Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

On the Effectiveness of Low Frequency Perturbations

Yash Sharma , **Gavin Weiguang Ding** and **Marcus A. Brubaker**
Borealis AI

Abstract

Carefully crafted, often imperceptible, adversarial perturbations have been shown to cause state-of-the-art models to yield extremely inaccurate outputs, rendering them unsuitable for safety-critical application domains. In addition, recent work has shown that constraining the attack space to a low frequency regime is particularly effective. Yet, it remains unclear whether this is due to generally constraining the attack search space or specifically removing high frequency components from consideration. By systematically controlling the frequency components of the perturbation, evaluating against the top-placing defense submissions in the NeurIPS 2017 competition, we empirically show that performance improvements in both the white-box and black-box transfer settings are yielded only when low frequency components are preserved. In fact, the defended models based on adversarial training are roughly as vulnerable to low frequency perturbations as undefended models, suggesting that the purported robustness of state-of-the-art ImageNet defenses is reliant upon adversarial perturbations being high frequency in nature. We do find that under L-inf-norm constraint 16/255, the competition distortion bound, low frequency perturbations are indeed perceptible. This questions the use of the L-inf-norm, in particular, as a distortion metric, and, in turn, suggests that explicitly considering the frequency space is promising for learning robust models which better align with human perception.

1 Introduction

Despite the impressive performance deep neural networks have shown, researchers have discovered that they are, in some sense, 'brittle'; small carefully crafted 'adversarial' perturbations to their inputs can result in wildly different outputs [Szegedy *et al.*, 2013]. Even worse, these perturbations have been shown to *transfer*: learned models can be successfully manipulated by adversarial perturbations generated by attacking distinct models. An attacker can discover a model's vulnerabilities even without access to it.

The goal of this paper is to investigate the relationship between a perturbation's frequency properties and its effectiveness, and is motivated by recent work showing the effectiveness of low frequency perturbations in particular. [Guo *et al.*, 2018] shows that constraining the perturbation to the low frequency subspace improves the query efficiency of the decision-based gradient-free boundary attack [Brendel *et al.*, 2017]. [Zhou *et al.*, 2018] achieves improved transferability by suppressing high frequency components of the perturbation. Similarly, [Sharma *et al.*, 2018] applied a 2D Gaussian filter on the gradient w.r.t. the input image during the iterative optimization process to win the CAAD 2018 competition[1].

However, two questions still remain unanswered:

1. is the effectiveness of low frequency perturbations simply due to the *reduced search space* or specifically due to the use of *low frequency components*? and
2. under what conditions are low frequency perturbations more effective than unconstrained perturbations?

To answer these questions, we design systematic experiments to test the effectiveness of perturbations manipulating specified frequency components, utilizing the discrete cosine transform (DCT). Testing against state-of-the-art ImageNet [Deng *et al.*, 2009] defense methods, we show that, when perturbations are constrained to the low frequency subspace, they are 1) generated faster; and are 2) more transferable. These results mirror the performance obtained when applying spatial smoothing or downsampling-upsampling operations. However, if perturbations are constrained to other frequency subspaces, they perform worse in general. This confirms that the effectiveness of low frequency perturbations is due to the application of a low-pass filter in the frequency domain of the perturbation rather than a general reduction in the dimensionality of the search space.

On the other hand, we also notice that the improved effectiveness of low frequency perturbations is only significant for defended models, but not for clean models. In fact, the state-of-the-art ImageNet defenses in test are roughly as vulnerable to low frequency perturbations as undefended models, suggesting that their purported robustness is reliant upon the assumption that adversarial perturbations are high frequency

[1]Competition on Adversarial Attacks and Defenses: http://hof.geekpwn.org/caad/en/index.html

in nature. As we show, this issue is not shared by the state-of-the-art on CIFAR-10 [Madry *et al.*, 2017], as the dataset is too low-dimensional for there to be a diverse frequency spectrum. Finally, based on the perceptual difference between the unconstrained and low frequency attacks, we discuss the problem of using the commonly used ℓ_∞ norm as a perceptual metric for quantifying robustness, illustrating the promise in utilizing frequency properties to learn robust models which better align with human perception. Our supplementary material is provided here[2].

2 Background

Generating adversarial examples is an optimization problem, while generating transferable adversarial examples is a generalization problem. The optimization variable is the perturbation, and the objective is to fool the model, while constraining (or minimizing) the magnitude of the perturbation. ℓ_p norms are typically used to quantify the strength of the perturbation; though they are well known to be poor perceptual metrics [Zhang *et al.*, 2018]. Constraint magnitudes used in practice are assumed to be small enough such that the ball is a subset of the imperceptible region.

