Exploring Adversarial Robustness of Vision Transformers in the Spectral Perspective

Gihyun Kim Juyeop Kim Jong-Seok Lee
Yonsei University, Republic of Korea
{kkh9314,juyeopkim,jong-seok.lee}@yonsei.ac.kr

Abstract

The Vision Transformer has emerged as a powerful tool for image classification tasks, surpassing the performance of convolutional neural networks (CNNs). Recently, many researchers have attempted to understand the robustness of Transformers against adversarial attacks. However, previous researches have focused solely on perturbations in the spatial domain. This paper proposes an additional perspective that explores the adversarial robustness of Transformers against frequency-selective perturbations in the spectral domain. To facilitate comparison between these two domains, an attack framework is formulated as a flexible tool for implementing attacks on images in the spatial and spectral domains. The experiments reveal that Transformers rely more on phase and low frequency information, which can render them more vulnerable to frequency-selective attacks than CNNs. This work offers new insights into the properties and adversarial robustness of Transformers.

1. Introduction

Convolution neural networks (CNNs) have served as the dominant architecture for computer vision for a long time. However, Transformer-based structures have recently emerged as another promising architecture [11], achieving even better performance than CNNs especially in image classification.

CNNs are known to be vulnerable to adversarial attacks, i.e., an imperceptible perturbation added to an image can fool a trained CNN so that it misclassifies the attacked image [2]. Investigating the robustness of a model against adversarial attacks is important because not only the vulnerability issue is critical in security-sensitive applications but also such investigation can lead to a better understanding of the operating mechanism of the model. Then, a naturally arising question is: how vulnerable are Transformers compared to CNNs?

The researches comparing adversarial robustness of CNNs and Transformers do not reach consistent conclusions. One group of studies claims that Transformers are more robust to adversarial attacks than CNNs [3, 6, 27, 32]. However, another group of studies claims that the two architectures have similar levels of robustness [5, 7, 25].

This paper aims to explore the adversarial robustness of Transformers from a previously unexplored perspective. It has been noted in previous studies that Transformers rely more on low frequency features [6, 28] while CNNs focus more on high frequency features [18, 35]. From this point of view, popular gradient-based attack methods, which is mostly used in the existing studies comparing adversarial robustness of CNNs and Transformers, tend to perturb high frequency features in images through spatial domain perturbations and this might cause CNNs to be fooled more easily than Transformers. To alleviate such bias, we formulate an attack framework that allows flexible perturbations in both spatial and spectral domains, with the hope to find certain types of adversarial perturbations for which Transformers become more vulnerable than CNNs. Note that we do not intend to develop a new stronger attack in the frequency domain, but aim to formulate a unified attack framework that can directly perturb the pixel values, magnitude spectrum, and phase spectrum of an image.

Figure 1 shows an example, where each of the magnitude, phase, and pixel components is perturbed using our attack framework for ResNet50 and ViT-B. It can be observed that attacking different components induces different distortion patterns in the image. The distortion pattern also varies depending on the target model. Thus, a standardized scale is required to make proper comparison among different target models and attack domains. We choose to utilize an image quality metric to measure the attack strength across various models and attack methods. In addition, we consider a wide range of attack strength because the superiority in terms of robustness between models may change depending on the amount of perturbation.

We conduct extensive experiments to compare the adversarial robustness of off-the-shelf pre-trained CNN and Transformer models. The results demonstrate that Trans-

formers are not necessarily more robust than CNNs, and in particular, Transformers tend to be more vulnerable to perturbations inserted to the magnitude and phase components in the frequency domain. Our contributions can be summarized as follows.

- To explore adversarial robustness in both the spatial domain and spectral domain, we formulate an attack framework that can perturb the magnitude and phase spectra in the spectral domain and the pixel values in the spatial domain. In particular, examining frequency-selective perturbations by the attacks is one of the key factors to a deeper understanding of the adversarial robustness of Transformers.
- Using the attack framework, we evaluate various models of CNN and Transformer over a wide range of image quality of attacked images. Relative vulnerability among the models is analyzed in various viewpoints such as attack type, attack strength, model size, and training data. As a main result, it is found that Transformers are particularly vulnerable to phase perturbations concentrated in the low frequency region.
- We conduct in-depth analyses to investigate the frequency-dependent behaviors and importance of spectral information in Transformers. Additionally, we explain the vulnerability of Transformers to the phase attack from the viewpoint of linearity of models and attacks.

The remainder of the paper is organized as follows. Section 2 briefly reviews the related works. In Section 3, we present our method to explore the robustness of Transformers. The results are presented in Section 4 with rich analysis. Finally, conclusion is made in Section 5.

2. Related Works

2.1. Vision Transformers

The vision Transformer (ViT) has appeared as a powerful neural architecture using the self-attention mechanism [11]. Several variants of ViT have been also proposed. The Swin Transformer [23] improves the efficiency over ViT using a shifted window scheme for self-attention. To resolve the issue that Transformers require a large dataset for training, the data-efficient image Transformer (DeiT) [34] is trained through distillation from a CNN teacher. Other variants include token-to-token ViT [41], pyramid ViT [36], Transformer in Transformer [17], cross-covariance image Transformer [4], etc.

