Impact of Spatial Frequency Based Constraints on Adversarial Robustness

Rémi Bernhard[1,2], Pierre-Alain Moellic[1,2], Martial Mermillod[3], Yannick Bourrier[3],
Romain Cohendet[4], Miguel Solinas[4], Marina Reyboz[4]

[1]CEA Tech, Centre CMP, Equipe Commune CEA Tech - Mines Saint-Etienne, F-13541
Gardanne, France
[2]Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France
[3] LPNC, CNRS, Université Grenoble Alpes, Université Savoie Mont Blanc, Grenoble, France
[4] Univ. Grenoble Alpes, CEA, List, F-38000 Grenoble, France

**remi.bernhard@cea.fr, pierre-alain.moellic@cea.fr, romain.cohendet@cea.fr,
miguelangel.solinas@cea.fr, marina.reyboz @cea.fr
martial.mermillod@univ-grenoble-alpes.fr, yannick.bourrier@univ-grenoble-alpes.fr**

April 27, 2021

1 Abstract

Adversarial examples mainly exploit changes to input pixels to which humans are not sensitive to, and arise from the fact that models make decisions based on uninterpretable features. Interestingly, cognitive science reports that the process of interpretability for human classification decision relies predominantly on low spatial frequency components. In this paper, we investigate the robustness to adversarial perturbations of models enforced during training to leverage information corresponding to different spatial frequency ranges. We show that it is tightly linked to the spatial frequency characteristics of the data at stake. Indeed, depending on the data set, the same constraint may results in very different level of robustness (up to 0.41 adversarial accuracy difference). To explain this phenomenon, we conduct several experiments to enlighten influential factors such as the level of sensitivity to high frequencies, and the transferability of adversarial perturbations between original and low-pass filtered inputs.

2 Introduction

Neural networks based models have been shown to reach impressive performances on challenging tasks, while being vulnerable to adversarial examples, i.e. maliciously crafted perturbations added to clean examples to fool a model at inference [1]. This phenomenon has shed the light on the fact that, to perform a specific task, machine learning models rely on different features or different feature processing from those humans rely on to make their decisions [2–4].

To make a model robust against an adversary in the white-box setting, many defenses have been developed including proactive [5–8] [5, 8] and reactive strategies [9, 10]. Few works underline the critical link between robustness and the interpretability of the features a model relies on: it has been shown that models trained with adversarial training [5] or randomized smoothing [8] exhibit interpretable gradients [11–13] and, inversely, a recent effort [14] demonstrates that a model trained to have explainable jacobian matrices presents adversarial robustness.

Furthermore, Zhang *et al.* [15] experimentally highlight the importance of low spatial frequency (hereafter, LSF) information, such as shape, for adversarially trained models, in opposition to concepts associated with high spatial frequency (hereafter, HSF) information. Experimental evidence in neural computation and cognitive psychology suggests the importance of LSF to perform efficient classification [16–18]. Therefore, a natural hypothesis would be that a model trained specifically to rely more on LSF information might present an improved adversarial robustness. This hypothesis has already been indirectly exploited by taking advantage of preprocessing defense schemes based on HSF components filtering [19–21]. Other defenses aim at training a model exploiting more human interpretable information, such as [22] by giving higher importance to information present in higher bit planes. However, these methods stay agnostic of the intrinsic spatial frequency characteristics of the data.

The objective of this work is twofold. First, we experimentally question some preconceived hypothesis related to adversarial examples, more particularly ones considering adversarial perturbations as a pure HSF phenomenon with
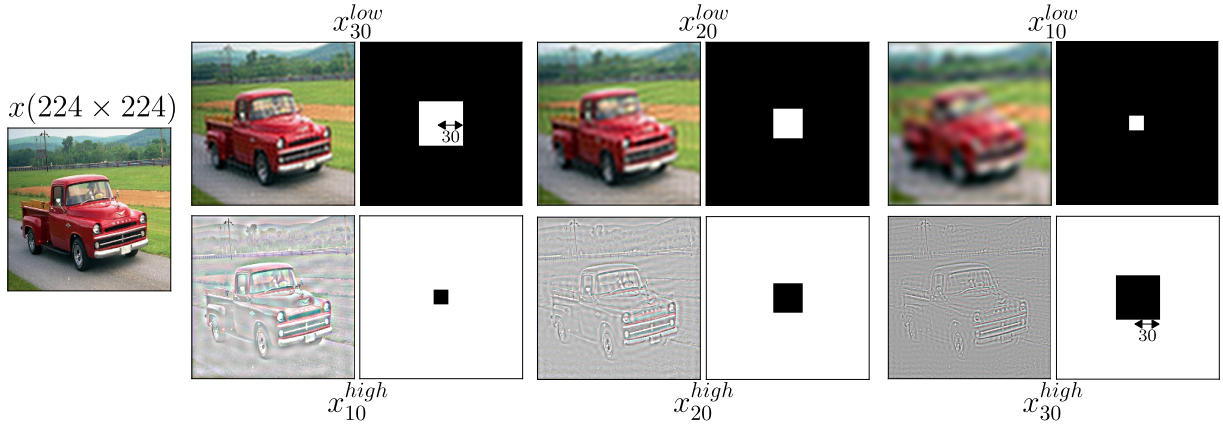
Figure 1: Original image (left). Low-pass filtered images with corresponding mask (first row). High-pass filtered images (magnified for visualization) with corresponding mask (bottom row). For the Fourier domain masks, white denotes value 1 and black value 0. **For LSF, low filtering intensity means restricted low-pass filtering. For HSF, high filtering intensity means restricted high-pass filtering.**

data-agnostic spatial frequency characteristics. Second, we aim at investigating the link between spatial frequency features of the information that a model uses to perform predictions *and* the robustness against adversarial perturbations offered by spatial frequency-based constraints. Our key contributions are:

- We show that a frequency-based regularization induces very different levels of robustness according to the frequency features of the data. As an example, a low-frequency constrained model on CIFAR10 (that covers a broad frequency spectrum) has no robustness, while reaching a 41 % true robustness for SVHN against the l_∞ PGD attack.