Adversarial perturbations can be crafted in not only the *white-box* setting [Carlini and Wagner, 2017b; Chen *et al.*, 2017a] but in limited access settings as well [Chen *et al.*, 2017b; Alzantot *et al.*, 2018a], when solely query access is allowed. When even that is not possible, attacks operate in the *black-box* setting, and must rely on transferability. Finally, adversarial perturbations are not a continuous phenomenon, recent work has shown applications in discrete settings (e.g. natural language) [Alzantot *et al.*, 2018b; Lei *et al.*, 2018].

Numerous approaches have been proposed as defenses, to limited success. Many have been found to be easily circumvented [Carlini and Wagner, 2017a; Sharma and Chen, 2018; Athalye *et al.*, 2018], while others have been unable to scale to high-dimensional complex datasets, e.g. ImageNet [Smith and Gal, 2018; Papernot and McDaniel, 2018; Li *et al.*, 2018; Schott *et al.*, 2018]. Adversarial training, training the model with adversarial examples [Goodfellow *et al.*, 2014; Tramèr *et al.*, 2017; Madry *et al.*, 2017; Ding *et al.*, 2018], has demonstrated improvement, but is limited to the properties of the perturbations used, e.g. training exclusively on ℓ_∞ does not provide robustness to perturbations generated under other distortion metrics [Sharma and Chen, 2017; Schott *et al.*, 2018]. In the NeurIPS 2017 ImageNet competition, winning defenses built upon these trained models to reduce their vulnerabilities [Kurakin *et al.*, 2018; Xie *et al.*, 2018].

3 Methods

3.1 Attacks

We consider ℓ_∞-norm constrained perturbations, where the perturbation δ satisfies $\|\delta\|_\infty \leq \epsilon$ with ϵ being the maximum perturbation magnitude, as the NeurIPS 2017 competition bounded δ with $\epsilon = 16$. The Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2014] provides a simple, one-step gradient-based perturbation of ℓ_∞ ϵ size as follows:

$$\delta_{\text{FGSM}} = s \cdot \epsilon \cdot \text{sign}(\nabla_x J(x, y)) \quad (1)$$

where x is the input image, J is the classification loss function, $\text{sign}(\cdot)$ is the element-wise sign function[3]. When y is the true label of x and $s = +1$, δ is the *non-targeted* attack for misclassification; when y is a *target* label other than the true label of x and $s = -1$, δ is the *targeted* attack for manipulating the network to wrongly predict y.

FGSM suffers from an "underfitting" problem when applied to non-linear loss function, as its formulation is dependent on a linearization of J about x. The Basic Iterative Method (BIM) [Kurakin *et al.*, 2016; Madry *et al.*, 2017], otherwise known as PGD (without random starts), runs FGSM for multiple iterations to rectify this problem. The top-placing attack in the previously mentioned NeurIPS 2017 competition, the Momentum Iterative Method (MIM) [Dong *et al.*, 2017], replaces the gradient $\nabla_x J(x, y)$ with a "momentum" term to prevent the "overfitting" problem, caused by poor local optima, in order to improve transferability. Thus, we use this method for our NeurIPS 2017 defense evaluation.

Figure 1: Masks used to constrain the frequency space where $n = 128$ and $d = 299$ (ImageNet). Red denotes frequency components of the perturbation which will be masked when generating the adversarial example, both during and after the optimization process.

Cln_1	`[InceptionV3]`
Cln_3	`[InceptionV3, InceptionV4, ResNetV2_101]`
Adv_1	`[AdvInceptionV3]`
Adv_3	`[AdvInceptionV3, Ens3AdvInceptionV3, Ens4AdvInceptionV4]`

Table 1: Models used for generating black-box transfer attacks.

3.2 Frequency Constraints

Our goal is to examine whether the effectiveness of low frequency perturbations is due to a reduced search space in general or due to the specific use of a low-pass filter in the frequency domain of the perturbation. To achieve this, we use the *discrete cosine transform* (DCT) [Rao and Yip, 2014] to constrain the perturbation to only modify certain frequency components of the input.

The DCT decomposes a signal into cosine wave components with different frequencies and amplitudes. Given a 2D image (or perturbation) $x \in \mathbb{R}^{d \times d}$, the DCT Transform of x is $v = \text{DCT}(x)$, where the entry $v_{i,j}$ is the magnitude of its corresponding basis functions.

[2] https://arxiv.org/abs/1903.00073 (appendix)

[3] $\text{sign} = 1$ if $x > 0$, $\text{sign} = -1$ if $x < 0$, $\text{sign} = 0$, if $x = 0$.

The numerical values of i and j represent the frequencies, i.e. smaller values represent lower frequencies and vice versa. The DCT is invertible, with an inverse transform $x = \text{IDCT}(v)$[4].