2.2. Adversarial Attack Methods

The goal of a typical adversarial attack is to change the classification result of a model by injecting a noise-like perturbation to the image while the perturbation is kept imperceptible in order not to be detected easily. The perturbation is usually found via gradient-based optimization. The fast gradient sign method (FGSM) [14] uses the sign of gradient of the classification loss. The projected gradient descent (PGD) method [24] implements a stronger attack by iteratively optimizing the perturbation. These attacks limit the L_p norm of the perturbation to control the attack strength.

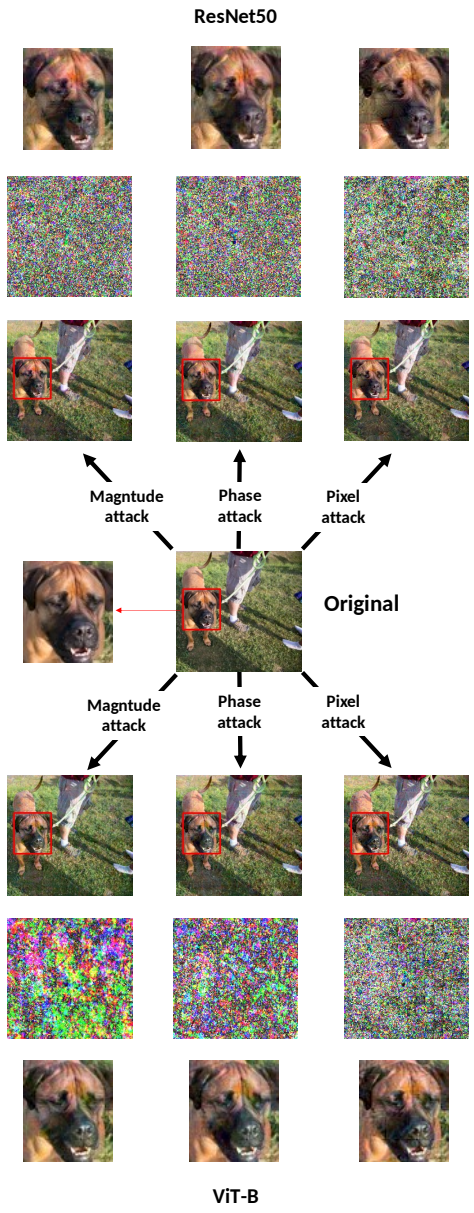
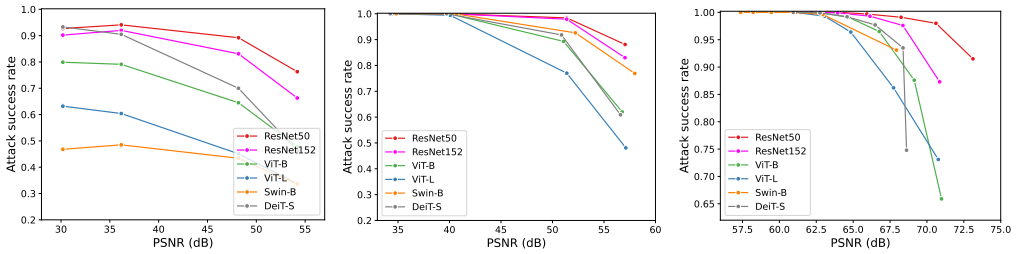


Figure 1. Example case of our attack perturbing the magnitude, phase, and pixel values, respectively, for ResNet50 and ViT-B. The attacked image, the difference between the original and attacked images (after amplification for visualization), and an enlarged area of the attacked image are shown in each case.



(a) FGSM (b) PGD (c) C&W

Figure 2. Example of evaluations showing that the adversarial vulnerability is dependent on the attack strength (represented as PSNR)

The C&W attack [8] optimizes the weighted sum of the amount of perturbation and the classification loss, which is known to be one of the strongest attacks. Note that our attack formulation (Section 3.2) is inspired by the C&W.

Frequency-domain filtering can be used to constrain perturbations only in certain frequency regions [16, 21, 33], which are still based on image-domain attacks. The work in [12] suggests an attack in the DCT domain for CNNs, which only drops certain frequency components via quantization. In [1], an attacked image is generated by swapping a frequency component (among four single-level wavelet components) of the original image with that of another image, which considers only a limited type of spectral perturbation and, furthermore, is agnostic to the target model and thus is not suitable for direct comparison of the robustness of CNNs and ViTs. Recently, direct perturbation in the frequency domain is also tried on CNNs [38], which is different from our method that can perturb the magnitude and phase components separately.

