Frequency Regularization for Improving Adversarial Robustness

Binxiao Huang, Chaofan Tao, Rui Lin, Ngai Wong

Department of Electrical and Electronic Engineering, The University of Hong Kong

{bxhuang, cftao, linrui, nwong}@eee.hku.hk

Abstract

Deep neural networks are incredibly vulnerable to crafted, human-imperceptible adversarial perturbations. Although adversarial training (AT) has proven to be an effective defense approach, we find that the AT-trained models heavily rely on the input low-frequency content for judgment, accounting for the low standard accuracy. To close the large gap between the standard and robust accuracies during AT, we investigate the frequency difference between clean and adversarial inputs, and propose a frequency regularization (FR) to align the output difference in the spectral domain. Besides, we find Stochastic Weight Averaging (SWA) (Izmailov et al. 2018), by smoothing the kernels over epochs[1], further improves the robustness. Among various defense schemes, our method achieves the strongest robustness against attacks by PGD-20, C&W and Autoattack, on a WideResNet trained on CIFAR-10 without any extra data.

Introduction

Deep neural networks (DNNs) have exhibited strong capabilities in various applications (He et al. 2016; Bin et al. 2019; Tao et al. 2020). However, research in adversarial learning shows that even well-trained DNNs are highly vulnerable to carefully crafted, human-imperceptible perturbations (Goodfellow, Shlens, and Szegedy 2014; Szegedy et al. 2013). Recently, various defense methods (Zhang et al. 2019; Wang et al. 2019; Wu, Xia, and Wang 2020) have been proposed to improve the robustness. Adversarial training (AT) (Madry et al. 2017), as a min-max saddle point problem, proves to be an effective and promising defense method without obfuscated gradients problems (Athalye, Carlini, and Wagner 2018). In the following, we denote the models obtained by natural training and AT as natural and robust models, respectively. For robust models, the accuracy achieved on natural and adversarial inputs are denoted as standard accuracy and robust accuracy, respectively. While AT improves robust accuracies, it generally sacrifices standard accuracies. Besides, frequency analysis (Wang et al. 2020) has been explored to yield new insights into DNNs. *In this work, we aim to answer the following questions using a frequency lens: 1) Why does AT reduce standard accuracy? and 2) how to improve the robustness by narrowing the gap between the standard and robust accuracies?*

To this end, we apply low-pass filtering (LPF) to the natural and adversarial inputs. Empirical results demonstrate that the robust model mainly relies on low-frequency content for prediction, which accounts for the low standard accuracy as high-frequency information is ignored. We also discover that the white-box attack can adapt its aggressive frequency distribution to the target model's frequency bias, thus explaining why white-box attacks are hard to defend. By visualizing the differences between the natural and adversarial inputs, we reveal that the differences are mainly concentrated in the low-frequency region. In order to close the accuracy gap, we propose a frequency regularization (FR) that aligns the outputs for natural and adversarial inputs in the frequency domain, leading to improvement in the robust accuracy. In addition, by observing that the robust model has a smoother kernel than its natural counterpart, we employ Stochastic Weight Averaging (SWA) (Izmailov et al. 2018) as a method of smoothing kernels over the training steps to further improve robustness.

To summarize, our work novelly adopts a frequency lens to: 1) explain the low standard accuracy of the robust model, and 2) propose a frequency-based regularization to significantly improve the robust accuracy.

Related works

Adversarial Defense. Among various defense methods that have been proposed to improve robustness (Szegedy et al. 2013; Madry et al. 2017), AT (Athalye, Carlini, and Wagner 2018) constitutes an effective and promising means. Typically, AT feeds adversarial inputs into a DNN to solve the following min-max optimization problem:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\left\|\mathbf{x}_i'-\mathbf{x}_i\right\|_p \leq \epsilon} \mathcal{L}\left(f_\theta\left(\mathbf{x}_i'\right), y_i\right), \tag{1}$$

where n is the number of training examples, $\mathbf{x}_i'$ is the adversarial input within the ϵ-ball (bounded by an L_p-norm) centered at the natural input $\mathbf{x}_i$, f_θ is the DNN with weight θ, $\mathcal{L}(\cdot)$ is the classification loss, e.g., cross-entropy (CE). Some recent results inspired by AT are also in place to further raise the robust accuracy: Zhang *et al.* (Zhang et al. 2019) identify a tradeoff between standard and robust accuracies that

[1] The final weight for evaluation is the average of the weights of multiple checkpoints during the training process.

