Adversarial amplitude swap towards robust image classifiers

Chun Yang Tan, Kazuhiko Kawamoto, Hiroshi Kera
Chiba University, Japan
chunyangtan@chiba-u.jp, kawa@faculty.chiba-u.jp, kera@chiba-u.jp

Abstract

The vulnerability of convolutional neural networks (CNNs) to image perturbations such as adversarial perturbations and common corruptions has recently been investigated from the perspective of frequency. In this study, we investigate the effect of the amplitude and phase spectra of adversarial images on the robustness of CNN classifiers. Extensive experiments revealed that the images generated by combining the amplitude spectrum of adversarial images and the phase spectrum of clean images accommodates moderate and general perturbations, and training with these images equips a CNN classifier with more general robustness, performing well under both adversarial perturbations and common corruptions. We also found that two types of overfitting (catastrophic overfitting and robust overfitting) can be circumvented by the aforementioned spectrum recombination. We believe that these results contribute to the understanding and the training of truly robust classifiers.

1. Introduction

Despite their state-of-the-art performance, convolutional neural networks (CNNs) have been found to be vulnerable to perturbations in images such as adversarial perturbations [6] and common corruptions [3]. Although such perturbations do not change the semantic information of the images, they substantially degrade the performance of CNN classifiers. Numerous data augmentation approaches had been introduced to improve the robustness of CNNs against different types of perturbations. For instance, adversarial training was proposed to improve the robustness against adversarial perturbations and APR [1] was proposed to improve the robustness against common corruptions. However, most of the methods only address a certain type of perturbations without considering the robustness in a broad sense.

The vulnerability of CNNs to image perturbations are often linked to the disparity of behaviors in the frequency domain between humans and CNNs in image recognition task.

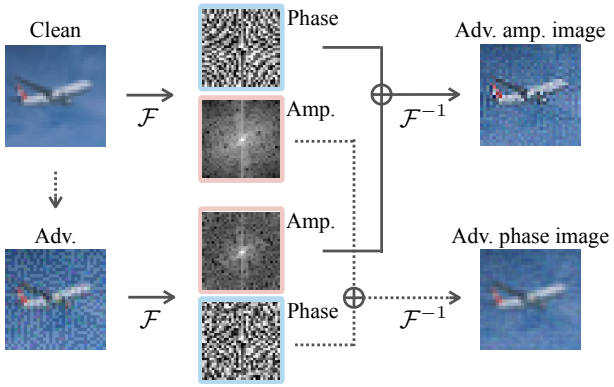


Figure 1. The pipeline of the proposed frequency-based data augmentation. The maps $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the discrete Fourier transform and its inverse. The amplitude spectra of a clean image and its adversarial image are swapped to generate adversarial amplitude and adversarial phase images.

For instance, Wang et al. [8] suggested that CNNs exploit high-frequency image components that are not perceptible to humans. Chen et al. [1] suggested that one can train a CNN classifier to be more robust against common corruptions by encouraging the model to focus more on the phase spectrum.

In this study, we investigate the nature of the amplitude and phase spectra of adversarial images. Inspired by [1], we propose a frequency-based data augmentation method in which the amplitude spectra are swapped between clean and adversarial images to generate adversarial amplitude and adversarial phase images (Figure 1). In particular, training with the former images provide more general robustness because it encourages a CNN classifier to focus on the phase spectra, leading to CNNs that are robust against common corruptions. In addition to that, adversarial amplitude images also contain the amplitude spectra of the adversarial images, which enable CNNs to be robust against adversarial perturbations. Through extensive experiments, we revealed that the training with the adversarial amplitude images achieved the followings:

- It led to CNN classifiers that are robust to both adver-

sarial perturbations and common corruptions, whereas the baseline methods only enhance either of them.

- It prevented the catastrophic overfitting [9] and the robust overfitting [4] during the adversarial training, whereas each of the other convention data augmentation methods, the random crop, and horizontal flip, did not.

- It also helped the CNN classifiers learn from moderate, strong, and even extremely strong adversarial images, whereas both the standard training with adversarial images and that with adversarial phase images suffer from catastrophic overfitting, leading to poor performance on adversarial perturbations other than those specifically trained against.