We remove certain frequency components of the perturbation δ by applying a mask to its DCT transform $\text{DCT}(\delta)$. We then reconstruct the perturbation by applying IDCT on the masked DCT transform. Specifically, the mask, $m \in \{0,1\}^{d \times d}$, is a 2D matrix image whose pixel values are 0's and 1's, and the "masking" is done by element-wise product.

We can then reconstruct the "transformed" perturbation by applying the IDCT to the masked $\text{DCT}(\delta)$. The entire transformation can then be represented as:

$$\text{FreqMask}(\delta) = \text{IDCT}(\text{Mask}(\text{DCT}(\delta))) \ . \tag{2}$$

Accordingly in our attack, we use the following gradient

$$\nabla_\delta J(x + \text{FreqMask}(\delta), y) \ .$$

We use 4 different types of FreqMask to constrain the perturbations, as shown in Figure 1. `DCT_High` only preserves high frequency components; `DCT_Low` only preserves low frequency components; `DCT_Mid` only preserves mid frequency components; and `DCT_Rand` preserves randomly sampled components. For reduced dimensionality n, we preserve $n \times n$ components. Recall that $v = \text{DCT}(x)$, `DCT_Low` preserves components $v_{i,j}$ if $1 \leq i, j \leq n$; `DCT_High` masks components if $1 \leq i, j \leq \sqrt{d^2 - n^2}$; `DCT_Mid` and `DCT_Rand` also preserve $n \times n$ components, the detailed generation processes can be found in the appendix. Figure 1 visualizes the masks when $d = 299$ (e.g. ImageNet) and $n = 128$. Note that when $n = 128$, we only preserve $128^2/299^2 \approx 18.3\%$ of the frequency components, a small fraction of the original unconstrained perturbation.

4 Results and Analyses

To evaluate the effectiveness of perturbations under different frequency constraints, we test against models and defenses from the NeurIPS 2017 Adversarial Attacks and Defences Competition [Kurakin *et al.*, 2018].

Threat Models

We evaluate attacks in both the non-targeted and targeted case, and measure the attack success rate (ASR) on 1000 test examples from the NeurIPS 2017 development toolkit[5]. We test on $\epsilon = 16/255$ (competition distortion bound) and iterations $= [1, 10]$ for the non-targeted case; $\epsilon = 32/255$ and iterations $= 10$ for the targeted case. The magnitude for the targeted case is larger since targeted attacks, particularly on ImageNet (1000 classes), are significantly harder. As can be seen in Figure 5 and 6, unconstrained adversarial perturbations generated under these distortion bounds are still imperceptible.

[4]DCT / IDCT is applied to each color channel independently.

[5]https://www.kaggle.com/c/6864/download/dev_toolkit.zip

Attacks

As described in Section 3, we experiment with the original unconstrained MIM and frequency constrained MIM with masks shown in Figure 1. For each mask type, we test $n = [256, 128, 64, 32]$ with $d = 299$. For `DCT_Rand`, we average results over 3 random seeds.

To describe the attack settings, we specify model placeholders A and B. We call an attack *white-box*, when we attack model A with the perturbation generated from A itself. We call an attack *grey-box*, when the perturbation is generated from A, but used to attack a "defended" A, where a defense module is prepended to A. We call an attack *black-box* (transfer), when the perturbation generated from A is used to attack distinct B, where B can be defended or not. Note that this is distinct from the black-box setting discussed in [Guo *et al.*, 2018], in which query access is allowed.

Target Models and Defenses for Evaluation

We evaluate each of the attack settings against the top defense solutions in the NeurIPS 2017 competition [Kurakin *et al.*, 2018]. Each of the top-4 NeurIPS 2017 defenses prepend a tuned (or trained) preprocessor to an ensemble of classifiers, which for all of them included the strongest available adversarially trained model: `EnsAdvInceptionResNetV2`[6] [Tramèr *et al.*, 2017]. Thus, we use `EnsAdvInceptionResNetV2` to benchmark the robustness[7] of adversarially trained models.

We then prepend the preprocessors from the top-4 NeurIPS 2017 defenses to `EnsAdvInceptionResNetV2`, and denote the defended models as D1, D2, D3, and D4, respectively. Regarding the preprocessors, D1 uses a trained denoiser where the loss function is defined as the difference between the target model's outputs activated by the clean image and denoised image [Liao *et al.*, 2017]; D2 uses random resizing and random padding [Xie *et al.*, 2017]; D3 uses a number of image transformations: shear, shift, zoom, and rotation [Thomas and Elibol, 2017]; and D4 simply uses median smoothing [Kurakin *et al.*, 2018].

For our representative cleanly trained model, we evaluate against the state-of-the-art `NasNetLarge_331`[8] [Zoph *et al.*, 2017]. We denote `EnsAdvInceptionResNetV2` as EnvAdv and `NasNetLarge_331` as NasNet for brevity.

