Appendix for *LEA*^{2}*: A Lightweight Ensemble Adversarial Attack via Non-overlapping Vulnerable Frequency Regions*

Yaguan Qian [*1], Shuke He[1], Jiaqiang Sha[1], Chenyu Zhao[1], Wei Wang[2], and Bin Wang[3]

[1]School of Science, Zhejiang University of Science and Technology, Hangzhou, China
[2]Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, China
[3]Zhejiang Key Laboratory of Multidimensional Perception Technology, Application and Cybersecurity, China

A. Preliminary

In this section, we first explain the Discrete Cosine Transform (DCT) used in our work and the specific location of each frequency band in the DCT block in Section A.1. Then, the description of the standard, robust, and weakly robust models are presented in Section A.2.

A.1. Discrete Cosine Transform (DCT)

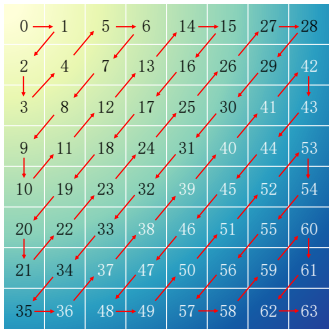


Figure 1. The standard 8×8 DCT block with all 64 frequencies arranged in zigzag order, the upper left corner and the lower right corner represent the lowest and highest frequency components in the DCT space, respectively.

Images are typically represented as pixel intensity values between 0 and 255 usually with multiple channels representing different colors. It is also possible to represent each channel of the image as a component of a set of frequencies. One way to convert pixel intensities to said representation is using the Discrete Cosine Transform (DCT) [9]. The DCT decomposes a signal into cosine wave components with different frequencies and amplitudes. More precisely, given a 2D image $x \in \mathbb{R}^{d\times d}$, define basis functions

$$\phi_d(i,j) = cos\left[\frac{\pi}{d}\left(i+\frac{1}{2}\right)j\right] \tag{1}$$

for $1 \le i,j \le d$. The DCT transform $V = \mathrm{D}(x)$ is:

$$V_{j_1,j_2} = N_{j_1}N_{j_2}\sum_{i_1=0}^{d-1}\sum_{i_2=0}^{d-1} x_{i_1,i_2}\phi_d(i_1,j_1)\,\phi_d(i_2,j_2) \tag{2}$$

where $N_j = \sqrt{\frac{1}{d}}$ if $j = 0$ and $N_j = \sqrt{\frac{2}{d}}$ otherwise. Here, N_{j_1}, N_{j_2} are normalization terms included to insure the transformation is isometric, *i.e.* $\|x\|_2 = \|\mathrm{D}(x)\|_2$. The entry $V_{i,j}$ corresponds to the magnitude of wave $\phi_d(i,j)$, with lower frequencies represented by lower i,j. Further, DCT is invertible, with inverse $x = \mathrm{D}^{-1}(V)$. For images containing multiple color channels, both DCT and IDCT can be applied channel-wise independently. In this work, The term *low-frequency* refers to frequency bands 0 to 9, *mid-frequency* refers to frequency bands 10 to 35, and *high-frequency* refer to frequency bands 36 to 63, as shown in Figure 1.

A.2. Robust, Weakly Robust, and Standard Models

Previous studies have only explored standard models and robust models, but they did not take into account the situation between two types of models. In order to comprehensively analyze the differences between various types of models from the frequency domain, we propose *weakly robust model*, which represents the models that have been adversarially trained but not converge. In this paper, for weakly robust models, we use the PGD attack [8] with $\epsilon = 4/255$ to train but not train to converge (20 epochs); for *robust models*, PGD, Mart [13] or Trades [17] with

*Corresponding author: qianyaguan@zust.edu.cn

Table 1. Attack success rates of white-box attacks, black-box attacks, and Gaussian noise on three standard models. For FGSM, PGD, TI-FGSM, and MI-FGSM, we set the maximum perturbation as $\epsilon = 4/255$ and for Gaussian noise $r \sim N\left(0, \sigma^2\right)$, we set $\sigma = 0.1$. The substitute model used for black-box attacks is the standard trained ResNet18.