- By analyzing the sensitivity of a model trained naturally (i.e. without spatial frequency-based procedures and hereafter noted as *regular* model) as well as adversarial transferability properties, we observe that enforcing a model to rely on LSF information is not a necessary condition to bring adversarial robustness.

- We notice that, depending on the data set complexity, some models spread over the whole frequency spectrum, and show that constraints spanning different frequency ranges can help improving robustness.

- We discuss combination with adversarial training [5] for future overall defense strategy.

The paper is organized as follows. After positioning our work in relation to the state-of-the-art in Section 3, we analyze, in Section 5, frequency properties and sensitivity of features learned by a model. Notably, we study to which extent features learned by a model are focused on low or high spatial frequency concepts, and the sensitivity of models to frequency constrained noise. In Section 6, we set forth interesting links between transferability of adversarial perturbations and frequency properties of the information models make use of. In Section 7, we design loss functions to force a model to extract features in particular frequency ranges, and observe the potential effects on adversarial robustness. We link these effects with analysis performed in Sections 5 and 6. We also show promising future work direction by binding our findings with Adversarial Training.

3   Related work

Recent efforts [4, 23] experimentally demonstrate that regular models predominantly exploit non-interpretable HSF components, and that robust models tend to use more concepts, such as shape, associated to LSF [15, 24], in a way similar to that of humans [16–18]. Therefore, a common belief is that the adversarial vulnerability of a model comes from the utilization of HSF components [25]. However, Yin *et al.* [4] show that adversarial perturbations cannot be viewed only as a HSF phenomenon: adversarially trained models, more sensitive to LSF, stay vulnerable to an adversary that optimally exploits this part of the frequency spectrum. Notably, Sharma *et al.* [26] make use of LSF constraints to craft adversarial examples against adversarially trained models on ImageNet and use fewer iterations than classical adversarial attacks.

As adversarial examples are the consequence of models relying on brittle and non-interpretable features [3], some papers exploit the idea of enforcing a model to use features to which humans are sensitive. During training, Yin *et al.* [4] add LSF noise and demonstrate that it does not necessarily improve robustness to LSF perturbations. As a explanation, the authors hypothesize that as natural images are more LSF concentrated, it is harder for a model to become invariant to these spatial frequencies and, then, to LSF perturbations. Before, Geirhos *et al.* [24] interestingly

took advantage of stylized images that keep only shape information within a data augmentation process to improve the robustness against common perturbations. Even if not related to frequency concerns, Addepalli *et al.* show in [22] that constraining a model to rely on the information contained in the higher bit planes only (as humans make decisions based on the information of large magnitude) has a positive impact on robustness.

4 Preliminaries

4.1 Notations

A neural network model M_θ, with parameters θ, classifies an input $x \in \mathbb{R}^d$ to a label $M_\theta(x) \in \{1 \dots C\}$. $L^E(\theta, x, y)$ denotes the cross-entropy loss for M_θ and (x, y) an input with its corresponding ground-truth label. The pre-softmax function of M_θ (the logits) is denoted as $f : \mathbb{R}^d \to \mathbb{R}^C$. We denote x_i^{low} and x_i^{high} respectively the low-pass and high-pass filtered versions of x at intensity i (see Section 4.2). The *LSF task* (resp. *HSF task*) refers to the classification task where inputs have been low-pass (resp. high-pass) filtered at some intensity, i.e. where input-label pairs correspond to (x_i^{low}, y) (resp. (x_i^{high}, y)) for some i. M_i^{low} (resp. M_i^{high}) denotes a model trained for the LSF (resp. HSF) task. The accuracy is denoted by Acc, and the adversarial accuracy (Acc_{adv}) denotes the accuracy of a model on a set of adversarial examples.

4.2 Filtering with the Fourier transform

Low or high-pass filtering is performed by deleting the undesired spatial frequencies in the Fourier domain thanks to a boolean mask $\Omega \in \{0, 1\}^{n \times n}$, similarly to [4, 23, 25, 26]. We note $\mathscr{F}$ and $\mathscr{F}^{-1}$ the Discrete Fourier Transform (DFT), and its inverse function, respectively. For a gray-scale image $a \in \mathbb{R}^{n \times n}$, $\mathscr{F}_c(a) \in \mathbb{C}^{n \times n}$ denotes the centered variant of $\mathscr{F}(a) \in \mathbb{C}^{n \times n}$ (i.e. coeffients for low frequencies are located at the center, and those for high-frequencies at the corners). The filtered image is then obtained classically as $x^{freq} = \mathscr{F}^{-1}(\mathscr{F}_c(a) \odot \Omega)$, with $\odot$ the Hadamard product. For a color image, the procedure is applied to each of the channels. We note x_i^{low} for Ω corresponding to 1's only in the $2i \times 2i$ square in the middle, and x_i^{high} for Ω corresponding to 1's only outside the $2i \times 2i$ square. In consequence, for low-pass (resp. high-pass) filtering, the smaller (resp. the higher) the intensity i, the stronger the filter is. We illustrate this process in Figure 1.