Table 1: Top-1 accuracy(%) of natural and robust ResNet18 models trained on CIFAR-10. The LPF row denotes the filter bandwidths applied to the inputs. The higher the value, the more information is retained (i.e. 32 means no filtering).

model	LPF	32	28	24	20	16
Natural	Clean	94.56	92.92	90.75	80.7	50.72
	PGD-20	0.0	2.17	23.15	39.25	32.15
Robust	Clean	80.55	80.50	80.17	79.27	77.4
	PGD-20	51.81	52.10	52.22	52.54	52.52

serves as a guiding principle for designing the defenses. Wu *et al.* (Wu, Xia, and Wang 2020) identify that the weight loss landscape is closely related to the robust generalization gap and propose an effective Adversarial Weight Perturbation (AWP) method to overcome the robust overfitting problems (Rice, Wong, and Kolter 2020).

Learning in the Frequency Domain. Frequency analysis provides a new perspective on the generalization behavior of DNNs. In (Rippel, Snoek, and Adams 2015), spectral pooling is designed to preserve more information than regular spatial-domain pooling. Tao *et al.* (Tao et al. 2022) propose a frequency-aware plug-in to remove redundant information effectively for quantization. Wang *et al.* (Wang et al. 2020) claim CNNs could capture human-imperceptible high-frequency components of images for prediction, and smooth convolutional kernels are beneficial for robustness.

Analysis

Reason for Low Standard Accuracy. To explore the importance of high- and low-frequency information for models, we apply different LPFs to the natural and adversarial inputs that are fed into the natural or robust models to calculate the corresponding standard and robust accuracies. The results are shown in Table 1, e.g., a LPF bandwidth of 16 means after a Fast Fourier Transform (FFT), only the 16×16 patch in the center (viz. low frequencies) is preserved, and all external values are zeroed. As the bandwidth of LPF decreases, the information retained in the images also decreases. For the robust model, even though a large amount of high-frequency information is removed, there is only a negligible reduction in standard accuracy and less than a 1% improvement in robust accuracy. This indicates that the robust model focuses primarily on low-frequency content for predictions, and the adversarial inputs rely on low-frequency components to exercise its aggressiveness. Furthermore, the standard accuracy of the robust model ($\approx$80%) is similar to that of the natural model fed with natural inputs at LPF 20 (80.7%). Such observation indicates that *the low standard accuracy in the robust model is due to the under-utilization of high-frequency components*.

White-box Attack. In the natural model, as high-frequency information is removed, standard accuracy drops sharply. This suggests that the natural model employs high-frequency information to make classification judgment, which is consistent with the findings of (Wang et al. 2020). Moreover, the accuracy subject to adversarial inputs is improved at lower LPF bandwidths, reflecting that the adversarial inputs of the natural model exhibit aggressiveness in both the high- and low-frequency regions.

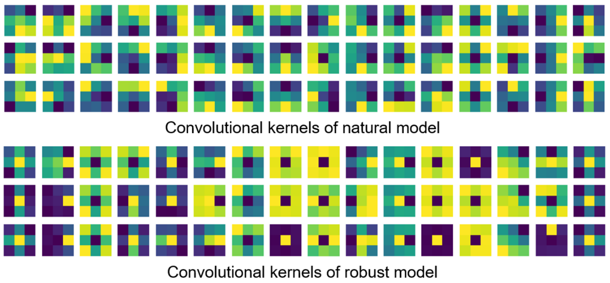
Convolutional kernels of natural model
Convolutional kernels of robust model

Figure 1: Visualization of first convolutional kernels (16 kernels each channel × 3 channels) of a natural (top) and robust (bottom) model. The latter has smoother kernels than the former.

For robust models that focus on the low-frequency information, the aggressiveness of the adversarial inputs is mainly concentrated in the low-frequency region. Whereas for natural models that utilize both high- and low-frequency information, the hostility is embedded in both high- and low-frequency regions. This suggests that the white-box attack can adapt its aggressive frequency distribution to the target model's frequency bias, thereby explaining why white-box attacks are so hard to defend.

Smooth Kernels. Wang *et al.* (Wang et al. 2020) introduce the concept of "smooth" kernel which has a smooth envelope on its spatial weights. If a kernel is smooth, it will see a reduced amount of high-frequency information. Along this line, it is articulated that smoothing the kernels' adjacent spatial values can help improve the adversarial robustness. Since the kernels of the first layer deal directly with the images, they can respond to the frequency bias of the information extracted from the images. In Figure 1, we visualize 16 randomly selected 3-channel kernels from the first layer of a natural (upper) and a robust model (lower), wherein the spatial size of each kernel is 3×3. The figure shows that the adjacent weights of kernels in the robust model change less dramatically, producing smoother kernels than the natural model counterparts. This implies that the robust model pays more attention to the low-frequency information, consistent with our previous discussion.

