towards LFC and HFC. Many renowned techniques in the ladder of accuracy seem to exploit HFC more or less.

### 5.1. Comparison of Different Heuristics

We test multiple heuristics by inspecting the prediction accuracy over LFC and HFC with multiple choices of  $r$  along the training process and plot the training curves.

**Batch Size:** We then investigate how the choices of batch size affect the generalization behaviors. We plot the results in Figure 4. As the figure shows, smaller batch size appears to excel in improving training and testing accuracy, while bigger batch size seems to stand out in closing the generalization gap. Also, it seems the generalization gap is closely related to the model’s tendency in capturing HFC: models trained with bigger epoch sizes are more invariant to HFC and introduce smaller differences in training accuracy and testing accuracy. The observed relation is intuitive because the smallest generalization gap will be achieved once the model behaves like a human (because it is the human who annotate the data).

The observation in Figure 4 also chips in the discussion in the previous section about “generalizable” features. Intuitively, with bigger epoch size, the features that can lead to steepest descent of the loss surface are more likely to be the “generalizable” patterns of the data, which are LFC.

**Heuristics:** We also test how different training methods react to LFC and HFC, including

- Dropout [26]: A heuristic that drops weights randomly during training. We apply dropout on fully-connected layers with  $p = 0.5$ .
- Mix-up [66]: A heuristic that linearly integrate samples and their labels during training. We apply it with standard hyperparameter  $\alpha = 0.5$ .
- BatchNorm [28]: A method that perform the normalization for each training mini-batch to accelerate Deep Network training process. It allows us to use a

much higher learning rate and reduce overfitting, similar with Dropout. We apply it with setting scale  $\gamma$  to 1 and offset  $\beta$  to 0.

- Adversarial Training [42]: A method that augments the data through adversarial examples generated by a threat model during training. It is widely considered as one of the most successful adversarial robustness (defense) method. Following the popular choice, we use PGD with  $\epsilon = 8/255$  ( $\epsilon = 0.03$ ) as the threat model.

We illustrate the results in Figure 5, where the first panel is the vanilla set-up, and then each one of the four heuristics are tested in the following four panels.

Dropout roughly behaves similarly to the vanilla set-up in our experiments. Mix-up delivers a similar prediction accuracy, however, it catches much more HFC, which is probably not surprising because the mix-up augmentation does not encourage anything about LFC explicitly, and the performance gain is likely due to attention towards HFC.

Adversarial training mostly behaves as expected: it reports a lower prediction accuracy, which is likely due to the trade-off between robustness and accuracy. It also reports a smaller generalization gap, which is likely as a result of picking up “generalizable” patterns, as verified by its invariance towards HFC (e.g.,  $r = 12$  or  $r = 16$ ). However, adversarial training seems to be sensitive to the HFC when  $r = 4$ , which is ignored even by the vanilla set-up.

The performance of BatchNorm is notable: compared to the vanilla set-up, BatchNorm picks more information in both LFC and HFC, especially when  $r = 4$  and  $r = 8$ . This BatchNorm’s tendency in capturing HFC is also related to observations that BatchNorm encourages adversarial vulnerability [18].

**Other Tests:** We have also tested other heuristics or methods by only changing along one dimension while the rest is fixed the same as the vanilla set-up in Section 4.

**Model architecture:** We tested LeNet [37], AlexNet [34], VGG [52], and ResNet [23]. The ResNet architecture seems advantageous toward previous inventions at different levels: it reports better vanilla test accuracy, smaller generalization gap (difference between training and testing accuracy), and a weaker tendency in capturing HFC.

**Optimizer:** We tested SGD, ADAM [32], AdaGrad [16], AdaDelta [64], and RMSprop. We notice that SGD seems to be the only one suffering from the tendency towards significantly capturing HFC, while the rest are on par within our experiments.