4.3 Data sets and models

We consider CIFAR10 [27], SVHN [28] and a custom-built data set, named *Small ImageNet*, built by extracting 10 meta classes from the ImageNet ILSVRC2012 benchmark. For each meta class, we extracted 3000 images from the original training set and 300 images from the non-blacklisted validation set. All input images are scaled to $[0, 1]$. CIFAR10 and SVHN, both involving color images of size 32×32 allow to make comparisons and draw conclusions about phenomenon observed. *Small ImageNet*, composed of 224×224 color images, enables to extrapolate results on images with higher definition. For SVHN we use a model inspired from VGG [29], for CIFAR10 we use a WideResNet28-8 model [30], and for *Small Imagenet* we consider a MobileNetV2 model [31]. All details about *Small Imagenet* and the training hyper-parameters are presented in the code repository of this work[1].

5 Frequency properties of data and models

In this section we aim at gaining insight on the way information learned by a classifier trained to solve the regular classification task contains information for the LSF and HSF tasks. These experiments allow to gain intuition on the way models leverage information with respect to data sets frequencies, and will help to better understand robustness dissimilarities which occur between different models trained with the same frequency-based constraints (later in Section 7).

5.1 Impact of filtered data sets.

We begin by evaluating CIFAR10 and SVHN regular models on low-pass and high-pass filtered images. Results are presented in Figure 2. For CIFAR10, we notice that the accuracy of a model decreases much slower than SVHN when evaluated on more and more high-pass filtered images. We obtain an opposite result for SVHN with low-pass filtered data. We can therefore assume that the informative features learned by the regular model are more focused on the LSF task for SVHN and are more spread between LSF and HSF tasks for CIFAR10. This analysis is consistent with the intrinsic frequency features of the data sets that is revealed by a classical Fourier analysis (presented in Figure 3, top row) that actually shows a quite narrow spectrum for SVHN (towards LSF) and a spread spectrum for CIFAR10.

---

[1]`https://gitlab.emse.fr/remi.bernhard/Frequency_and_Robustness/`
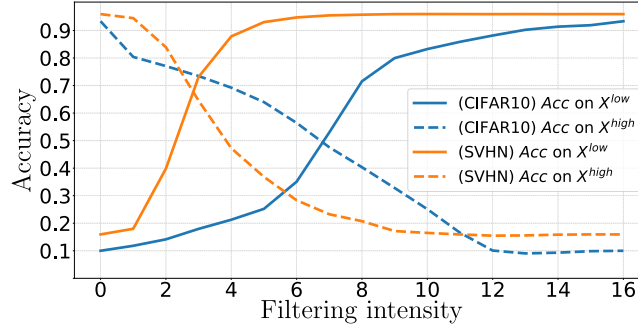
Figure 2: CIFAR10 and SVHN. Accuracy of a regular model on low-pass and high-pass filtered data set, for different filtering intensities.
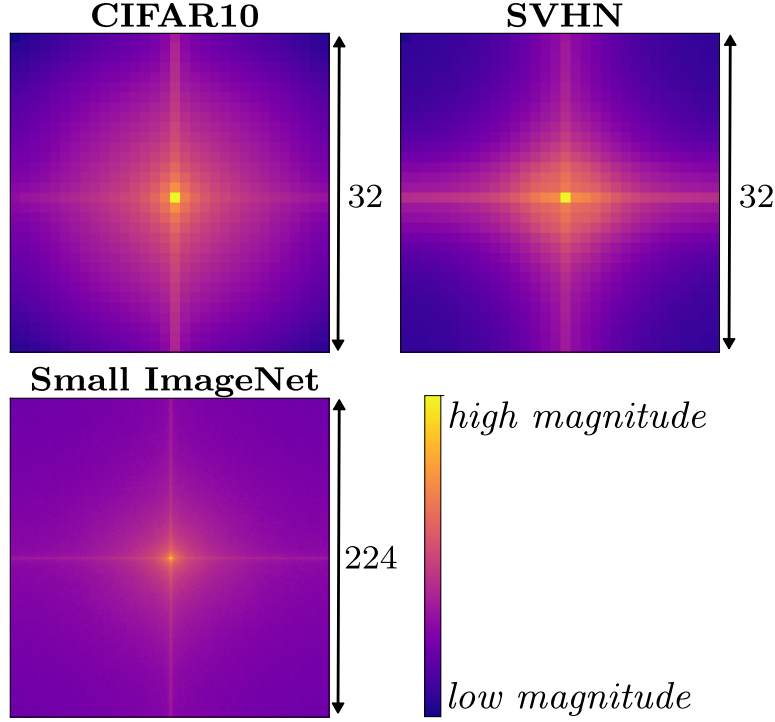

Figure 3: CIFAR10, SVHN and Small Imagenet data sets. Magnitude of the Fourier spectrum for clean images. Low frequencies are at the center, and high frequencies at the corners.

The accuracy of models M_i^{low} and M_i^{high} for various i, presented in Figure 4, further highlights this phenomenon. The accuracy reached by models M_i^{low} decreases much slower for SVHN than for CIFAR10, as the filtering intensity increases (and the inverse phenomenon is observable for high-pass filtering). This agrees with previous results (Figure 2), i.e. the useful information for the classification task is more distributed in the frequency spectrum for CIFAR10 compared to SVHN for which the information is predominantly concentrated in the low frequencies.