2.3. Adversarial Robustness of Transformers

As mentioned in the introduction, two conflicting conclusions exist in the literature regarding the relative adversarial robustness of CNNs and Transformers. In one group of studies, it is claimed that Transformers are more robust than CNNs. The studies in [3, 6, 30, 32] commonly make a conclusion that Transformers are more robust against gradient-based attacks including FGSM, PGD, and C&W because CNNs rely on high frequency information while Transformers do not. In [20], it is suggested that the severe nonlinearity of the input-output relationship of Transformers causes their higher robustness than CNNs. When adversarial training is considered, it is recognized that ViTs exhibit more robust generalization than CNNs [22]. Another group of studies argues that Transformers are as vulnerable to attacks as CNNs. The work in [25] finds that ViTs are not advantageous over ResNet in terms of robustness against various attack methods such as FGSM, PGD, and C&W. In [7], it is shown that CNNs and Transformers are similarly vulnerable against various natural and adversarial perturbations. In [5], it is attempted to compare CNNs and Transformers on a common training setup, from which it is concluded that they have similar adversarial robustness. An attack perturbing single patches is designed in [13] to induce vulnerability of ViTs, and similarly the patch attack [19] is applied to Transformers in [15].

2.4. Our Distinguished Contributions

Our work is distinguished from the previous works as follows. (1) Compared to some previous works where only a limited number of models are compared [3, 6, 13, 30] or a limited range of attack strength is considered [5, 7, 15, 25], we consider the trade-off characteristics between vulnerability and attack strength in an extensive manner for diverse CNNs and Transformers over a wide range of image quality degradation. (2) Some previous works explain different levels of vulnerability of CNNs and Transformers in terms of their reliance on different frequency components [3, 6, 30, 32]. While this conclusion is based on the results of attacks in the spatial domain, we directly impose perturbations in the spectral domain to implement frequency-selective attacks for Transformers. (3) In [13] and [15], it is shown that localized perturbations on image patches effectively attack Transformers due to the patch-wise self-attention mechanism of Transformers. However, successful patch perturbations are usually visible, which is undesirable as adversarial attacks, whereas we show that Transformers can become vulnerable by spectral-domain perturbations inducing imperceptible global distortion in the images.

3. Method

3.1. Consideration of Attack Strength

In many popular attack methods, the attack strength can be controlled by certain parameters (e.g., L_∞ norm of perturbation in FGSM, the balancing parameter between the amount of distortion and the change of the classification loss in C&W). As an attack becomes strong, the target model becomes more vulnerable, but the image distortion becomes more perceptible. We notice that depending on the considered attack strength, the superiority of one model to another in terms of robustness may vary.

Figure 2 shows an example of this issue. For each trained model, we apply the FGSM, PGD, or C&W attack to the images from the NeurIPS 2017 Adversarial Challenge [39]. For FGSM and PGD, the L_∞ norm constraint of the perturbation varies among $\{0.1/255, 0.5/255, 1/255, 4/255, 8/255\}$. For C&W, the balance parameter between the amount of perturbation and the cross-entropy varies among $\{1, 0.4, 0.2, 0.1, 0.05, 0.01\}$. The figure shows the attack success rate (ASR) of the attacked images with respect to the attack strength. Here, the image quality of the attacked images (peak signal-to-noise ratio (PSNR) in this figure) is used to represent the attack strength. Overall, ResNet models tend to be more vulnerable than Transformers in all attacks by showing higher ASR, which is consistent with the results in [3, 6, 32]. When we have a look at the details, observations demonstrating the vulnerability dependent on the attack strength can be also made. For instance, in Figure 2a, DeiT-S is more robust than both ResNet models over 40 dB but they show similar vulnerability at 30-35 dB; in Figure 2b, all models are similarly vulnerable showing almost 100% of ASR for low PSNR values, but their vulnerability becomes different after about 40 dB; Figure 2c shows a similar trend to Figure 2b but in a higher PSNR range. These results demonstrate that it is important to examine a wide range of attack strength with various models in order to better understand the vulnerability of different models.

3.2. Attack Method

Different from the existing works, we formulate a unified attack framework that is capable of perturbing images both spatially and spectrally. There are two key motivations behind our approach. First, the previous studies point out that CNNs tend to rely on high frequency information in images, which may be why they appear more vulnerable than Transformers [6,32]. In other words, the popular attack methods injecting high frequency noise may be unfavorable to CNNs. The unified framework tries to alleviate such an inherent bias by enabling to flexibly perturb images in both spatial location-selective and frequency-selective manners. Second, we aim to analyze the mechanisms of Transformers in various viewpoints through the results of attacks applied in different domains.

The Fourier transform of an image X can be written by

$$\mathcal{F}\{X\} = M \cdot e^{j\phi}, \tag{1}$$

where M and ϕ are the magnitude and phase spectra, respectively. The attacked image X' is obtained by the combination of multiplicative magnitude perturbation[1] δ_{mag}, additive phase perturbation δ_{phase}, and additive pixel perturbation δ_{pixel} as follows:

$$\tilde{X}' = \mathcal{F}^{-1}\left\{\text{clip}_{0,\infty}(M \otimes \delta_{\text{mag}}) \cdot e^{j(\phi+\delta_{\text{phase}})}\right\} + \delta_{\text{pixel}}, \tag{2}$$

$$X' = \text{clip}_{0,1}(\tilde{X}'), \tag{3}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform, $\otimes$ is the element-wise multiplication, and $\text{clip}_{a,b}(x)$ limits the value of each element of x within a and b. Here, we assume that the pixel values are normalized within 0 and 1. Note that δ_{mag} and δ_{phase} are kept to be symmetric in order to ensure the resulting image after the inverse Fourier transform to have real-valued pixel values.