### 5.2. A hypothesis on Batch Normalization

Based on the observation, we hypothesized that one of BatchNorm’s advantage is, through normalization, to align the distributional disparities of different predictive signals. For example, HFC usually shows smaller magnitude than

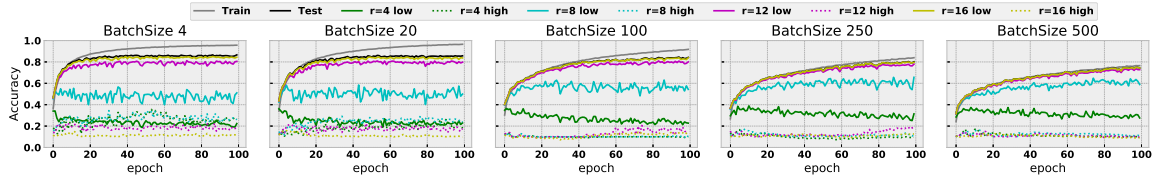


Figure 4. Plots of accuracy of different epoch sizes along the epochs for train, test data, as well as LFC and HFC with different radii.

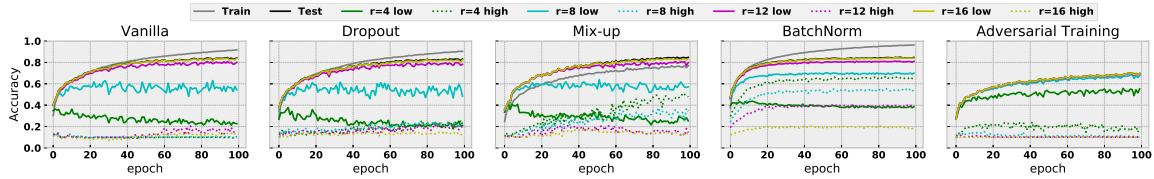


Figure 5. Plots of accuracy of different heuristics along the epochs for train, test data, as well as LFC and HFC with different radii.

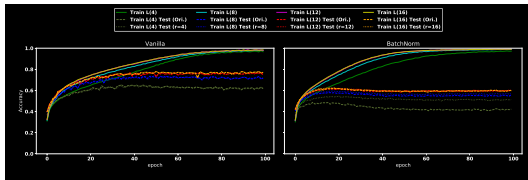


Figure 6. Comparison of models with vs. without BatchNorm trained with LFC data.

LFC, so a model trained without BatchNorm may not easily pick up these HFC. Therefore, the higher convergence speed may also be considered as a direct result of capturing different predictive signals simultaneously.

To verify this hypothesis, we compare the performance of models trained with vs. without BatchNorm over LFC data and plot the results in Figure 6.

As Figure 6 shows, when the model is trained with only LFC, BatchNorm does not always help improve the predictive performance, either tested by original data or by corresponding LFC data. Also, the smaller the radius is, the less the BatchNorm helps. Also, in our setting, BatchNorm does not generalize as well as the vanilla setting, which may raise a question about the benefit of BatchNorm.

However, BatchNorm still seems to at least boost the convergence of training accuracy. Interestingly, the acceleration is the smallest when  $r = 4$ . This observation further aligns with our hypothesis: if one of BatchNorm’s advantage is to encourage the model to capture different predictive signals, the performance gain of BatchNorm is the most limited when the model is trained with LFC when  $r = 4$ .

## 6. Adversarial Attack & Defense

As one may notice, our observation of HFC can be directly linked to the phenomenon of “adversarial example”: if the prediction relies on HFC, then perturbation of HFC will significantly alter the model’s response, but such perturbation may not be observed to human at all, creating the

unintuitive behavior of neural networks.

This section is devoted to study the relationship between adversarial robustness and model’s tendency in exploiting HFC. We first discuss the linkage between the “smoothness” of convolutional kernels and model’s sensitivity towards HFC (§6.1), which serves the tool for our follow-up analysis. With such tool, we first show that adversarially robust models tend to have “smooth” kernels (§6.2), and then demonstrate that directly smoothing the kernels (without training) can help improve the adversarial robustness towards some attacks (§6.3).