5.2 Sensitivity to high spatial frequency noise.

We investigate the sensitivity of regular models to perturbations in specific frequencies. For that purpose, we use a similar procedure as in [4]: the sensitivity is measured as the error rate of the model on a set of examples perturbed with noise located only in those spatial frequencies. More precisely, for a clean input image $x \in \mathbb{R}^{n \times n \times c}$, each channel is perturbed independently by the addition of a noise $rvU_{i,j}$, where $U_{i,j}$ is a Fourier basis matrix for coordinates (i, j) in the Fourier domain, r is chosen randomly in $\{-1, 1\}$, and v controls the magnitude of the perturbation. We then measure the error rate of a model on $1,000$ well-classified test set examples perturbed with this type of noise, as a function of (i, j). We focus our analysis on CIFAR10 and SVHN, two data sets with the same image size, and we experimentally set $v = 4$. Results are presented in Figure 5 (top row). We notice that a CIFAR10 regular model is more sensitive to LSF and HSF noise, while a SVHN regular model is more sensitive to LSF and not to high and very high frequencies.
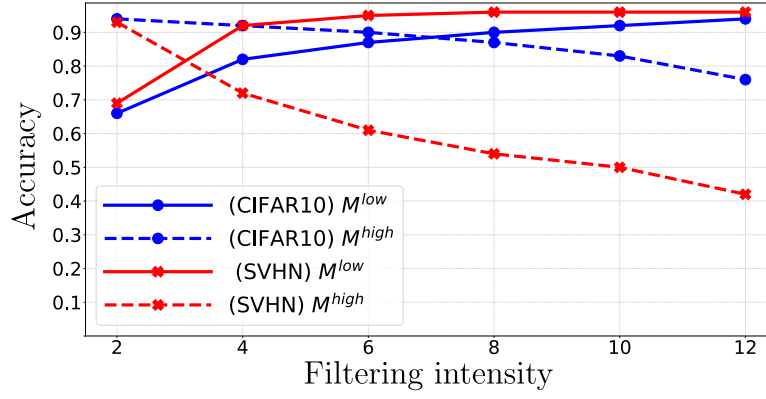
Figure 4: CIFAR10 and SVHN. Test set accuracy of models trained on filtered data sets. (See Figure 1 for filtering intensity effect)
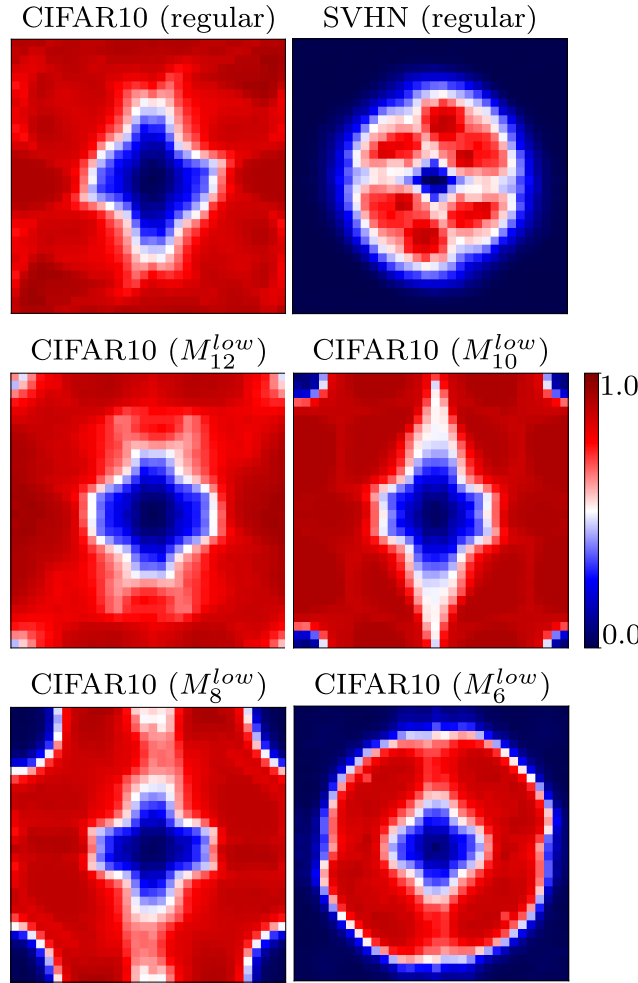

Figure 5: CIFAR10 and SVHN. Error rate of models on data perturbed with Fourier constrained noise. Value at (i, j) measures the error rate on data perturbed with noise along Fourier basis matrice $U_{i,j}$. Low frequencies are located at the center, and high frequencies at the corner. High values (red color) denote a high sensibility, and low values (blue color) denote a low sensibility.

For CIFAR10, we also investigate the sensitivity of models M_i^{low}. Results are presented in Figure 5 (middle and bottom rows). Interestingly, we notice that the models M_i^{low} become less sensitive to HSF as i decreases (i.e. as the low-pass filtering intensity increases). Therefore, for CIFAR10, training a model on low-pass filtered images brings robustness against HSF perturbations. However, this robustness would be useless if adversarial perturbations were to rely on a broader spectrum than only HSF. We investigate this in the following section.

6 Transferability analysis

The transferability of adversarial perturbations between two models has been explained by shared non-robust useful features [3], that is features sensitive to adversarial perturbations and exploited for the prediction by both models. Therefore, to gain a better understanding of the nature of features at stake as well as frequency properties of adversarial perturbations, we study the transferability of adversarial examples between a regular model and models trained for the LSF or HSF task (i.e. M^{low} and M^{high}). For these M^{low} and M^{high} models, a pre-processing layer is added before the model to evaluate it on non-filtered inputs. We use the l_∞ DIM attack [32] (a state-of-the-art gradient-based attack tuned for transferability), with 40 iterations, a probability $p = 0.8$ and a l_∞ perturbation budget of 0.03.