We consider attacks employing one among δ_{mag}, δ_{phase}, and δ_{pixel}, denoted as "magnitude attack," "phase attack," and "pixel attack," respectively. It is also possible to employ two or all types of perturbation at the same time, the results of which are in Supplementary Material.

The process to optimize the perturbations is inspired by the C&W attack [8]. In other words, we minimize the L_2 difference between the original and attacked images to keep the amount of perturbation as small as possible, while the cross-entropy (CE) loss is maximized to fool the classifier. Thus, the loss function of the unified attack framework is given by

$$\text{Loss} = \lambda \cdot L_2(X', X) - \text{CE}(f(X'), y), \tag{4}$$

where λ is a parameter balancing the L_2 difference and CE, which controls the attack strength, $f(\cdot)$ is the classifier, and y is the ground truth. This loss can be minimized by a gradient-descent approach to obtain δ_{mag}, δ_{phase}, and/or δ_{pixel}, and consequently the attacked image X'.

4. Experiments

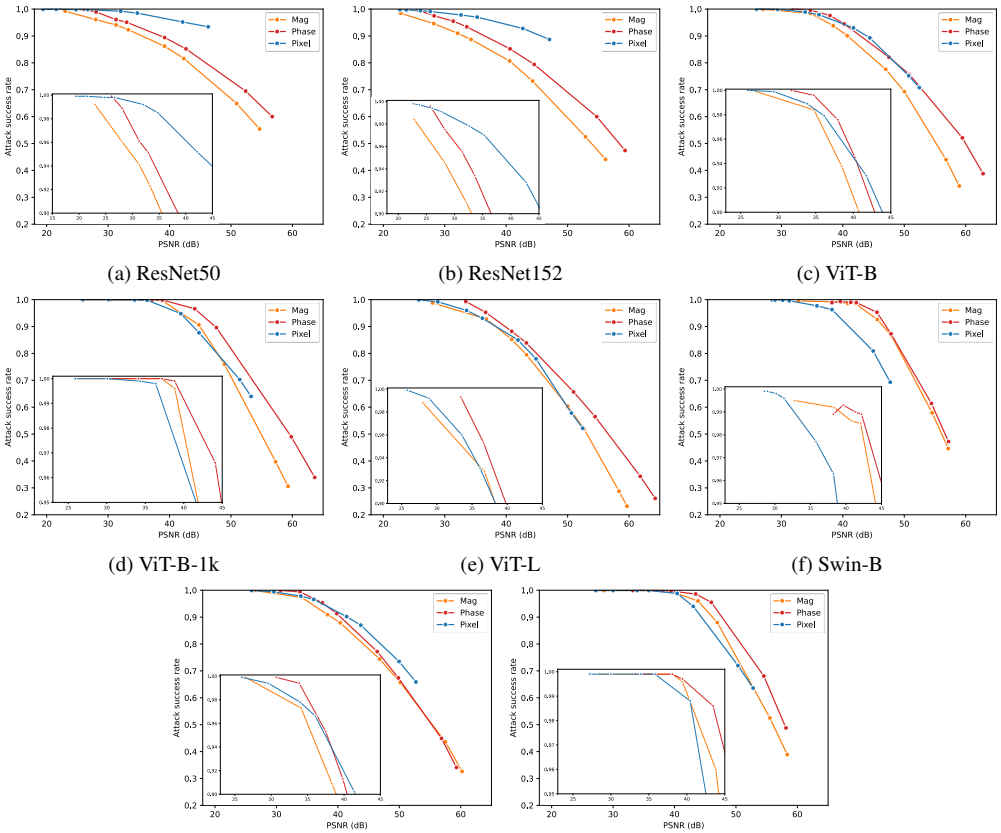
4.1. Setup

We aim to benchmark the adversarial robustness of the pre-trained models that serve as off-the-shelf solutions in general image classification applications. We consider ResNet50, ResNet152 as CNNs, and ViT-B/16, ViT-L/16, DeiT-S, and Swin-B as Transformers. Here, S, B and L mean small, base, and large, and /16 means the patch size. ResNet50 and ResNet152 are from the torchvision models [29] trained on ImageNet-1k [31]. ViT-B, ViT-L, DeiT-S, and Swin-B are from the timm module [40], which are pre-trained on ImageNet-21k [10] and finetuned on ImageNet-1k. ViT trained on ImageNet-1k and DeiT-S without distillation are also considered.

To evaluate the adversarial robustness of the models, the image dataset from the NeurIPS 2017 Adversarial Challenge [39] is used.

To obtain perturbations by minimizing the loss in Eq. (4), we use the Adam optimizer with a fixed learning rate

[1] We also tried an additive magnitude perturbation but it was not optimized well because the magnitude spectrum has values over a wide range.