### 6.1. Kernel Smoothness vs. Image Frequency

As convolutional theorem [6] states, the convolution operation of images is equivalent to the element-wise multiplication of image frequency domain. Therefore, roughly, if a convolutional kernel has negligible weight at the high-end of the frequency domain, it will weigh HFC accordingly. This may only apply to the convolutional kernel at the first layer because the kernels at higher layer do not directly with the data, thus the relationship is not clear.

Therefore, we argue that, to push the model to ignore the HFC, one can consider to force the model to learn the convolutional kernels that have only negligible weights at the high-end of the frequency domain.

Intuitively (from signal processing knowledge), if the convolutional kernel is “smooth”, which means that there is no dramatics fluctuations between adjacent weights, the corresponding frequency domain will see a negligible amount of high-frequency signals. The connections have been mathematically proved [47, 55], but these proved exact relationships are out of the scope of this paper.

### 6.2. Robust Models Have Smooth Kernels

To understand the connection between “smoothness” and adversarial robustness, we visualize the convolutional kernels at the first layer of the models trained in the

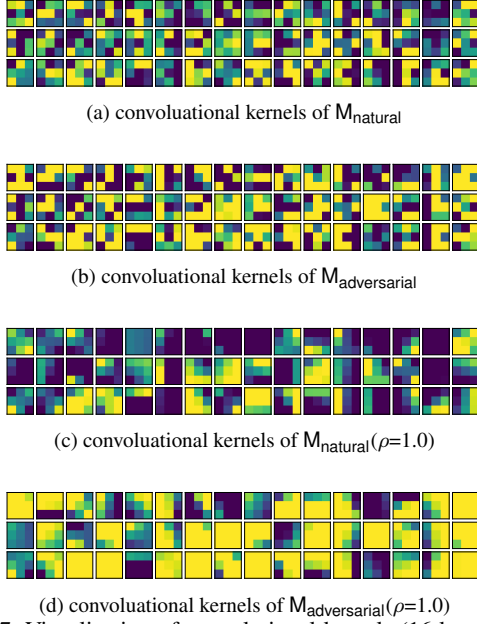


Figure 7. Visualization of convolutional kernels (16 kernels each channel  $\times$  3 channels at the first layer) of models.

vanilla manner ( $M_{\text{natural}}$ ) and trained with adversarial training ( $M_{\text{adversarial}}$ ) in Figure 7 (a) and (b).

Comparing Figure 7(a) and Figure 7(b), we can see that the kernels of  $M_{\text{adversarial}}$  tend to show a more smooth pattern, which can be observed by noticing that the adjacent weights of kernels of  $M_{\text{adversarial}}$  tend to share the same color. The visualization may not be very clear because the convolutional kernel is only  $[3 \times 3]$  in ResNet, the message is delivered more clearly in Appendix with other architecture when the first layer has kernel of the size  $[5 \times 5]$ .

### 6.3. Smoothing Kernels Improves Adversarial Robustness

The intuitive argument in §6.1 and empirical findings in §6.2 directly lead to a question of whether we can improve the adversarial robustness of models by smoothing the convolutional kernels at the first layer.

Following the discussion, we introduce an extremely simple method that appears to improve the adversarial robustness against FGSM [21] and PGD [36]. For a convolutional kernel  $\mathbf{w}$ , we use  $i$  and  $j$  to denote its column and row indices, thus  $\mathbf{w}_{i,j}$  denotes the value at  $i^{\text{th}}$  row and  $j^{\text{th}}$  column. If we use  $\mathcal{N}(i, j)$  to denote the set of the spatial neighbors of  $(i, j)$ , our method is simply:

$$\mathbf{w}_{i,j} = \mathbf{w}_{i,j} + \sum_{(h,k) \in \mathcal{N}(i,j)} \rho \mathbf{w}_{h,k}, \quad (3)$$

where  $\rho$  is a hyperparameter of our method. We fix  $\mathcal{N}(i, j)$  to have eight neighbors. If  $(i, j)$  is at the edge, then we

simply generate the out-of-boundary values by duplicating the values on the boundary.