We first remind that for models M_i^{low}, a low filtering intensity i means that the model is trained on a strict low-pass filtered data set (see Figure 1) and as the filtering intensity i increases, more and more HSF are considered. Thus, for example, adversarial examples crafted on M_2^{low} exploit non-robust features learned with only LSF information and adversarial examples crafted on M_{12}^{low} will have the possibility to take advantage of non-robust features determined on a broader range of frequencies. On the contrary, for models M_i^{high}, a high filtering intensity i means that only HSF are kept (see Figure 1), and a low intensity represents a larger spectrum, gathering more LSF. Here, adversarial examples crafted on M_{12}^{high} will exploit non-robust features learned exclusively on HSF information and adversarial examples crafted on M_2^{high} will rely on non-robust features from a broader spectrum. We illustrate this in Figure 6 *(a)* and *(b)*, by presenting the magnitude of the Fourier spectrum of adversarial perturbation crafted on the regular model and on M_6^{low}. We observe that the perturbation is quite uniformly distributed along the spectrum (confirming a result from [4]) for the regular model and predominantly focused on the low frequencies for M_6^{low}. This is an important point, as it shows that adversarial examples are not HSF phenomena but may rely on a large range of spatial frequencies to efficiently fool a model.

Furthermore, analyzing the transferability results of Figure 7, we can provide interesting information about the nature of the robust and non robust features of a model as well as the properties of adversarial perturbations. Importantly, we observe similar behaviors between CIFAR10, SVHN and Small ImageNet.

A first conclusion comes from the two-way transferability between the regular model M (note *Base*) and models M_i^{low} (blue curves). Indeed, we notice that the stronger the low-pass filtering (smaller i), the lower the transferability of adversarial perturbations. This indicates that the regular classification task and the LSF task share predominantly robust useful features.

Secondly, an important outcome arises from the dissimilarity between the two orange curves for each data set. On one hand, the solid orange curve attests the impact of non-robust features exploiting HSF, highlighted by an almost constant success of adversarial examples crafted from the base model against models M_i^{high} (about 0.1, 0.2 and 0.5 of adversarial accuracy for CIFAR10, SVHN and Small ImageNet respectively). On another hand, the dotted orange curve indicates that, as the high-pass filtering becomes more restrictive (i.e. as the i value increases), the transferability of adversarial examples crafted on M_i^{high} to the regular model decreases. These observations support the claim that, to be efficient, adversarial perturbations must exploit a wide part of the spectrum, and therefore cannot be only focused on HSF. This is particularly the case for Small ImageNet, as *i)* the accuracy from a regular to a M^{high} model is higher than for CIFAR10 and SVHN, and *ii)* the transferability of adversarial examples crafted on M_i^{high} is already poor for $i = 5$.

We can summarize our conclusions as follows:

- Robustness is strongly related to features that rely on LSF information.

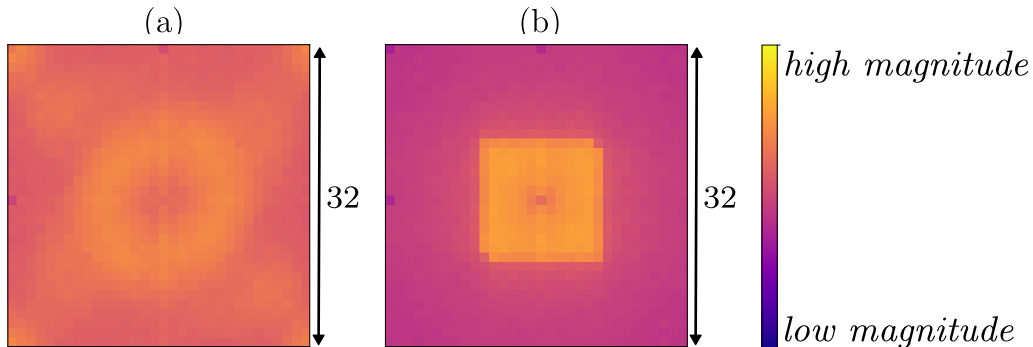- Adversarial perturbation are not efficient when focused only on HSF.

Figure 6: CIFAR10. Magnitude of the Fourier spectrum for the adversarial perturbation (a) Regular model, (b) Model trained on low-pass filtered dataset at intensity 6
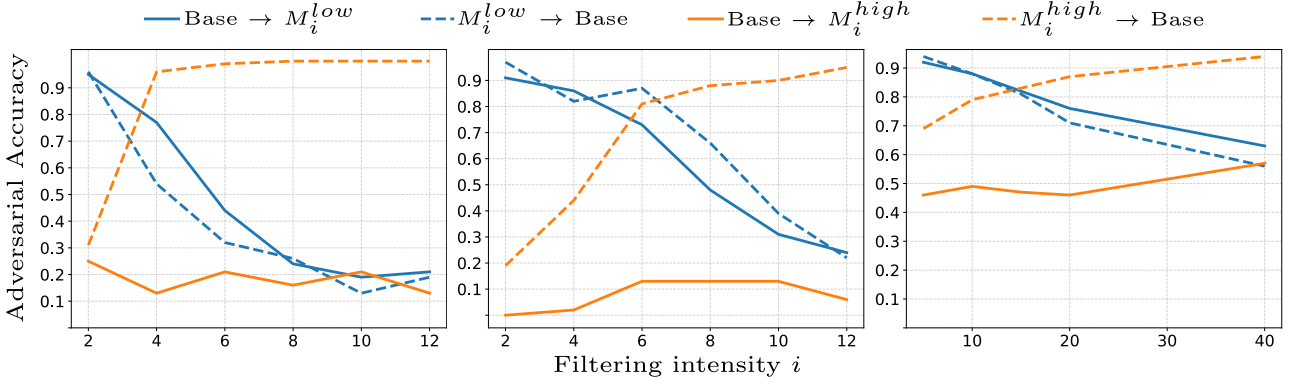
Figure 7: SVHN (left), CIFAR10 (middle), Small ImageNet (right). Transferability analysis between base model and models trained on a filtered data set. (See Figure 1 for filtering intensity effect)

In the next section, we propose to look deeper into the impact of frequency-based constraints when training a regular model in order to investigate if it can help increasing its robustness.