(a) ResNet50 (b) ResNet152 (c) ViT-B

(d) ViT-B-1k (e) ViT-L (f) Swin-B

(g) DeiT-S (h) DeiT-S with no distillation

Figure 3. Comparison of different attacks for each model. The range for high attack success rates is enlarged for better visualization.

of 5×10^{-3} and a weight decay parameter of 5×10^{-6}. The maximum number of iterations is set to 1000. We also set a termination condition that the optimization stops when the loss does not improve for five consecutive iterations. We vary the value of λ as $\{1, 10^3, 5\times10^3, 10^4, 5\times10^4, 10^5, 5\times 10^5, 10^6\}$ to control the attack strength.

We employ image quality metrics to enable representation of the attack strength commonly across different attacks introduced in the previous section, including PSNR, multi-scale structural similarity index measure (MS-SSIM) [37], mean deviation similarity index (MDSI) [26], and learned perceptual image patch similarity (LPIPS) [42]. Here, the results using PSNR are shown; those using the other metrics are provided in Supplementary Material, which show similar trends to those using PSNR.

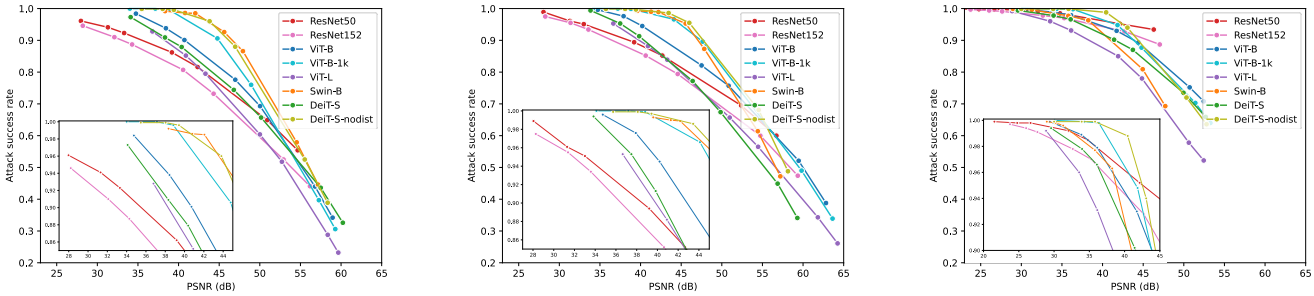
4.2. Results

We present the results in two perspectives: (1) comparison of different attacks for each model, and (2) comparison of different models for each attack type.

4.2.1 Comparison of Attacks

We first compare various attacks implemented using our unified attack framework. Figure 3 shows the results in terms of ASR with respect to PSNR. All the attacks can achieve (almost) 100% of ASR at the lower extremes of PSNR for all models. However, their relative effectiveness varies depending on the model. For ResNet50 and ResNet152, the pixel attack appears to be the strongest, whereas the phase attack is the strongest for ViTs and Swin-B. This shows that the flexible frequency domain attack is able to overcome the limitation of the pixel attack primarily perturbing high frequency information, and becomes a potent tool for attacking Transformers. We assume that the vulnerability of DeiT-S to the pixel attack is attributed to its training method, which involves distillation from a CNN teacher. This assumption is supported by the observation that DeiT-S without distillation is also more vulnerable to the phase attack, which is consistent with the vulnerability of ViTs and Swin-B.

It is observed that the phase attack is mostly stronger than the magnitude attack. Further analysis on this is presented in Section 4.4. Attention maps before and after at-



(a) Magnitude attack (b) Phase attack (c) Pixel attack

Figure 4. Comparison of different models for each attack type. The range for high attack success rates is enlarged for better visualization.

tacks are also compared in Supplementary Material.

4.2.2 Comparison of Vulnerability of Models

In Figure 4, ASR of the models is compared with respect to PSNR for each attack type. Notably, when ASR is relatively high, Transformers are either equally or more robust than ResNet models under the pixel attack but are more vulnerable to the magnitude and phase attacks. Again, this is attributed to the flexible perturbations in the frequency domain, which will be further analyzed later. At PSNR over 45-50 dB, where ASR is low, the observed trends do not hold anymore, i.e., some Transformers (ViTs and DeiT-S) become similarly robust to ResNet models under the magnitude and phase attacks, and ResNet models become more vulnerable to all Transformers under the pixel attack. When Swin-B is compared to ViT-B and ViT-L, the former shows higher vulnerability than the latter for the magnitude and phase attacks. When the model size is concerned, larger models (ViT-L and ResNet152) are more robust than smaller models (ViT-B and ResNet50) under all attacks. While [7] also observed a similar trend using pixel-domain attacks (FGSM and PGD), we find that the same also holds for the attacks in the spectral domain.