In other words, we try to smooth the kernel through simply reducing the adjacent differences by mixing the adjacent values. The method barely has any computational load, but appears to improve the adversarial robustness of  $M_{\text{natural}}$  and  $M_{\text{adversarial}}$  towards FGSM and PGD, even when  $M_{\text{adversarial}}$  is trained with PGD as the threat model.

In Figure 7, we visualize the convolutional kernels with our method applied to  $M_{\text{natural}}$  and  $M_{\text{adversarial}}$  with  $\rho = 1.0$ , denoted as  $M_{\text{natural}}(\rho = 1.0)$  and  $M_{\text{adversarial}}(\rho = 1.0)$ , respectively. As the visualization shows, the resulting kernels tend to show a significantly smoother pattern.

We test the robustness of the models smoothed by our method against FGSM and PGD with different choices of  $\epsilon$ , where the maximum of perturbation is 1.0. As Table 2 shows, when our smoothing method is applied, the performance of clean accuracy directly plunges, but the performance of adversarial robustness improves. In particular, our method helps when the perturbation is allowed to be relatively large. For example, when  $\epsilon = 0.09$  (roughly 23/255),  $M_{\text{natural}}(\rho = 1.0)$  even outperforms  $M_{\text{adversarial}}$ . In general, our method can easily improve the adversarial robustness of  $M_{\text{natural}}$ , but can only improve upon  $M_{\text{adversarial}}$  in the case where  $\epsilon$  is larger, which is probably because the  $M_{\text{adversarial}}$  is trained with PGD( $\epsilon = 0.03$ ) as the threat model.

## 7. Beyond Image Classification

We aim to explore more than image classification tasks. We investigate in the object detection task. We use RetinaNet [40] with ResNet50 [23] + FPN [39] as the backbone. We train the model with COCO detection train set [41] and perform inference in its validation set, which includes 5000 images, and achieve an MAP of 35.6%.

Then we choose  $r = 128$  and maps the images into  $\mathbf{x}_l$  and  $\mathbf{x}_h$  and test with the same model and get 27.5% MAP with LFC and 10.7% MAP with HFC. The performance drop from 35.6% to 27.5% intrigues us so we further study whether the same drop should be expected from human.

### 7.1. Performance Drop on LFC

The performance drop from the  $\mathbf{x}$  to  $\mathbf{x}_l$  may be expected because  $\mathbf{x}_l$  may not have the rich information from the original images when HFC are dropped. In particular, different from image classification, HFC may play a significant role in depicting some objects, especially the smaller ones.

Figure 8 illustrates a few examples, where some objects are recognized worse in terms of MAP scores when the input images are replaced by the low-frequent counterparts. This disparity may be expected because the low-frequent images tend to be blurry and some objects may not be clear to a human either (as the left image represents).

|                                       | Clean        | FGSM              |                   |                   | PGD               |                   |                   |
|---------------------------------------|--------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|                                       |              | $\epsilon = 0.03$ | $\epsilon = 0.06$ | $\epsilon = 0.09$ | $\epsilon = 0.03$ | $\epsilon = 0.06$ | $\epsilon = 0.09$ |
| $M_{\text{natural}}$                  | <b>0.856</b> | 0.107             | 0.069             | 0.044             | 0.003             | 0.002             | 0.002             |
| $M_{\text{natural}}(\rho = 0.10)$     | 0.815        | 0.149             | 0.105             | 0.073             | 0.009             | 0.002             | 0.001             |
| $M_{\text{natural}}(\rho = 0.25)$     | 0.743        | 0.16              | 0.11              | 0.079             | 0.021             | 0.005             | 0.005             |
| $M_{\text{natural}}(\rho = 0.50)$     | 0.674        | 0.17              | 0.11              | 0.083             | 0.031             | 0.016             | 0.014             |
| $M_{\text{natural}}(\rho = 1.0)$      | 0.631        | <b>0.171</b>      | <b>0.14</b>       | <b>0.127</b>      | <b>0.086</b>      | <b>0.078</b>      | <b>0.078</b>      |
| $M_{\text{adversarial}}$              | <b>0.707</b> | <b>0.435</b>      | <b>0.232</b>      | 0.137             | <b>0.403</b>      | 0.138             | 0.038             |
| $M_{\text{adversarial}}(\rho = 0.10)$ | 0.691        | 0.412             | 0.192             | 0.109             | 0.379             | 0.13              | 0.047             |
| $M_{\text{adversarial}}(\rho = 0.25)$ | 0.667        | 0.385             | 0.176             | 0.097             | 0.352             | 0.116             | 0.04              |
| $M_{\text{adversarial}}(\rho = 0.50)$ | 0.653        | 0.365             | 0.18              | 0.106             | 0.334             | 0.121             | 0.062             |
| $M_{\text{adversarial}}(\rho = 1.0)$  | 0.638        | 0.356             | 0.223             | <b>0.186</b>      | 0.337             | <b>0.175</b>      | <b>0.131</b>      |