7 Adversarial robustness of frequency-constrained models

Following the precedent observations, we investigate loss functions designed to make a model to leverage informative features for both the regular task and the LSF and/or HSF task and evaluate how it impacts the adversarial robustness. The adversarial robustness is evaluated with the adversarial accuracy (Acc_{add}) considering an attacker in the white-box setting, under the common l_∞ threat model. Adversarial examples are crafted with the l_∞ PGD attack [5], with a perturbation budget of $\epsilon = 0.03$. To provide the more accurate evaluation of robustness as possible, notably to assess that no gradient masking occurs, we follow state-of-the-art guidelines from [33]. Detailed architectures and setups are presented in the code repository of this work.

7.1 Frequency-based regularization

To enforce a model M_θ to rely on information relative to the LSF or HSF task, we define the following loss function L^{freq}, which acts on the logits.

$$
\begin{aligned}
L_{i,j}^{freq}(\theta, x, y) = L^E(\theta, x, y) &+ \lambda_1 \left\| f(x) - f(x_i^{low}) \right\|_2^2 \\
&+ \lambda_2 \left\| f(x) - f(x_j^{high}) \right\|_2^2
\end{aligned}
\tag{1}
$$

For readability, when $\lambda_1 = 0$ (i.e. constraint is only focused on the HSF task) the loss is simply noted as L_i^{high}, and for $\lambda_2 = 0$ as L_i^{low}. During training, the cross-entropy part L^E makes the model to learn useful features to solve the regular classification task. The second part (moderated by λ_1 and λ_2) constraints the model to extract useful features coherently to the LSF and HSF tasks.

For CIFAR10 and SVHN, the loss functions considered are L_i^{low} and L_i^{high} for $i \in \{2, 4, 6, 8, 10\}$, and $L_{i,j}^{freq}$ for $(i, j) \in \{(10, 4), (4, 12), (6, 3), (6, 10), (8, 6), (6, 8)\}$. For Small ImageNet, as each model training is costly, we consider the loss functions L_i^{low}, L_i^{high} and $L_{i,j}^{freq}$ for representatives values in $\{10, 20, 40, 60\}$. For conciseness purpose when presenting results, the subscript '$*$' means that equal results are reached whatever the intensity of the filtering.

7.2 Do the intrinsic frequency properties of the data bias the level of adversarial robustness?

Results for L^{low} and L^{high}, presented in Table 1, allow for a first important observation. Indeed, we see that the same constraint induces very different effects on the robustness, depending on the data set at stake. For CIFAR10 and Small ImageNet we observe no robustness when considering separate losses L_*^{low} or L_*^{high}. On the contrary, models trained on SVHN with L^{low} present an interesting level of adversarial robustness depending on the intensity level. This observation appears as coherent with different frequency properties of the data sets highlighted in Section 5: the information learned by models trained on CIFAR10 and Small ImageNet are spread over the whole frequency spectrum, and – on the contrary – focused on LSF for SVHN. To further investigate this link, we proceed to check if a CIFAR10 model trained with L^{low} would still reach some robustness on low-pass filtered data (i.e. train with

SVHN

	Acc	Constraints	Acc_{adv}	Constraints
Best	0.96	$L_{i>2}^{low}$	0.41	L_6^{low}
Worst	0.93	L_2^{low}	0.0	L_{10}^{low}

CIFAR10

	Acc	Constraints	Acc_{adv}	Constraints
Best	0.92	L_2^{low}, L_{10}^{high}	0.0	L_*^{low}, L_*^{high}
Worst	0.74	L_2^{low}	0.0	L_*^{low}, L_*^{high}

Small Imagenet

	Acc	Constraints	Acc_{adv}	Constraints
Best	0.92	L_{40}^{low}	0.0	L_*^{low}, L_*^{high}
Worst	0.88	L_{40}^{high}	0.0	L_*^{low}, L_*^{high}

Table 1: CIFAR10, SVHN and Small ImageNet. Best and worst accuracy on clean (Acc) and adversarial examples (Acc_{adv}) of models trained with the constrained loss L_i^{low} or L_i^{high}.

SVHN

	Acc	Constraints	Acc_{adv}	Constraints
Best	0.96	$L_{10,4}^{freq}$, $L_{6,*}^{freq}$, $L_{8,6}^{freq}$	0.30	$L_{6,3}^{freq}$
Worst	0.95	$L_{4,12}^{freq}$	0.0	$L_{10,4}^{freq}$, $L_{6,10}^{freq}$, $L_{6,8}^{freq}$

CIFAR10

	Acc	Constraints	Acc_{adv}	Constraints
Best	0.95	$L_{5,3}^{freq}, L_{8,6}^{freq}$	0.12	$L_{5,3}^{freq}, L_*^{high}$
Worst	0.92	$L_{4,12}^{freq}$	0.03	$L_{4,12}^{freq}, L_{6,8}^{freq}$

Small ImageNet

	Acc	Constraints	Acc_{adv}	Constraints
Best	0.92	$L_{60,*}^{freq}$	0.36	$L_{40,20}^{freq}$
Worst	0.89	$L_{10,60}^{freq}$	0.0	$L_{60,*}^{freq}$

Table 2: CIFAR10, SVHN and Small ImageNet. Best and worst accuracy on clean (Acc) and adversarial examples (Acc_{adv}) of models trained with the constrained loss $L_{i,j}^{freq}$.