We also compare ViTs trained in different environments, i.e. ViT-B and ViT-B-1k. It is observed that ViT-B is more robust than ViT-B-1k below 45 dB, but becomes more vulnerable when PSNR increases, particularly for the pixel attack. It is worth noting that [7] also pointed out the advantage of training with a larger dataset to enhance robustness; we additionally find that the benefit of a larger dataset is even more prominent for the magnitude and phase attacks than for the pixel attack, but the benefit disappears when the amount of perturbation is small.

DeiT-S behaves more like ResNet than the other Transformers due to the distillation using CNN. In Supplementary Material, the models pre-trained on the same dataset (i.e., ImageNet-1k) are compared, where similar trends to the above results are observed.

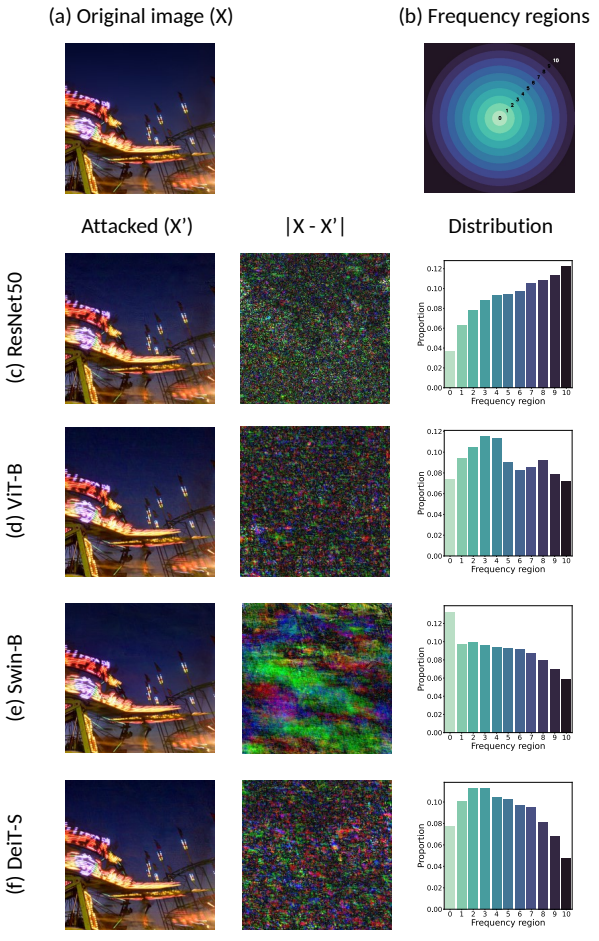
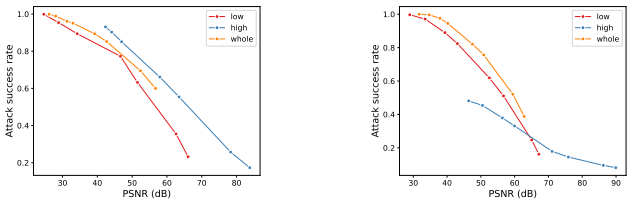


Figure 5. Example of the attacked image under the phase attack, distortion in the pixel domain (magnified by $\times 20$), and distribution of the distortion over different frequency regions.

4.3. Analysis

We further investigate the particular vulnerability of Transformers to the phase attack.



(a) ResNet50 (b) ViT-B

Figure 6. Results of the phase attack when the perturbation is restricted to reside only in the low or high frequency band. The case without restriction is also shown as 'whole.'

4.3.1 Frequency Analysis of Perturbations

In order to investigate the effect of the phase attack on the spectral characteristics of images, we apply the Fourier transform to the difference between the original and attacked images, and analyze the magnitudes in different frequency regions, which are defined as illustrated in Figure 5(b). Figures 5(c) to 5(f) show the attacked image, distortion in the pixel domain, and magnitude distribution. The averaged results over images are shown in Supplementary Material. In the case of ResNet50, the distortion is concentrated on the high frequency regions, whereas low frequency regions are mainly distorted in the other models. Since CNNs and Transformers rely more on high and low frequency information, respectively [6, 18, 28, 35], the attack effectively injects perturbations in such vulnerable frequency regions. Consequently, the distortion pattern significantly differs according to those properties.

4.3.2 Frequency-Restricted Attacks

We examine the case where the perturbation is applied to a limited frequency band. Figure 6 compares the phase attack when the phase perturbation is restricted to only the low frequency band (regions 1 and 2 in Figure 5(b)) or the high frequency band (region 10 in Figure 5(b)). Perturbing only the high frequency band is more effective than perturbing the whole band for ResNet50, as it is particularly vulnerable to high frequency perturbations. However, for ViT-B, perturbing only the high frequency band is the least effective and the case without restriction (i.e., 'whole') implements the strongest attack because effective perturbations need to be applied to low-intermediate frequency regions as shown in Figure 5(e).