Table 2. Prediction performance of models against different adversarial attacks with different  $\epsilon$ .

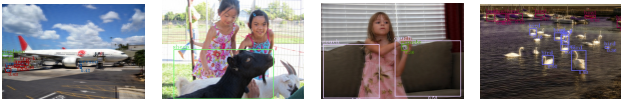


Figure 8. Some objects are recognized worse (lower MAP scores) when the experiments are repeated with low-frequency images. Marked objects are the ones that induce differences.



Figure 9. Some objects are recognized better (higher MAP scores) when the experiments are repeated with low-frequency images. Marked objects are the ones that induce differences.

## 7.2. Performance Gain on LFC

However, the disparity gets interesting when we inspect the performance gap in the opposite direction. We identified 1684 images that for each of these images, the some objects are recognized better (high MAP scores) in comparison to the original images.

The results are shown in Figure 9. There seems no apparent reasons why these objects are recognized better in low-frequency images, when inspected by human. These observations strengthen our argument in the perceptual disparity between CNN and human also exist in more advanced computer vision tasks other than image classification.

## 8. Discussion: Are HFC just Noises?

To answer this question, we experiment with another frequently used image denoising method: truncated singular value decomposition (SVD). We decompose the image and separate the image into one reconstructed with dominant singular values and one with trailing singular values. With this set-up, we find much fewer images supporting the story in Figure 2. Our observations suggest the signal CNN exploit is more than just random “noises”.

Code used in the paper: <https://github.com/HaohanWang/HFC>

## 9. Conclusion & Outlook

We investigated how image frequency spectrum affects the generalization behavior of CNN, leading to multiple interesting explanations of the generalization behaviors of neural networks from a new perspective: there are multiple signals in the data, and not all of them align with human’s visual preference. As the paper comprehensively covers many topics, we briefly reiterate the main lessons learned:

- CNN may capture HFC that are misaligned with human visual preference (§3), resulting in generalization mysteries such as the paradox of learning label-shuffled data (§4) and adversarial vulnerability (§6).
  - Heuristics that improve accuracy (*e.g.*, Mix-up and BatchNorm) may encourage capturing HFC (§5). Due to the trade-off between accuracy and robustness (§3), we may have to rethink the value of them.
  - Adversarially robust models tend to have smooth convolutional kernels, the reverse is not always true (§6).
  - Similar phenomena are noticed in the context of object detection (§7), with more conclusions yet to be drawn.
- Looking forward, we hope our work serves as a call towards future era of computer vision research, where the state-of-the-art is not as important as we thought.
- A single numeric on the leaderboard, while can significantly boost the research towards a direction, does not reliably reflect the alignment between models and human, while such an alignment is arguably paramount.
  - We hope our work will set forth towards a new testing scenario where the performance of low-frequency counterparts needs to be reported together with the performance of the original images.
  - Explicit inductive bias considering how a human views the data (*e.g.*, [58, 57]) may play a significant role in the future. In particular, neuroscience literature have shown that human tend to rely on low-frequency signals in recognizing objects [4, 5], which may inspire development of future methods.