The experimental results show that (i) the amplitude spectrum of adversarial images accommodates moderate and general perturbations that helped classifiers to equip more general robustness, and (ii) the phase spectrum of adversarial images tended to be moderate but still adversarial, which may improve the adversarial robustness of classifiers but also retain the risk of catastrophic overfitting. We believe that this study deepens the understanding of general perturbations in images, particularly from the frequency perspective, and contributes to the future development of truly robust image classifiers.

2. Adversarial Amplitude Swap

To train a CNN classifier to be robust against both common corruptions and adversarial perturbations, we consider that both the amplitude and phase spectra of images play important roles. We propose a new frequency-based data augmentation method to swap the amplitude spectrum of the former with that of the latter to generate two augmented images: an adversarial amplitude image, which has the amplitude spectrum of the adversarial image and the phase spectrum of the clean image, and an adversarial phase image, which is the opposite (Figure 1). Formally, the process of the adversarial amplitude swap is performed as follows. First, given a clean image $\boldsymbol{x}$, an adversarial image $\boldsymbol{x}_{\text{adv}}$ is generated. Then, the discrete Fourier transform (DFT) is applied to the two images to obtain the amplitude–phase decompositions, $(\mathcal{A}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{x}))$ and $(\mathcal{A}(\boldsymbol{x}_{\text{adv}}), \mathcal{P}(\boldsymbol{x}_{\text{adv}}))$. The adversarial amplitude and adversarial phase images, x_{AA} and x_{AP}, are then constructed by the inverse DFT of $(\mathcal{A}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{x}_{\text{adv}}))$ and $(\mathcal{A}(\boldsymbol{x}_{\text{adv}}), \mathcal{P}(\boldsymbol{x}))$, respectively; namely,

$$\boldsymbol{x}_{\text{AA}} = \mathcal{F}^{-1}\left(\mathcal{A}(\boldsymbol{x}_{\text{adv}}) \cdot e^{i \cdot \mathcal{P}(\boldsymbol{x})}\right), \tag{1}$$

$$\boldsymbol{x}_{\text{AP}} = \mathcal{F}^{-1}\left(\mathcal{A}(\boldsymbol{x}) \cdot e^{i \cdot \mathcal{P}(\boldsymbol{x}_{\text{adv}})}\right). \tag{2}$$

Algorithm 1: Adversarial amplitude swap

Input: $\boldsymbol{x}$: clean image
Output: $\boldsymbol{x}_{\text{AA}}$: adversarial amplitude image, $\boldsymbol{x}_{\text{AP}}$: adversarial phase image

$\boldsymbol{x}_{\text{adv}} \leftarrow \text{AdversarialAttack}(\boldsymbol{x})$. // E.g., by FGSM.
$\mathcal{A}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{x}) \leftarrow \text{DFT}(\boldsymbol{x})$
$\mathcal{A}(\boldsymbol{x}_{\text{adv}}), \mathcal{P}(\boldsymbol{x}_{\text{adv}}) \leftarrow \text{DFT}(\boldsymbol{x}_{\text{adv}})$
$\boldsymbol{x}_{\text{AA}} \leftarrow \text{InvDFT}(\mathcal{A}(\boldsymbol{x}_{\text{adv}}), \mathcal{P}(\boldsymbol{x}))$
$\boldsymbol{x}_{\text{AP}} \leftarrow \text{InvDFT}(\mathcal{A}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{x}_{\text{adv}}))$

Table 1. The classification accuracy (%) of WideResNet40-2 classifiers trained on CIFAR-10 with different combination of images. The FGSM attack with $\epsilon = 8/255$ was used in training. The top-2 results are indicated in bold while the best results are underlined.