On the other hand, SWA (Izmailov et al. 2018), which averages the values of weights over time (epochs) along the natural training trajectory, proves to be an effective method to improve the generalization of the models. Here, we utilize SWA in AT as a method of *smoothing kernels* in the training time-axis dimension to mitigate the robust overfitting problems (Chen et al. 2020). Ablation study is shown in experiments to confirm the benefits of SWA in terms of robustness.

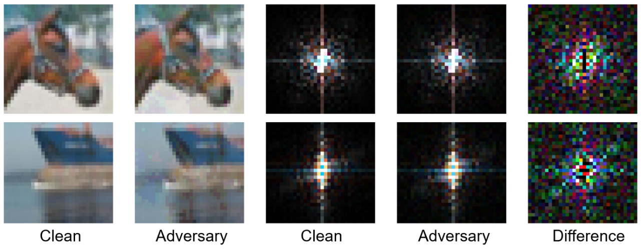

Figure 2: Visualization of natural and adversarial inputs in the spatial (left) and frequency domains (middle), and the absolute difference (right) after normalization in the frequency domain, with low frequency in the center and high frequency around. The brighter the pixel, the higher the frequency amplitude. The differences are mainly concentrated in the low-frequency region.

Frequency Regularization

To narrow the gap between standard and robust accuracies of the robust model, we need to identify the differences between natural and adversarial inputs. Figure 2 illustrates the natural and adversarial inputs in the spatial and spectral domains based on an adversarially trained model on the CIFAR-10 dataset. Because the adversarial inputs need to satisfy the l_∞ norm constraints, the changes in the spatial domain are rather small, and one can still recognize the horse and the ship before and after the perturbation. In the frequency domain, as shown in Figure 2, the differences between the natural and adversarial inputs are mainly distributed in the low-frequency region, with smaller amplitudes in the high-frequency region. Combined with the previous findings that robust models rely primarily on low-frequency information for prediction, it is easy to understand that the differences in the low-frequency region lead to a large accuracy gap. This further validates that, for the robust model, adversarial inputs rely mainly on low-frequency information to execute their aggressiveness.

Inspired by these findings, we propose that if a model can be trained to limit such frequency differences and achieve similar spectral domain outputs, then robust accuracy can be improved by approaching standard accuracy. To do so, we devise a simple yet effective frequency regularization (FR) to align the outputs for natural and adversarial inputs in the frequency domain. The optimization goal of the proposed AT with FR is:

$$\mathcal{L}_{AT} = \mathcal{L}_{CE} + \lambda \cdot \frac{1}{n}\sum_{i=1}^{n} \mathit{Dis}(\mathcal{F}(f(x_i)), \mathcal{F}(f(x_i'))), \quad (2)$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss, λ (defaulted at 0.1) denotes the FR coefficient, *Dis* denotes the distance function ($\mathcal{L}_1$ is used), and $\mathcal{F}$ denotes the Discrete Fourier Transform (DFT). The distance function is applied to the real and imaginary parts of the complex numbers after the DFT, respectively, and the results are summed. With FR, the robust accuracy against the PGD-20 attack on CIFAR-10 is substantially improved from 55.01% to 59.49%.

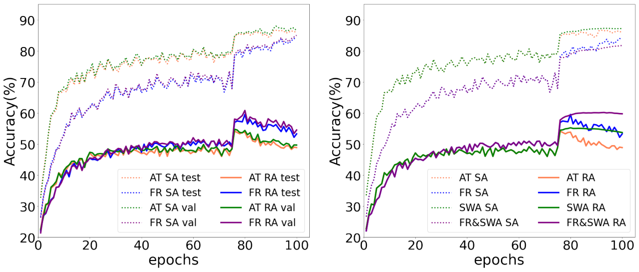

Figure 3: **Left**: Standard accuracy (dashed line) and robust accuracy (solid line) on validation and test sets over epochs for AT-trained WideResNet-34-10 on CIFAR-10. FR denotes frequency regularization, SA and RA denote the standard and robust accuracies; **Right**: Ablation studies to demonstrate the effect of FR and SWA on model performance.