## References

- [1] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3389–3398, 2018. 2
- [2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 2
- [3] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 233–242. JMLR. org, 2017. 2
- [4] Bhuvanesh Awasthi, Jason Friedman, and Mark A Williams. Faster, stronger, lateralized: low spatial frequency information supports face processing. *Neuropsychologia*, 49(13):3583–3590, 2011. 8
- [5] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617, 2004. 8
- [6] Ronald Newbold Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986. 6
- [7] Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018. 1
- [8] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. 3
- [9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 2
- [10] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018. 2
- [11] Jinghui Chen and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint arXiv:1806.06763*, 2018. 2
- [12] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2
- [13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 4
- [14] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 4
- [15] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1019–1028. JMLR. org, 2017. 2
- [16] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 5
- [17] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017. 2
- [18] Angus Galloway, Anna Golubeva, Thomas Tanay, Medhat Moussa, and Graham W Taylor. Batch normalization is a cause of adversarial vulnerability. *arXiv preprint arXiv:1905.02161*, 2019. 5
- [19] 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. 1
- [20] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. PMLR, 2010. 4
- [21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples (2014). In *International Conference on Learning Representations*, 2015. 1, 2, 3, 7
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4
- [23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 7
- [24] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th {USENIX} Workshop on Offensive Technologies ({WOOT} 17)*, 2017. 2
- [25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2
- [26] Geoffrey E Hinton, Alexander Krizhevsky, Ilya Sutskever, and Nitish Srivastva. System and method for addressing overfitting in a neural network, Aug. 2 2016. US Patent 9,406,017. 5
- [27] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019. 1
- [28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5
- [29] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017. 1
- [30] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017. 1
- [31] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and

- sharp minima. In *International Conference on Learning Representations*, 2017. 1
- [32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5
- [33] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Cite-seer, 2009. 4
- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5
- [35] David Krueger, Nicolas Ballas, Stanislaw Jastrzebski, Devansh Arpit, Maxinder S Kanwal, Tegan Maharaj, Emmanuel Bengio, Asja Fischer, and Aaron Courville. Deep nets don’t learn via memorization. 2017. 2
- [36] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Workshop of International Conference on Learning Representations*, 2017. 2, 7
- [37] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 5
- [38] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, 2017. 2
- [39] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7
- [40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7
- [41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7
- [42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 5
- [43] Saeed Mahloujifar, Dimitrios I. Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure, 2018. 1
- [44] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 135–147. ACM, 2017. 2
- [45] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017. 2
- [46] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017. 1
- [47] Sergei Sergeevich Platonov. The fourier transform of functions satisfying the lipschitz condition on rank 1 symmetric spaces. *Siberian Mathematical Journal*, 46(6):1108–1118, 2005. 6
- [48] Nasim Rahaman, Devansh Arpit, Aristide Baratin, Felix Draxler, Min Lin, Fred A Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of deep neural networks. *arXiv preprint arXiv:1806.08734*, 2018. 4
- [49] Andras Rozsa, Manuel Günther, and Terrance E Boulton. Are accuracy and robustness correlated. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 227–232. IEEE, 2016. 2
- [50] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data, 2018. 1
- [51] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance, 2019. 1
- [52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5
- [53] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 2019. 2
- [54] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2, 3
- [55] Edward C Titchmarsh. Introduction to the theory of fourier integrals. 1948. 6
- [56] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. 2
- [57] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power, 2019. 8
- [58] Haohan Wang, Zexue He, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019. 1, 8
- [59] Haohan Wang, Aaksha Meghawati, Louis-Philippe Morency, and Eric P Xing. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 949–954. IEEE, 2017. 1
- [60] Haohan Wang and Bhiksha Raj. On the origin of deep learning. *arXiv preprint arXiv:1702.07800*, 2017. 5
- [61] Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017. 2
- [62] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples

- with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. [2](#)
- [63] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. [4](#)
- [64] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. [5](#)
- [65] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. [1](#), [2](#), [3](#), [4](#)
- [66] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. [5](#)
- [67] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019. [2](#)