Source Models for Perturbation Generation

For white-box attacks, we evaluate perturbations generated from NasNet and EnsAdv to attack themselves respectively. For grey-box attacks, we use perturbations generated from EnsAdv to attack D1, D2, D3, and D4 respectively. For black-box attacks, since the models for generating the perturbations need to be distinct from the ones being attacked, we use 4 different sources (ensembles) which vary in ensemble size and whether the models are adversarially trained or cleanly trained, as shown in Table 1. In summary, for black-box attacks, perturbations generated from Adv_1, Adv_3, Cln_1,

[6]https://github.com/tensorflow/models/tree/master/research/adv_imagenet_models

[7]`EnsAdvInceptionResNetV2` is to be attacked.

[8]https://github.com/tensorflow/models/tree/master/research/slim

and Cln_3 are used to attack NasNet, EnsAdv, D1, D2, D3, and D4.

4.1 Overview of the Results

As described, we test the unconstrained and constrained perturbations in the white-box, grey-box, and black-box scenarios. Representative results are shown in Figure 2a, 2b, 2c, and 2d. In each of these plots, the vertical axis is attack success rate (ASR), while the horizontal indicates the number of frequency components kept (Dimensionality). Unconstrained MIM is shown as a horizontal line across the dimensionality axis for ease of comparison. In each figure, the plots are, from left to right, non-targeted attack with iterations $= 1$, non-targeted with iterations $= 10$, and targeted with iterations $= 10$. From these figures, we can see that DCT_Low always outperforms the other frequency constraints, including DCT_High, DCT_Mid and DCT_Rand.

In the appendix, we show results where the perturbation is constrained using a spatial smoothing filter or a downsampling-upsampling filter (perturbation resized with bilinear interpolation). The performance mirrors that of Figure 2a, 2b, 2c, and 2d, further confirming that the effectiveness of low frequency perturbations is not due to a general restriction of search space, but due to the low frequency regime itself. Thus, in our remaining experiments, we focus on low frequency constrained perturbations induced with DCT_Low.

We compare ASR and relative changes across all black-box transfer pairs between standard unconstrained MIM and MIM constrained with DCT_Low $n = 128$, on non-targeted attacks with both iterations $= 1$ and iterations $= 10$. This comparison is visualized in Figure 3 and 4. We also show that these results do not transfer to the significantly lower-dimensional CIFAR-10 dataset ($d = 32$, minimum n used in ImageNet experiments), as the rich frequency spectrum of natural images is no longer present.

4.2 Observations and Analyses

White-box Evaluation

Figure 2a and 2b show the white-box ASRs on EnsAdv and NasNet respectively. For EnsAdv, we can see that DCT_Low improves ASR in the non-targeted case with iterations $= 1$ and in the targeted case with iterations $= 10$, but not in the non-targeted case with iterations $= 10$. However, in this case, DCT_Low still outperforms other frequency constraints and does not significantly deviate from unconstrained MIM's performance. When the number of iterations is large enough that unconstrained MIM can succeed consistently, constraining the space only limits the attack, but otherwise, the low frequency prior is effective. Therefore, low frequency perturbations are more "iteration efficient", as they can be found more easily with a less exhaustive search, which is practically helpful computationally.

However, for white-box attacks on NasNet in Figure 2b, we see that although DCT_Low still outperforms the other frequency constraints, it does perform worse than unconstrained MIM. Comparing Figure 2a and 2b, it is clear that DCT_Low performs similarly against the adversarially trained model as with the cleanly trained model, the difference here is due to unconstrained MIM performing significantly better against the cleanly trained model than against the adversarially trained model. This implies that the low frequency prior is useful against defended models, in particular, since it exploits the space where adversarial training, which is necessarily imperfect, fails to reduce vulnerabilities.

Grey-box Evaluation

As previously mentioned, in the grey-box case, we generate the perturbations from the undefended EnsAdv and use them to attack D1, D2, D3 and D4 (which include preprocessors prepended to EnsAdv). Figure 2c shows the ASR results averaged over D1∼4. DCT_Low outperforms unconstrained MIM by large margins in all cases. Comparing Figure 2a with Figure 2c, the larger difference between unconstrained MIM and DCT_Low in the grey-box case reflects the fact that the top NeurIPS 2017 defenses are not nearly as effective against low frequency perturbations as they are against standard unconstrained attacks. In fact, DCT_Low yields the same ASR on D1, the winning defense submission in the NeurIPS 2017 competition, as on the adversarially trained model without the preprocessor prepended; the preprocessors are not effective (at all) at protecting the model from low frequency perturbations, even in the targeted case, where success is only yielded if the model is fooled to predict, out of all 1000 class labels, the specified target label. Results against the individual defenses are presented in the appendix.