4.3.3 Linearity of Models and Attacks

We further analyze the vulnerability of Transformers to the phase attack in the viewpoint of the linearity of models and attacks. A noticeable difference between the attacks, as shown in Eq. 2, is whether the input is *linearly* perturbed

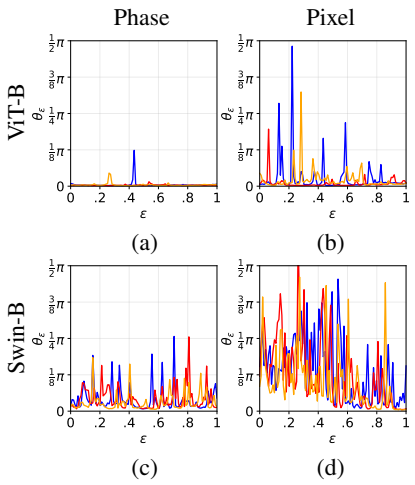


Figure 7. Direction changes of output features (θ_ϵ) for ViT-B and Swin-B. Different colors indicate the results of different images.

or not. If a classifier f is linear, its input-output relationship will be linear, i.e.,

$$g(X + \epsilon \cdot \delta) = g(X) + \epsilon \cdot g(\delta), \tag{5}$$

where X is an input image, δ is a perturbation, ϵ is a scalar, and g is the output at the penultimate layer of f (i.e., output feature). For such a classifier, δ can effectively move $g(X+\epsilon\cdot\delta)$ along the determined adversarial direction $g(\delta)$ in the feature space as intended. In contrast, if f is nonlinear, $g(X + \epsilon \cdot \delta)$ will not directly follow the direction of $g(\delta)$ in the feature space, which will weaken the attack. This explanation coincides with [14], which showed that adversarial examples are a result of models being too linear.

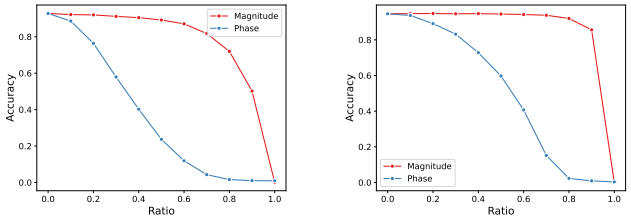
To experimentally examine the linearity of models under different attacks, we use the method in [20]. Consider $X_\epsilon = X + \epsilon \cdot \delta$, where $0 \leq \epsilon \leq 1$. As ϵ increases, X moves on the straight line along the direction determined by δ in the input space. For a chosen ϵ, we examine the following quantity:

$$\theta_\epsilon = \pi - \arccos(\hat{g}_{\epsilon^-} \cdot \hat{g}_{\epsilon^+}), \tag{6}$$

where

$$\hat{g}_{\epsilon^\pm} = \frac{g(X_{\epsilon\pm\Delta\epsilon}) - g(X_\epsilon)}{||g(X_{\epsilon\pm\Delta\epsilon}) - g(X_\epsilon)||_2}. \tag{7}$$

$\Delta\epsilon$ (> 0) controls the amount of forward and backward shifts along the direction of δ in the input space. $\hat{g}_{\epsilon^\pm}$ signifies the corresponding shifts in the feature space (with normalization). θ_ϵ indicates the orientation of these shifts in the feature space, determining whether the linear displacements in the input space are also maintained in the feature space.



(a) ResNet50 (b) ViT-B

Figure 8. Classification accuracy with respect to the ratio of reduction in the magnitude or phase spectrum.

Model	Phase (%)	Magnitude (%)	Else (%)
ResNet50	2.04	0.11	97.85
ViT-B	33.92	0.26	65.82

Table 1. Proportions of magnitude-phase-recombined images that are classified to the classes of the magnitude or phase images, or some other classes.

Figure 7 shows θ_ϵ of ViT-B and Swin-B for $\delta = X' - X$, where X is an image, X' is the attacked result of X with $\lambda = 5 \times 10^3$, and $\Delta\epsilon = 0.001$. For the pixel attack, high peaks of θ_ϵ appear as ϵ changes from 0 to 1 (Figures 7(b) and 7(d)), i.e., the models are relatively nonlinear (as also shown in [20]) and thus are robust to the pixel attack that is linear. However, for the phase attack, which perturbs the input in a nonlinear manner, θ_ϵ remains as small values (Figures 7(a) and 7(c)), meaning that the features are more effectively moved in the feature space in the adversarial direction and thus become more vulnerable.

4.4. Dependence on Magnitude and Phase

In most results above, the magnitude attack appears to be weaker than the phase attack. We investigate this phenomenon further.

4.4.1 Sensitivity to Reduced Magnitude and Phase

We conduct an experiment to evaluate the impact of gradually reducing the magnitude or phase spectrum on the classification accuracy without any attack applied (i.e., $M' = M \times (1-r)$ or $\phi' = \phi \times (1-r)$, where $r \in \{0, 0.1, ..., 1\}$). The results, as shown in Figure 8, demonstrate that both ResNet50 and ViT-B are more sensitive to phase reduction than magnitude reduction. Notably, even with 90% of the magnitude reduced, ViT-B still achieves an accuracy of 85.6%. These results suggest that the corruption of the phase spectrum has a more significant impact on model performance than that of the magnitude, which explains the higher vulnerability of the models to the phase attack.