	Combination of Training Data				
	Clean	APR	C&Adv	C&AA	C&AP
Clean	**94.1**	**94.3**	86.1	91.3	88.3
FGSM ($\epsilon_0 = 8$)	**66.8**	64.5	63.7	**72.4**	65.7
FGSM ($\epsilon_0 = 32$)	43.5	51.0	53.5	**60.3**	**54.3**
PGD-l_∞	0.2	0.3	**45.7**	28.7	**38.4**
PGD-l_2	2.4	7.0	**53.8**	51.7	**54.2**
Corrupted-1	**90.4**	**92.2**	86.2	89.4	87.9
Corrupted-2	88.1	**90.9**	85.0	**88.2**	86.7
Corrupted-3	85.9	**89.7**	83.8	**86.9**	85.4
Corrupted-4	83.2	**87.7**	82.2	**85.3**	83.8
Corrupted-5	79.3	**85.1**	79.5	**82.4**	81.2

Pseudo-code for this process is provided in Algorithm 1.

Note that the adversarial image changes at each training step because its generation depends on the classifier. Thus, $\boldsymbol{x}_{\text{AA}}$ has a static phase spectrum derived from $\boldsymbol{x}$, and a stochastic amplitude spectrum derived from different $\boldsymbol{x}_{\text{adv}}$, along the training. Therefore, training with $\boldsymbol{x}_{\text{AA}}$ encourages CNN classifiers to learn the static semantic features from the phase spectrum and also resist the stochastic adversarial features in the amplitude spectrum. Similarly, $\boldsymbol{x}_{\text{AP}}$ has a static amplitude spectrum of a clean image and the stochastic phase spectrum of the adversarial image, which encourages the classifiers to learn more on the amplitude spectrum and resist the adversarial features in the phase spectrum.

3. Experiments

We conducted multiple experiments on CIFAR-10 and CIFAR-100 datasets. To measure the robustness of models against adversarial perturbations, we use FGSM with $\epsilon = \epsilon_0/255$ for $\epsilon_0 \in \{8, 32\}$ and PGD-l_∞ with $\epsilon = 8/255$, the step size $\alpha = 0.1$, and the number of iterations $i_{\text{iters}} = 20$, and PGD-l_2 with $\epsilon = 0.5, \alpha = 0.1, i_{\text{iters}} = 20$. For the robustness

Table 2. The classification accuracy (%) of WideResNet40-2 classifiers trained on CIFAR-100 with different combination of images. The FGSM attack with $\epsilon = 8/255$ was used in training. The top-2 results are indicated in bold while the best results are underlined.

	Combination of Training Data				
	Clean	APR	C&Adv	C&AA	C&AP
Clean	**72.5**	**72.0**	63.7	66.3	61.3
FGSM ($\epsilon_0 = 8$)	27.5	31.7	**70.1**	**38.8**	31.1
FGSM ($\epsilon_0 = 32$)	12.7	20.4	**41.9**	**27.3**	21.1
PGD-l_∞	0.0	0.0	1.3	**8.9**	**13.3**
PGD-l_2	0.1	1.0	0.5	**19.5**	**24.3**
Corrupted-1	**70.3**	**73.5**	67.2	68.9	66.2
Corrupted-2	66.8	**71.7**	65.4	**67.0**	64.6
Corrupted-3	63.9	**69.8**	63.5	**65.4**	62.9
Corrupted-4	60.3	**67.2**	61.1	**63.3**	61.1
Corrupted-5	55.2	**63.4**	57.5	**59.7**	57.9

Table 3. The classification accuracy (%) of WideResNet-40-2 classifiers trained on CIFAR-10 with different combination of images. The PGD-l_∞ attack with $\epsilon = 8/255, \alpha = 2/255, i_{\text{iters}} = 10$ was used in training. The top-2 results are indicated in bold, and the best results are underlined.