$L_i^{low}(\theta, x^{low}, y)$). In other words, we try to exclude the high spatial frequencies, which are predominantly non robust, and that we assume to explain this complete lack of robustness. To that purpose, we train models with L_i^{low} ($i = 6, 8, 10, 12$) on low-pass filtered versions of CIFAR10 ($i = 2, 4, 10$). Thus, we ensure that the model learns features strictly focused in the LSF. Interestingly, we measure a slight but true and non-negligible robustness for some models with loss L^{low} (up to an adversarial accuracy of 0.11).

As robustness is noticed for SVHN with L^{low} or for CIFAR10 with L^{low} on low-pass filtered data, a first conclusion is that a model relying predominantly on useful features of the LSF task can be made more robust with low-frequency based constraint (L^{low}). Moreover, these models also share a non-sensitivity to high and very high frequencies (cf. Section 5), which highlights another influential factor on the robustness of a model trained with the loss L^{low}.

However, is a frequency-based regularization suitable for complex data sets such as CIFAR10 or Small ImageNet? In Table 2, we present results with a constraint spanning a wider spectrum ($\lambda_1 > 0$, $\lambda_2 > 0$). For SVHN, as expected, the combined constraints does not enable to reach better robustness compared to the loss L^{low}, as the combination of the two constraints is not compatible with the intrinsic frequency properties of the data set: information is predominantly concentrated in the low frequencies. For Small ImageNet and CIFAR10 the combined constraint forces the model to learn informative features for the LSF and HSF tasks, which are mainly robust as features informative of the LSF task are mostly robust, as shown in Section 6 for CIFAR10. Interestingly, stronger level of robustness is observed for Small ImageNet and we hypothesize that the impact of the combined constraint is all the more efficient that the frequency spectrum is wider.

Loss L_{AT}: $Acc = 0.84$, $Ac_{adv} = 0.55$				
Loss $L_i^{AT,low}$				
$i =$	2	4	8	10
Acc	0.80	0.79	0.81	0.82
Acc_{adv}	0.59	0.60	0.57	0.56
Loss $L_i^{AT,high}$				
$i =$	4	**6**	8	10
Acc	0.84	**0.84**	0.85	0.83
Acc_{adv}	0.58	**0.6**	0.55	0.60
Loss $L_{i,j}^{AT,freq}$				
$(i,j) =$	(10,4)	(2,10)	(4,8)	(8,6)
Acc	0.84	0.8	0.81	0.82
Acc_{adv}	0.58	0.61	0.58	0.6

Table 3: CIFAR10. Accuracy on clean (Acc) and adversarial (Acc_{adv}) examples of models trained with loss functions $L_i^{AT,low}$ ($\lambda_2 = 0$), $L_i^{AT,high}$ ($\lambda_1 = 0$) and $L_{i,j}^{AT,freq}$

Loss L_{AT}: $Acc = 0.93$, $Ac_{adv} = 0.54$				
Loss $L_i^{AT,low}$				
$i =$	2	4	8	10
Acc	0.87	0.9	0.90	0.90
Acc_{adv}	0.59	0.59	0.60	0.61
Loss $L_i^{AT,high}$				
$i =$	4	6	8	10
Acc	0.87	0.8	0.88	0.87
Acc_{adv}	0.56	0.58	0.57	0.57
Loss $L_{i,j}^{AT,freq}$				
$(i,j) =$	**(10,4)**	(2,10)	(4,8)	(8,6)
Acc	**0.92**	0.92	0.91	0.91
Acc_{adv}	**0.66**	0.6	0.64	0.66

Table 4: SVHN. Accuracy on clean (Acc) and adversarial (Acc_{adv}) examples of models trained with loss functions $L_i^{AT,low}$ ($\lambda_2 = 0$), $L_i^{AT,high}$ ($\lambda_1 = 0$) and $L_{i,j}^{AT,freq}$

7.3 Is frequency-based regularization compatible with adversarial training ?

Motivated by the true robustness observed when training some models with loss functions L^{low}, L^{high} or L^{freq}, we study the combination of these losses with Adversarial Training (hereafter AT) [5], a common and widely used approach when defending against an adversary in the white-box setting. This study is of particular interest as AT is shown to be related to frequency concern as it makes a model rely more on low-frequency concepts and being less sensitive to perturbations in the high-frequencies [15, 24]. AT consists of training a model on adversarial examples generated online and designed to induce a worst-case loss. Considering a l_∞ bound ϵ for adversarial perturbations and an input-label pair (x, y), an adversarial example $x' = x + \delta$ is generated following:

$$\delta = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\max} \quad L^E(\theta, x + \delta, y) \tag{2}$$

The model is then trained minimizing the loss $L_{AT} = L^E(\theta, x + \delta, y)$. For the combination of Adversarial Training with frequency-based constraints, we define the loss functions $L^{AT,freq}$:

$$L_{i,j}^{AT,freq}(\theta, x, y) = L_{i,j}^{freq}(\theta, x + \delta, y) \tag{3}$$

As AT can be prone to robust overfitting [34], we perform early-stopping relatively to the error on an hold-out set of test set examples when training models with these loss functions. We evaluate the robustness against the l_∞ PGD attack with $\epsilon = 0.03$ for CIFAR10 and SVHN, considering the same sanity checks and attack parameters as the ones used in Section 7, to ensure the tightest evaluation of robustness as possible. Results are presented in Tables 3 and 4 for CIFAR10 and SVHN, respectively. We highlight the fact that, when dealing with Adversarial Training, there is an

inherent trade-off between accuracy and robustness [6]. Therefore, for a fair comparison, we focus on the benefits of the addition of frequency-based constraints, only for no (or up to 0.1%) loss of natural accuracy (highlighted in bold in Tables 3 and 4). Particularly, for CIFAR10, a model trained with loss $L_{AT,all,2,10}$ outperforms AT by 0.06 on adversarial accuracy. For SVHN, this improvement reaches 0.12 for a model trained with loss $L_{AT,all,10,4}$.