Dataset	Target Model	White-Box			Black-Box			r
		FGSM [4]	PGD [8]	Average	TI-FGSM [2]	MI-FGSM [1]	Average	
CIFAR-10	ResNet20 [7]	66.14%	82.10%	74.12%	28.86%	83.11%	55.99%	73.78%
	WideResNet [16]	70.53%	100%	85.26%	29.98%	88.99%	59.49%	74.86%
	VGG16 [10]	44.56%	99.97%	72.26%	22.83%	62.10%	42.47%	72.18%
CIFAR-100	ResNet20 [7]	79.54%	91.07%	85.30%	46.56%	84.23%	45.40%	88.02%
	WideResNet [16]	61.20%	99.95%	80.57%	39.91%	92.09%	66.00%	70.89%
	VGG16 [10]	71.18%	99.82%	85.5%	38.53%	74.48%	56.51%	83.48%

Table 2. The accuracy of all the models used in our experiments on the testing sets of different datasets.

CIFAR-10		CIFAR-100		ImageNet-30	
Model	Accuracy	Model	Accuracy	Model	Accuracy
ResNet18	95.57%	ResNet18	76.70%	ResNet18	89.67%
ResNet20	94.60%	ResNet20	72.21%	ResNet50	89.40%
ResNet34	95.40%	ResNet34	74.09%	WideResNet101	90.32%
WideResNet	95.59%	WideResNet	71.72%	Densenet121	88.20%
VGG16	94.27%	VGG16	74.27%	VGG16	92.07%
PGD-ResNet18	82.81%	PGD-ResNet18	54.42%	PGD-ResNet18	82.33%
PGD-WideResNet	83.64%	PGD-WideResNet	47.45%	PGD-ResNet50	83.53%
Trades-ResNet18	83.28%	Trades-ResNet18	55.17%	Trades-ResNet18	83.20%
Mart-ResNet18	81.57%	Mart-ResNet18	47.81%	Mart-VGG16	80.33%
Weak-ResNet18	83.32%	Weak-ResNet18	58.14%	Weak-ResNet18	81.87%

$\epsilon = 8/255$ are used in training, and the models are trained to converge (50 epochs); for *standard models*, which means standard trained models. To further analyze the characteristics of these models in the frequency domain, we use the special frequency perturbations δ_f that are restricted to a specific frequency domain, where the maximum perturbation $\epsilon = 8/255$. The attack success rates when δ_f attacks standard models, weakly robust models, and robust models are visualized in Section 4.1.

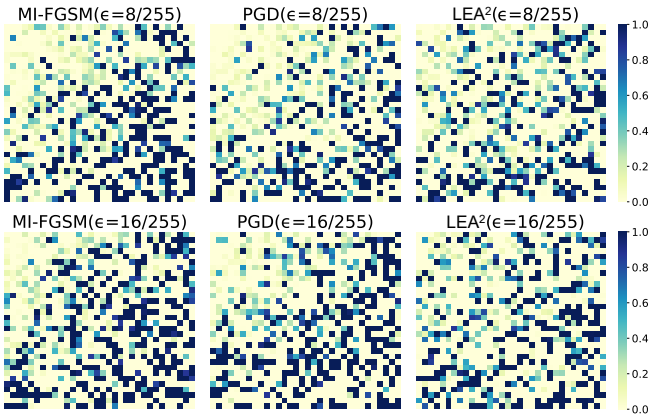


Figure 2. RCT map of various attacks on CIFAR-100 test sets. The upper left corner and lower right corner represent low and high frequency, respectively.