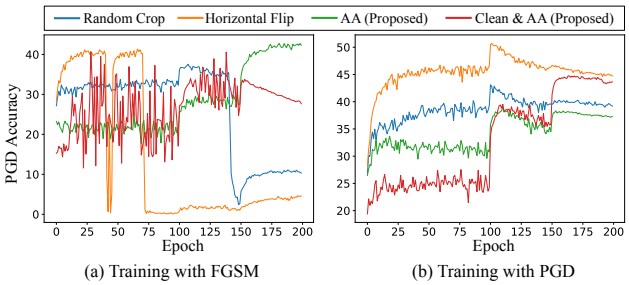
	Combination of Training Data				
	Clean	APR	C&Adv	C&AA	C&AP
Clean	**94.1**	**94.3**	83.8	88.5	86.2
FGSM ($\epsilon_0 = 8$)	**66.8**	64.5	55.2	**66.2**	62.8
FGSM ($\epsilon_0 = 32$)	43.5	41.0	38.3	**56.7**	**51.9**
PGD-l_∞	0.2	0.3	**46.2**	**42.6**	39.6
PGD-l_2	2.4	7.0	**52.1**	**52.7**	51.6
Corrupted-1	**90.4**	**92.2**	84.5	88.0	86.4
Corrupted-2	**88.1**	**90.9**	83.3	86.9	85.3
Corrupted-3	**85.9**	**89.7**	82.2	85.7	84.1
Corrupted-4	83.2	**87.7**	80.7	**84.1**	82.5
Corrupted-5	79.3	**85.1**	77.7	**81.3**	79.7

against common corruptions, we evaluated methods on the CIFAR-10-C and CIFAR-100-C datasets, which contains testset with 15 different types of noises, each appearing at five severity levels or intensities. Level-1 represents the lowest severity and level-5 represents the highest severity.

We trained classifiers with WideResNet-40-2][10]. We adopted two data-augmentation-based methods as the baselines, the APR method introduced in [1] and the standard training with clean and adversarial images (C&Adv), which corresponds to GoodFellow's adversarial training with the weight of $w = 0.5$ [2].

3.1. CIFAR-10 & CIFAR-100 Image Classification

Table 1 shows the results of models trained on the CIFAR-10 dataset. The fast gradient sign method (FGSM) attack with budget $\epsilon = 8/255$ was used in training. Notably, the model trained with clean and adversarial amplitude images (C&AA) achieved an overall improvement in the robustness tests compared to the model trained with clean images (Clean). The APR model specialized for common corruptions achieved the highest robustness against common corruptions across all severity levels but were still vulnerable to adversarial perturbations, particularly to those by the projected gradient descent (PGD) [5]. In contrast, the C&Adv, which is specialized for adversarial perturbations were robust against PGD perturbations but not common corruptions. Training with clean and adversarial phase images (C&AP), which contain the adversarial phase spectra showed the same trend as that of C&Adv.



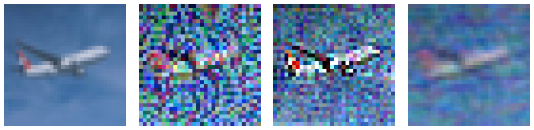
(a) Training with FGSM (b) Training with PGD

Figure 2. WideResNet-40-2 classifiers trained with conventional and proposed data augmentation methods. (a) Training with FGSM perturbations. (b) Training with PGD perturbations.

We also evaluated the methods on the CIFAR-100 dataset using the same experimental setup (Table 2). Similar trends were observed, in which the C&AA model achieved improvements in the robustness against both adversarial perturbations and common corruptions, while both the baseline models, APR and C&Adv only improved the robustness against either of the perturbations. To further evaluate the proposed method, we used the PGD-l_∞ with $\epsilon = 8/255$, step size $\alpha = 0.1$, number of iterations $i_{\text{iters}} = 10$ in training. The results are shown in Table 3. We observed the same trend where C&AA achieved improvement in general robustness while the baseline models did not.

3.2. Catastrophic and Robust Overfitting

It is known that data augmentation is effective in preventing robust overfitting during adversarial training [7]. We compared the proposed method with two conventional data augmentation methods, random crop and random horizontal flip. Figure 2 shows the model robustness against PGD perturbations when WideResNet-40-2 classifiers are trained using different data augmentation methods on the FGSM and the PGD perturbed images. Notably, both the models trained with either of the conventional data augmentation methods suffered from both catastrophic and robust



(a) Clean (b) Adv. (c) AA (d) AP

Figure 3. Examples of images generated by PGD-l_∞ with $\epsilon = 0.5, \alpha = 0.1, i_{\text{iters}} = 10$. (a) clean image, (b) adversarial image, (c) adversarial amplitude image, and (d) adversarial phase image.

Table 4. The classification accuracy (%) of WideResNet-40-2 on CIFAR-10. PGD-l_∞ with $\epsilon = 0.5, \alpha = 0.1, i_{\text{iters}} = 10$ was used in training. The best results are indicated in bold.