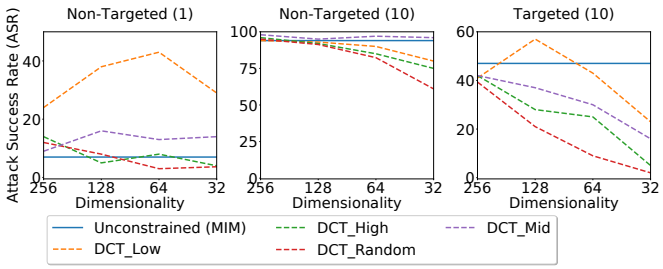
Black-box Evaluation (Defended)

For assessing black-box transferability, we use Cln_1, Cln_3, Adv_1, Adv_3 in Table 1 as the source models for generating perturbations, and attack EnsAdv and D1∼4, resulting in 20 source-target pairs in total. The average ASR results over these pairs are reported in Figure 2d. In the non-targeted case, we again see that DCT_Low significantly outperforms unconstrained MIM. However, in the targeted case, constraining to the low frequency subspace does not enable MIM to succeed in transferring to distinct black-box defended models due to the difficult nature of targeted transfer.

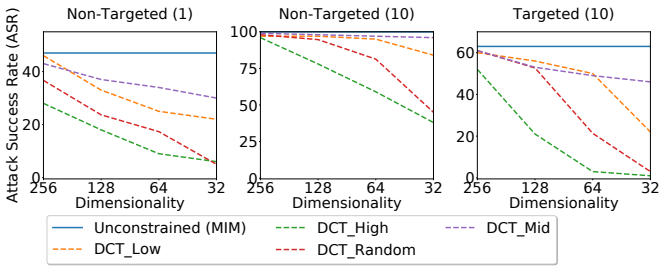
Next, we look at individual source-target pairs. For each pair, we compare DCT_Low ($n = 128$) with unconstrained MIM in the non-targeted case with iterations $= 1$ and iterations $= 10$. Results for all frequency configurations with varied dimensionality are provided in the appendix. Figure 3 shows the transferability matrices for all source-target pairs, where for each subplot, the row indices denote source models, and the column indices denote target models. The value (and associated color) in each gridcell represent the ASR for the specified source-target pair. For Figure 4, the gridcell values represent the relative difference in ASR between the target model and the cleanly trained model (Cln)[9], using the source model of the corresponding row.

Comparing (a) to (b) and (c) to (d) in Figure 3, it is clear that low frequency perturbations are much more effective than unconstrained MIM against defended models. Specifically, we can see that DCT_Low is significantly more effective than unconstrained MIM against EnsAdv, and D1∼4 provide almost no additional robustness to EnsAdv when low

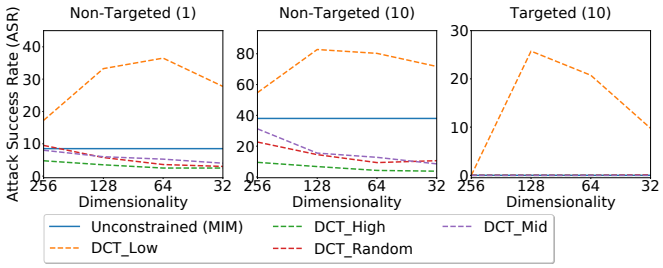
[9]The relative difference for the target model = (ASR on the target model - ASR on Cln) / ASR on Cln.



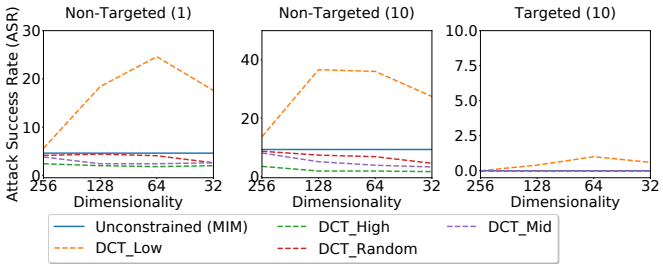
(a) **White-box** attack on adversarially trained model, EnsAdv.



(b) **White-box** attack on standard cleanly trained model, NasNet.

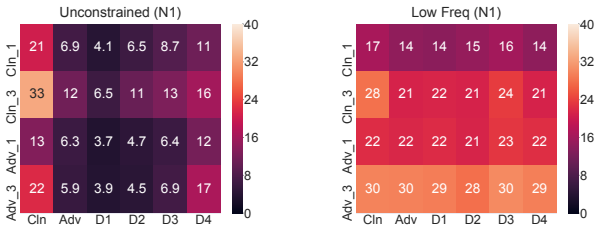


(c) **Grey-box** attack on top-4 NeurIPS 2017 defenses prepended to adversarially trained model.