B. More Studies about Gaussian Noise

In order to explore whether Gaussian noise can achieve the same attack effect as the perturbations generated based on the standard model. We evaluated the effect of white-box attacks, black-box attacks, and Gaussian noise on three standard models—ResNet20 [7], WideResNet [16] and VGG16 [10]—which reach about 95% and 75% accuracy on CIFAR-10 and CIFAR-100, respectively. As shown in Table 1, the Gaussian noise can achieve a higher attack success rate than black-box attacks and is comparable to white-box attacks, which means that it's feasible that using Gaussian noise to replace the effect of vulnerable high-frequency regions (*i.e.*, $\mathcal{B}_{h_{standard}}$).

C. Experiment

C.1. Models

We list all the models used in our experiments here again for more friendly reading. And more details of these models are provided.

The accuracy of all the models used in our experiments are shown in Table 2. All models are trained using the SGD optimizer with Nesterov momentum 0.9 [12], weight decay 5×10^{-4}. We further employ cyclic learning rates, which can drastically reduce the number of epochs required for

Table 3. The success rate of various attacks on standard models with JPEG compression [6] on CIFAR-100. The best results are indicated in bold.

Dataset	Attack	ResNet20			VGG16		
		Clean	JPEG-75	JPEG-50	Clean	JPEG-75	JPEG-50
CIFAR-100	TI-FGSM [2]	76.47%	64.53%	47.37%	69.42%	60.74%	57.71%
	MI-FGSM [1]	93.15%	76.90%	36.46%	92.17%	72.89%	48.89%
	DI-FGSM [15]	96.43%	80.31%	55.70%	95.75%	81.75%	65.45%
	MI-FGSMens [1]	95.77%	82.19%	54.47%	95.08%	82.02%	64.88%
	DI-FGSMens [15]	**98.11%**	**87.71%**	70.76%	**98.03%**	**89.15%**	78.14%
	LEA^2(ours)	90.48%	85.47%	**77.60%**	84.87%	84.57%	**82.40%**

Table 4. The attack success rate of various attacks on advanced defense models. The best results are indicated in bold.

Dataset	Attack	AT	Trades	JPEG-75	JPEG-50	TVM	FS	Spatial Smoothing
ImageNet-30	FGSM [4]	32.69%	28.23%	33.11%	32.74%	32.64%	33.63%	33.06%
	PGD [8]	48.87%	32.81%	49.47%	48.31%	30.39%	49.68%	42.94%
	TI-FGSM [2]	43.16%	37.31%	43.03%	43.38%	**48.78%**	43.18%	46.21%
	MI-FGSM [1]	46.06%	34.02%	45.93%	45.56%	35.73%	46.46%	44.25%
	DI-FGSM [15]	48.23%	36.43%	48.18%	48.31%	39.21%	47.83%	47.10%
	MI-FGSMens [1]	27.28%	24.42%	36.60%	26.05%	21.45%	28.14%	25.67%
	DI-FGSMens [15]	27.08%	25.71%	30.55%	20.08%	24.35%	32.15%	27.56%
	LA [5]	41.96%	33.16%	42.55%	42.66%	37.06%	42.37%	44.81%
	LEA^2(ours)	**59.97%**	**39.81%**	**59.39%**	**58.79%**	46.06%	**59.88%**	**55.92%**

training deep networks [11]. A simple cyclic learning rate schedules the learning rate linearly from zero, to a maximum learning rate, and back down to zero and allows architectures to converge to the benchmark accuracy in tens of epochs instead of hundreds [14]. For PGD [8], Trades [17], and Mart [13] adversarial training with $\epsilon = 8/255$, we set the number of epochs as 50, batch size as 128, and maximum learning rate as 0.2 for CIFAR-10 and CIFAR-100. For ImageNet-30, we set the number of epochs as 50, batch size as 64, and maximum learning rate as 0.03. For weakly robust models, the training attack is PGD with random start, we set perturbation budget as $\epsilon = 4/255$, step size $\alpha = 2/255$, and the number of epochs as 20.