Experiments

Experimental Settings. We take WideResNet-34-10 as a default model and adopt an SGD optimizer with a momentum of 0.9 and a global weight decay of 5×10^{-4}. The model is trained for 100 epochs with a batch size of 128 on one 3090 GPU. The initial learning rate is 0.1, decays to one-tenth at 75th and 90th epochs, respectively. All experiments are performed on the CIFAR-10 dataset, which contains 50k training (randomly split into a training set and a validation set at a 9:1 ratio) and 10k test examples. No extra data are used. We use PGD-10 AT as a standard training method. The robust accuracy of the PGD-20 attack equipped with random-start is taken as the main basis for robustness analysis. The attack step size is α = 2/255 and maximum l_∞ norm-bounded perturbation ϵ = 8/255. SWA is used since the first epoch where the learning rate drops and continues until the end with a cycle length 1.

Experimental Results. We evaluate the robust accuracy against several popular attack methods, including FGSM (Goodfellow, Shlens, and Szegedy 2014), PGD (Madry et al. 2017), C&W (Carlini and Wagner 2017) and AA (Croce and Hein 2020), shown in Table 2. Following the default setting of AT, the attack step size is 2/255, and the maximum l_∞ bounded perturbation is 8/255. The standard and robust accuracies are used as the evaluation metrics.

As shown in Figure 3, our method succeeds in closing the gap from 29.61% to 20.97% with a 5.11% improvement in robust accuracy against PGD-20 and a 3.53% drop in standard accuracy. This matches the generally accepted theory that there is a trade-off between standard and robust accuracies. The ablation experiments show that FR (59.49%) plays a major role in improving robust accuracy, while SWA (55.18%) is utilized here to alleviate the overfitting problem. The scheme that combines both of them achieves the best 60.12% and 54.35% robust accuracy against PGD-20 attack and Autoattack, respectively.

Table 2: Top-1 robust accuracy(%) of the WideResNet-34-10 model on the CIFAR-10. Bold numbers indicate the best.

Method	Clean	FGSM	PGD-20	C&W	AA
PGD-AT (Rice, Wong, and Kolter 2020)	84.62	60.17	55.01	53.32	51.42
TRADES (Zhang et al. 2019)	84.65	61.32	56.33	54.20	53.08
MART (Wang et al. 2019)	84.17	61.61	58.56	54.58	51.10
AWP (Wu, Xia, and Wang 2020)	85.57	**62.90**	58.14	55.96	54.04
AT-SWA	**86.17**	61.20	55.18	54.57	52.25
AT-FR(ours)	80.59	61.47	59.49	54.33	52.06
AT-FR-SWA(ours)	81.09	62.49	**60.12**	**56.14**	**54.35**

Conclusion

This work reveals that an adversarially trained model focuses primarily on low-frequency content for predictions, which accounts for the low standard accuracy due to underutilization of high-frequency information. To this end, we devise a frequency regularization to align the logits for natural and adversarial inputs in the spectral domain. SWA is adopted temporally to smooth the weights, improving the robustness further. Experiments show that the proposed method can substantially improve the robust accuracy. We further find that the white-box attack can adapt its aggressive frequency distribution to the target model's frequency bias, which explains why white-box attacks are hard to defend. It is believed these findings can shed light on the design of robust DNNs.

Acknowledgement

This work is supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under the General Research Fund (GRF) projects 17206020 and 17209721.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Bin, Y.; Yang, Y.; Tao, C.; Huang, Z.; Li, J.; and Shen, H. T. 2019. Mr-net: Exploiting mutual relation for visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8110–8117.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Chen, T.; Zhang, Z.; Liu, S.; Chang, S.; and Wang, Z. 2020. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2206–2216. PMLR.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D.; and Wilson, A. G. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Rice, L.; Wong, E.; and Kolter, Z. 2020. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 8093–8104. PMLR.

Rippel, O.; Snoek, J.; and Adams, R. P. 2015. Spectral Representations for Convolutional Neural Networks. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28, 2449–2457. Curran Associates, Inc.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tao, C.; Jiang, Q.; Duan, L.; and Luo, P. 2020. Dynamic and static context-aware lstm for multi-agent motion prediction. In *European Conference on Computer Vision*, 547–563. Springer.

Tao, C.; Lin, R.; Chen, Q.; Zhang, Z.; Luo, P.; and Wong, N. 2022. FAT: Frequency-Aware Transformation for Bridging Full-Precision and Low-Precision Deep Representations. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8684–8694.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*.

Wu, D.; Xia, S.-T.; and Wang, Y. 2020. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33: 2958–2969.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.