	Combination of Training Data		
	Clean&Adv	Clean&AA	Clean&AP
Clean	90.0	**93.7**	89.9
FGSM ($\epsilon_0 = 8$)	33.6	**71.1**	61.3
PGD-l_2	0.1	**15.5**	0.9
Corrupted	71.0	**86.2**	78.1

overfitting. However, when FGSM perturbations were used in training, the model trained with adversarial amplitude images (AA) did not suffer from catastrophic overfitting, leading to improvements in accuracy at the 100th and 150th epochs, and eventually outperformed both the conventional data augmentation methods. When PGD perturbions were used in training, the model trained with clean and adversarial amplitude images (Clean & AA) did not suffer from robust overfitting. Although it started at a relatively low accuracy during the first 100 epochs, the model could pick up the adversarial features from the amplitude spectrum after the tuning of learning rates at the 100th and 150th epochs.

3.3. Extreme Cases

We trained WideResNet-40-2 on CIFAR-10 with an extremely strong adversarial attack. PGD-l_∞ with $\epsilon = 0.5, \alpha = 0.1, i_{\text{iters}} = 10$, which allows the adversary to change at most 50% of the target image, was used in training. These adversarial images can barely be recognized by humans. By swapping the amplitude spectra, the adversarial amplitude and adversarial phase images became recognizable (Figure 3). Table 4 shows the results. When adversarial images were used, the model struggled to learn adversarial features, and hence it failed to be robust against adversarial perturbations. In contrast, adversarial amplitude images, which contain only the adversarial amplitude spectrum, trained the model to be robust against both adversarial perturbations and common corruptions.

3.4. Adversarial Features of Images

We demonstrated the efficacy of the proposed method in training a robust CNN against both adversarial perturbations and common corruptions. Despite the ability to improve adversarial robustness, adversarial amplitude images did not serve as a strong attack compared to original adversarial images, as shown in Table 5. Nevertheless, the CNN classifiers trained with adversarial amplitude images showed comparable or, in some cases, even better adversarial robustness compared to the standard adversarial training.

Table 5. Classification error of classifiers trained with clean images on several types of adversarially perturbed images.

	Networks	Error Rates (%)			
		Clean	Adv	AA	AP
FGSM	ResNet-18	5.2	34.4	22.6	27.0
	WideResNet	6.0	33.2	21.5	26.3
PGD-l_∞	ResNet-18	5.2	99.9	37.8	76.4
	WideResNet	6.0	99.8	36.0	76.5

4. Conclusion

In this study, we have investigated the nature of the amplitude and phase spectra of adversarial images. We have proposed a frequency-based data augmentation method, in which the amplitude spectra are swapped between clean and adversarial images. We have demonstrated that training with adversarial amplitude images led to CNN classifiers that are robust against both adversarial perturbations and common corruptions. The experimental results have shown that the amplitude spectrum of adversarial images accommodates moderate and general perturbations that can help CNN classifiers to equip more general robustness. Despite having a weaker fooling ability, adversarial amplitude images served as better training images that helped CNN classifiers achieve more general robustness. We further demonstrated that with the proposed data augmentation method, CNNs can even learn from some extreme adversarial examples that humans can barely recognize. We believe that these findings will be crucial for the future development of truly robust CNN classifiers.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22H03658.

References

[1] Chen, G., Peng, P., Ma, L., Li, J., Du, L., Tian, Y.: Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 458–467 (2021) 1, 3

[2] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015) 3

[3] Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019) 1

[4] Leslie Rice, Eric Wong, Z.K.: Overfitting in adversarially robust deep learning. In: Proceedings of the 37th International Conference on Machine Learning (PMLR), pp. 8093–8104 (2020) 2

[5] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018) 3

[6] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2014) 1

[7] Tack, J., Yu, S., Jeong, J., Kim, M., Hwang, S.J., Shin, J.: Consistency regularization for adversarial robustness. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) *in press* (2022) 3

[8] Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8684–8694 (2020) 1

[9] Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020) 2

[10] Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 87.1–87.12 (2016) 3