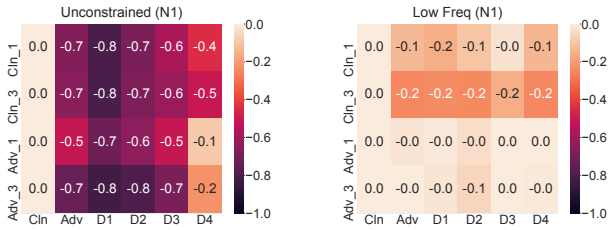
(d) **Black-box** attack on sources (Table 1) transferred to defenses (EnsAdv + D1∼4)

Figure 2: Number of iterations in parentheses. Non-targeted with $\epsilon = 16/255$, targeted with $\epsilon = 32/255$.



(a) (b) (c) (d)

Figure 3: Transferability matrices with attack success rates (ASRs), comparing unconstrained MIM with low frequency constrained `DCT_Low` ($n = 128$) in the non-targeted case. First row is with iterations $= 1$, second is with iterations $= 10$. The column Cln is NasNet, Adv is EnsAdv.



(a) (b) (c) (d)

Figure 4: Transferability matrices with attack relative difference in ASR with the Cln model (first column). Rows and columns in each subfigure is indexed in the same way as Figure 3.

frequency perturbations are applied.

Black-box Evaluation (Undefended)

However, we do observe that `DCT_Low` does not improve black-box transfer between undefended cleanly trained models, which can be seen by comparing indices (Cln_1,Cln) and (Cln_3,Cln) between Figure 3 (a) and (b), as well as (c) and (d). As discussed when comparing white-box performance against cleanly trained and adversarially trained models, low frequency constraints are not generally more effective, but instead exploit the vulnerabilities in currently proposed defenses.