4.4.2 Magnitude-Phase Recombination

Inspired by [9], we conduct an experiment where the magnitude component of one image and the phase component of another image are recombined in the frequency domain and the classification result of this recombined image from a model is tested. For all possible image pairs, Table 1 shows the proportion of the images that are classified as the class of the magnitude image or phase image, or none of the two classes. For ResNet50, most of the cases (97.85%) do not follow either the classes of the magnitude or the phase, while for the rest, more images follow the phase classes (2.04%) than the magnitude classes (0.11%). For ViT-B, however, a considerable amount of recombined images (33.92%) are classified as the class of the phase images, while only 0.26% of the images follow the classes of the magnitude images. These results highlight the relative importance of the phase information, providing an explanation on the particular vulnerability of Transformers to the phase attack and the higher strength of the phase attack than the magnitude attack for Transformers.

5. Conclusion

We comparatively investigated the adversarial robustness of CNNs and Transformers using the unified attack framework with consideration of the relationship between the attack strength and ASR. Our study provides a unique contribution to the field by revealing that the vulnerability of models to adversarial attacks is highly dependent on the type of attack and the frequency regions where the perturbations are injected. Specifically, we found that Transformers are more vulnerable to the phase and magnitude attacks that mainly inject perturbations in the low frequency regions, while CNNs are more vulnerable to the pixel attack that injects perturbations mainly in the high frequency regions. We provided an explanation for this difference in the viewpoint of linearity of the models and attacks.

Our results provide insights into the underlying mechanisms of adversarial attacks and highlight the importance of considering the frequency domain when evaluating and improving the robustness of deep learning models. Furthermore, we observed that the phase information plays a more important role in classification for both CNNs and Transformers than the magnitude information, and reliance on the phase information is more prominent in Transformers.

As a future work, it would be interesting to explore the components that make ViTs vulnerable to the phase attack, based on which robust ViT structures could be developed.

Acknowledgements

This work was supported by Artificial Intelligence Graduate School Program, Yonsei University under Grant 2020-0-01361.

References

[1] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE Transactions on Image Processing*, 31:7338–7349, 2022. 3

[2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 1

[3] Ahmed Aldahdooh, Wassim Hamidouche, and Olivier Deforges. Reveal of vision transformers robustness against adversarial attacks. *arXiv preprint arXiv:2106.03734*, 2021. 1, 3, 4

[4] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. XCIT: Cross-covariance image transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 20014–20027, 2021. 2

[5] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than CNNs? In *Advances in Neural Information Processing Systems*, volume 34, pages 26831–26843, 2021. 1, 3

[6] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs. In *Proceedings of the 32nd British Machine Vision Conference*, 2021. 1, 3, 4, 7

[7] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021. 1, 3, 6

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 3, 4

[9] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021. 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*, 2021. 1, 2

[12] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to DNNs by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7506–7515, 2021. 3

[13] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. 2, 7

[15] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *Proceedings of the European Conference on Computer Vision*, pages 404–421, 2022. 3

[16] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Proceedings of the Uncertainty in Artificial Intelligence*, pages 1127–1137, 2020. 3

[17] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. In *Advances in Neural Information Processing Systems*, volume 34, pages 15908–15919, 2021. 2

[18] Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017. 1, 7

[19] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *Proceedings of International Conference on Machine Learning*, pages 2507–2515, 2018. 3

[20] Juyeop Kim, Junha Park, Songkuk Kim, and Jong-Seok Lee. Curved representation space of vision transformers. *arXiv preprint arXiv:2210.05742*, 2022. 3, 7, 8

[21] Xiu-Chuan Li, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. F-mixup: Attack CNNs from Fourier perspective. In *Proceedings of the 25th International Conference on Pattern Recognition*, pages 541–548, 2020. 3

[22] Aishan Liu, Shiyu Tang, Siyuan Liang, Ruihao Gong, Boxi Wu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architectural design and adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4096–4107, 2023. 3

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018. 2

[25] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 1, 3

[26] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *IEEE Access*, 4:5579–5590, 2016. 5

[27] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 23296–23308, 2021. 1

[28] Namuk Park and Songkuk Kim. How do vision transformers work? In *Proceedings of the 10th International Conference on Learning Representations*, 2022. 1, 7

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 4

[30] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 3

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 4

[32] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 1, 3, 4

[33] Yash Sharma, Gavin Weiguang Ding, and Marcus A Brubaker. On the effectiveness of low frequency perturbations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3389–3396, 2019. 3

[34] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10347–10357, 2021. 2

[35] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 1, 7

[36] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021. 2

[37] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Proceedings of the Asilomar Conference on Signals, Systems & Computers*, volume 2, pages 1398–1402, 2003. 5

[38] Zerui Wen. Fourier attack–a more efficient adversarial attack method. In *Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence*, pages 125–130, 2022. 3

[39] Ross Wightman. NIPS 2017 adversarial competition (pytorch). https://github.com/rwightman/pytorch-nips2017-adversarial, 2017. 4

[40] Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 4

[41] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 558–567, 2021. 2

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5