From these results, two important and dependent observations can be made. In first place, it shows that existing defense scheme can in fact benefit from constraints related to frequency properties. However, in second place, it shows that the intrinsic effects of the defense scheme (AT) has an influence on the effectiveness of such frequency-based regularization. Indeed, relatively to each data set, there is no connection between the frequency constraints which allow the loss function $L^{AT,freq}$ to outperform Adversarial Training, and the ones which bring robustness when considered alone (7.2). Indeed, as an example, for CIFAR10, no robustness was brought by constraint on the HSF task (Table 1), whereas this is the loss $L_i^{AT,high}$ which outperforms Adversarial Training. A possible explanation is very likely to come from Adversarial Training biasing the sensitivity of models towards the low frequencies. Therefore, when combining frequency related constraint with a defense scheme, not only do the frequency properties of the data at stake play an important role, but frequency related effects of the defense itself has a strong influence.

8 Conclusion

In this paper, we investigate through experiments the link between frequency-based processing and adversarial robustness. Particularly, this allows to gain insight on the strong influence of frequency properties of the data at stake that may be very specific according to the application domain. Notably, we found that models relying predominantly on useful features for the LSF task, and with a non-sensitivity to high frequency noise show robustness when constrained to rely on useful information for the LSF task. Interestingly, when the information encompassed in the images is spread over the whole frequency spectrum, a constraint spanning a wide frequency spectrum is a viable solution. Moreover, the efficiency of frequency-based regularization when combined with existing defense schemes is strongly dependent of the nature of these schemes. These experiences as well as the conclusions of previous efforts in neural computation and cognitive psychology highlight the fact that the intrinsic frequency characteristics of data must be necessarily considered when designing robust defense strategies against integrity-based attacks of supervised models.

Acknowledgments

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014.

[2] J. Jo and Y. Bengio, "Measuring the tendency of cnns to learn surface statistical regularities," *arXiv preprint arXiv:1711.11561*, 2017.

[3] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

[4] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A fourier perspective on model robustness in computer vision," in *Advances in Neural Information Processing Systems 32*, 2019.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[6] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 2019.

[7] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 2019.

[8] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.

[9] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 135–147, ACM, 2017.

[10] U. Hwang, J. Park, H. Jang, S. Yoon, and N. I. Cho, "Puvae: A variational autoencoder to purify adversarial examples," *IEEE Access*, vol. 7, pp. 126582–126593, 2019.

[11] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2019.

[12] C. Etmann, S. Lunz, P. Maass, and C. Schoenlieb, "On the connection between adversarial robustness and saliency map interpretability," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 1823–1832, 2019.

[13] S. Kaur, J. Cohen, and Z. C. Lipton, "Are perceptually-aligned gradients a general property of robust classifiers?," in *Advances in Neural Information Processing Systems*, 2019.

[14] A. Chan, Y. Tay, Y. S. Ong, and J. Fu, "Jacobian adversarially regularized networks for robustness," in *International Conference on Learning Representations*, 2020.

[15] T. Zhang and Z. Zhu, "Interpreting adversarially trained convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[16] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition," *Psychological science*, vol. 5, no. 4, pp. 195–200, 1994.

[17] M. Mermillod, P. Bonin, L. Mondillon, D. Alleysson, and N. Vermeulen, "Coarse scales are sufficient for efficient categorization of emotional facial expressions: Evidence from neural computation," *Neurocomputing*, vol. 73, no. 13-15, pp. 2522–2531, 2010.

[18] R. M. French, M. Mermillod, A. Chauvin, P. C. Quinn, and D. Mareschal, "The importance of starting blurry: Simulating improved basic-level category learning in infants due to weak visual acuity," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 24, 2002.

[19] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau, "Shield: Fast, practical defense and vaccination for deep learning using jpeg compression," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

[20] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: Dnn-oriented jpeg compression against adversarial examples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Z. Zhang, C. Jung, and X. Liang, "Adversarial defense by suppressing high-frequency components," *arXiv preprint arXiv:1908.06566*, 2019.

[22] S. Addepalli, B. Vivek, A. Baburaj, G. Sriramanan, and R. Venkatesh Babu, "Towards achieving adversarial robustness by enforcing feature consistency across bit planes," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[23] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *International Conference on Learning Representations*, 2019.

[25] Z. Wang, Y. Yang, A. Shrivastava, V. Rawal, and Z. Ding, "Towards frequency-based explanation for robust cnn," *arXiv preprint arXiv:2005.03141*, 2020.

[26] Y. Sharma, G. W. Ding, and M. A. Brubaker, "On the effectiveness of low frequency perturbations," *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[27] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., University of Toronto, 2009.

[28] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop Deep Learning & Unsupervised Feature Learning*, 2011.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[30] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference (BVCM)*, 2016.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[32] C. Xie, Z. Zhang, J. Wang, Y. Zhou, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2019.

[33] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[34] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, 2020.