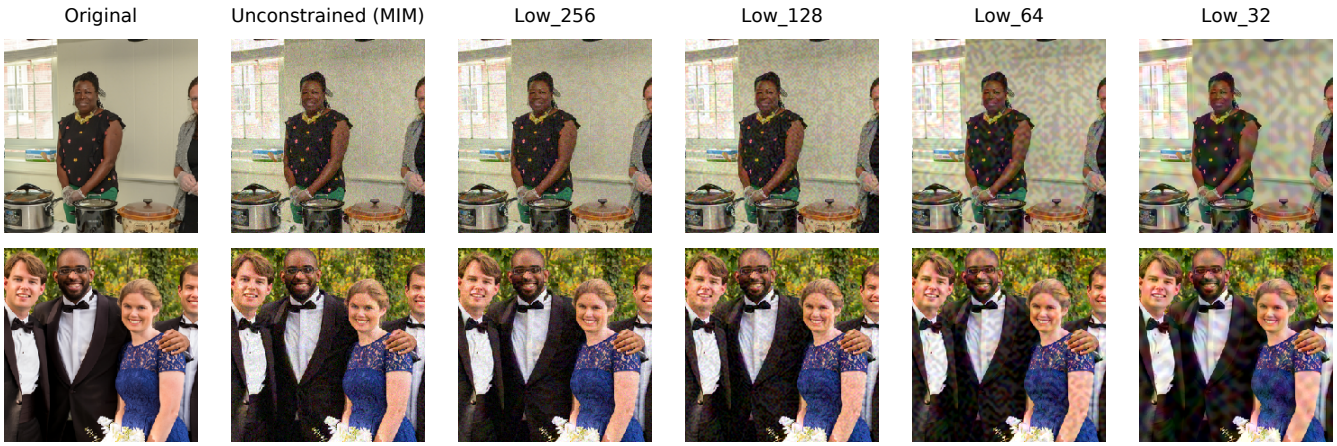


Figure 5: Adversarial examples generated with ℓ_∞ $\epsilon = 16/255$ distortion

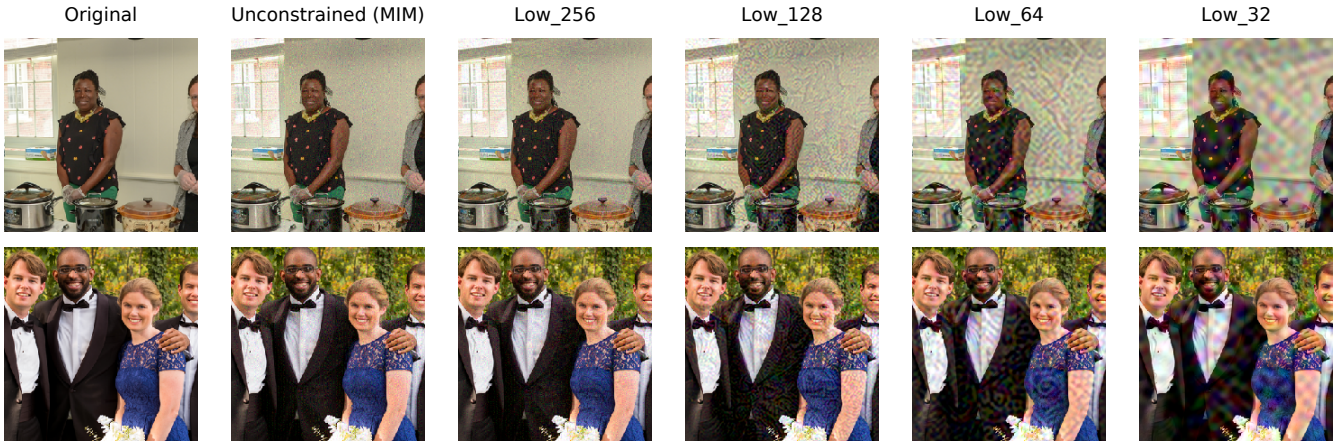


Figure 6: Adversarial examples generated with ℓ_∞ $\epsilon = 32/255$ distortion

4.3 Effectiveness of Low Frequency on Undefended Models v.s. Defended Models

In the last section, we showed that DCT_Low is highly effective against adversarially trained models and top-performing preprocessor-based defenses, in the white-box, grey-box and black-box cases. However, low frequency does not help when only cleanly trained models are involved, i.e. white-box on clean models and black-box transfer between clean models. To explain this phenomenon, we hypothesize that the state-of-the-art ImageNet defenses considered here do not reduce vulnerabilities within the low frequency subspace, and thus DCT_Low is roughly as effective against defended models as against clean models, a property not seen when evaluating with standard, unconstrained attacks.

This can be most clearly seen in Figure 4, which presents the normalized difference between ASR on each of the target models with ASR on the cleanly trained model. The difference is consistently smaller for DCT_Low than for unconstrained MIM, and nearly nonexistent when the perturbations were generated against adversarially trained (defended) models (Adv_1,Adv_3). Thus, as discussed, defended models are roughly as vulnerable as undefended models when encountered by low frequency perturbations.

Dim	White (Adv)	Black (Adv)	Black (Cln)
32	54.6	38.1	14.4
24	48.1	33.1	14.4
16	46.4	28.8	14.4
8	37.0	25.4	14.4
4	26.5	20.0	14.0

Table 2: Non-targeted attack success rate (ASR) with iterations $=$ 10 and $\epsilon = 8/255$ of DCT_Low in the white-box and black-box settings (transfer from distinct adversarially trained and cleanly trained models of the same architecture) against adversarially trained model with 12.9% test error [Madry *et al.*, 2017].

4.4 Effectiveness of Low Frequency on CIFAR-10

We test the effectiveness of low frequency perturbations on the much lower-dimensional than ImageNet, CIFAR-10 dataset ($d = 299$ to $d = 32$), attacking the state-of-the-

art adversarially trained model [Madry *et al.*, 2017]. Experiment results with 1000 test examples can be seen in Table 2. Constraining the adversary used for training (non-targeted PGD [Kurakin *et al.*, 2016; Madry *et al.*, 2017]; iterations $= 10$ and $\epsilon = 8/255$) with `DCT_Low`, and evaluating both in the white-box and black-box settings (transfer from distinct adversarially trained and cleanly trained models of the same architecture), we observe that dimensionality reduction only hurts performance. This suggests that the notion of low frequency perturbations is not only constrained to the computer vision domain, but also only induces problems for robustness in the realm of high-dimensional natural images.

5 Discussion

Our experiments show that the results seen in recent work on the effectiveness of constraining the attack space to low frequency components [Guo *et al.*, 2018; Zhou *et al.*, 2018; Sharma *et al.*, 2018] are not due to generally reducing the size of the attack search space. When evaluating against state-of-the-art adversarially trained models and winning defense submissions in the NeurIPS 2017 competition in the white-box, grey-box, and black-box settings, significant improvements are only yielded when low frequency components of the perturbation are preserved. Low frequency perturbations are so effective that they render state-of-the-art ImageNet defenses to be roughly as vulnerable as undefended, cleanly trained models under attack.

However, we also noticed that low frequency perturbations do not improve performance when defended models are not involved, seen in evaluating white-box performance against and black-box transfer between cleanly trained models. Low frequency perturbations do not yield faster white-box attacks on clean models, nor do they provide more effective transfer between clean models.

Our results suggest that the state-of-the-art ImageNet defenses, based on necessarily imperfect adversarial training, only significantly reduce vulnerability outside of the low frequency subspace, but not so much within. Against defenses, low frequency perturbations are more effective than unconstrained ones since they exploit the vulnerabilities which purportedly robust models share. Against undefended models, constraining to a subspace of significantly reduced dimensionality is unhelpful, since undefended models share vulnerabilities beyond the low frequency subspace. Understanding whether this observed vulnerability in defenses is caused by an intrinsic difficulty to being robust in the low frequency subspace, or simply due to the (adversarial) training procedure rarely sampling from the low frequency region is an interesting direction for further work.

Are Low frequency Perturbations Perceptible?

Our results show that the robustness of currently proposed ImageNet defenses relies on the assumption that adversarial perturbations are high frequency in nature. Though the adversarial defense problem is not constrained to achieving robustness to imperceptible perturbations, this is a reasonable first step. Thus, in Figure 5, we visualize low frequency constrained adversarial examples under the competition ℓ_∞-norm constraint $\epsilon = 16/255$. Though the perturbations do not significantly change human perceptual judgement, e.g. the top example still appears to be a standing woman, the perturbations with $n \leq 128$ are indeed perceptible.

Although it is well-known that ℓ_p-norms (in input space) are far from metrics aligned with human perception, exemplified by their widespread use, it is still assumed that with a small enough bound (e.g. ℓ_∞ $\epsilon = 16/255$), the resulting ball will constitute a subset of the imperceptible region. The fact that low frequency perturbations are fairly visible challenges this common belief. In addition, if the goal is robustness to imperceptible perturbations, our study suggests this might be achieved, without adversarial training, by relying on low frequency components, yielding a much more computationally practical training procedure. In all, we hope our study encourages researchers to not only consider the frequency space, but perceptual priors in general, when bounding perturbations and proposing tractable, reliable defenses.

References

[Alzantot *et al.*, 2018a] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, and Mani Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. *arXiv preprint arXiv:1805.11090*, 2018.

[Alzantot *et al.*, 2018b] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018.

[Athalye *et al.*, 2018] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[Brendel *et al.*, 2017] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

[Carlini and Wagner, 2017a] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017.

[Carlini and Wagner, 2017b] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2017.

[Chen *et al.*, 2017a] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: Elastic-net attacks to deep neural networks via adversarial examples. *arXiv preprint arXiv:1709.0414*, 2017.

[Chen *et al.*, 2017b] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec '17*, pages 15–26. ACM, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[Ding *et al.*, 2018] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Max-margin adversarial (mma) training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018.

[Dong *et al.*, 2017] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *arXiv preprint arXiv:1710.06081*, 2017.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Guo *et al.*, 2018] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*, 2018.

[Kurakin *et al.*, 2016] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018.

[Lei *et al.*, 2018] Qi Lei, Lingfei Wu, Pin-Yu Chen, Alexandros G. Dimakis, Inderjit S. Dhillon, and Michael Witbrock. Discrete attacks and submodular optimization with applications to text classification. *arXiv preprint arXiv:1812.00151*, 2018.

[Li *et al.*, 2018] Yingzhen Li, John Bradshaw, and Yash Sharma. Are generative classifiers more robust to adversarial attacks? *arXiv preprint arXiv:1802.06552*, 2018.

[Liao *et al.*, 2017] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. *arXiv preprint arXiv:1712.02976*, 2017.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Papernot and McDaniel, 2018] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[Rao and Yip, 2014] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[Schott *et al.*, 2018] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

[Sharma and Chen, 2017] Yash Sharma and Pin-Yu Chen. Attacking the madry defense model with l1-based adversarial examples. *arXiv preprint arXiv:1710.10733*, 2017.

[Sharma and Chen, 2018] Yash Sharma and Pin-Yu Chen. Bypassing feature squeezing by increasing adversary strength. *arXiv preprint arXiv:1803.09868*, 2018.

[Sharma *et al.*, 2018] Yash Sharma, Tien-Dung Le, and Moustafa Alzantot. Caad 2018: Generating transferable adversarial examples. *arXiv preprint arXiv:1810.01268*, 2018.

[Smith and Gal, 2018] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

[Szegedy *et al.*, 2013] 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.

[Thomas and Elibol, 2017] Anil Thomas and Oguz Elibol. Defense against adversarial attack: 3rd place. https://github.com/anlthms/nips-2017, 2017.

[Tramèr *et al.*, 2017] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[Xie *et al.*, 2017] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

[Xie *et al.*, 2018] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint arXiv:1801.03924*, 2018.

[Zhou *et al.*, 2018] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang;. Transferable adversarial perturbations. In *ECCV (14)*, pages 471–486. Springer, 2018.

[Zoph *et al.*, 2017] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *arXiv preprint arXiv:1707.07012*, 2017.