C.2. Evaluation Metrics

We use (1) attack success rate (ASR) to evaluate the attack performance of adversarial examples, (2) l_2 norm to measure the perturbation amplitude, and (3) the structural similarity (SSIM) index as a measurement of the similarity between original images and adversarial examples. The formula is as follows:

$$SSIM\left(x, x'\right) = \left[l\left(x, x'\right)\right]^{\alpha} \left[c\left(x, x'\right)\right]^{\beta} \left[s\left(x, x'\right)\right]^{\gamma} \quad (3)$$

where $\alpha, \beta, \gamma > 0$, $l\left(x, x'\right)$ is brightness comparison, $c\left(x, x'\right)$ is contrast comparison,and $s\left(x, x'\right)$ is structure comparison:

$$l\left(x, x'\right) = \frac{2\mu_x \mu_{x'} + c_1}{\mu_x^2 + \mu_{x'}^2 + c_1} \quad (4)$$

$$c\left(x, x'\right) = \frac{2\sigma_{xx'} + c_2}{\sigma_x^2 + \sigma_{x'}^2 + c_2} \quad (5)$$

$$s\left(x, x'\right) = \frac{\sigma_{xx'} + c_3}{\sigma_x \sigma_{x'} + c_3} \quad (6)$$

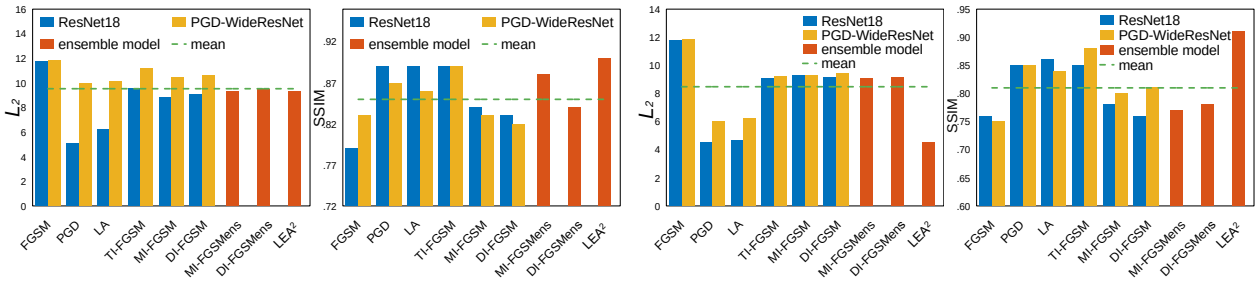
where μ_x and $\mu_{x'}$ represent the average of x and x' respectively, σ_x and $\sigma_{x'}$ represent the standard deviation of x and x' respectively, $\sigma_{xx'}$ represents the covariance of x and x', and c_1, c_2 and c_3 are constants.

C.3. RCT map of various attacks on CIFAR-100

Here, we show more studies of perturbations generated by the black-box MI-FGSM attack, white-box PGD attack, and our attack LEA^2 in the frequency domain on the CIFAR-100 test dataset, as shown in Figure 2. The perturbations generated by LEA^2 are distributed throughout the entire frequency region. In contrast, the perturbations generated by MI-FGSM and PGD are more concentrated in the high-frequency regions, and almost no perturbation is generated in the low-frequency regions. This is consistent with the conclusion in Section 4.3.

C.4. More Experiments

We also test the performance of various black-box attacks on CIFAR-100 and ImageNet. We first test the performance of various attacks on the standard models with JPEG defense [3] (see Table 3). Then, we conduct experiments



(a) CIFAR-10 (b) CIFAR-100

Figure 3. Perceptual similarity measurement of nine adversarial attacks on CIFAR-10 and CIFAR-100. The blue color denotes adversarial examples generated based on standard model, the yellow color denotes adversarial examples generated based on robust model, the orange color denotes ensemble attacks, and the green dotted line denotes the mean of the l_2 distance or SSIM of the first six attacks.

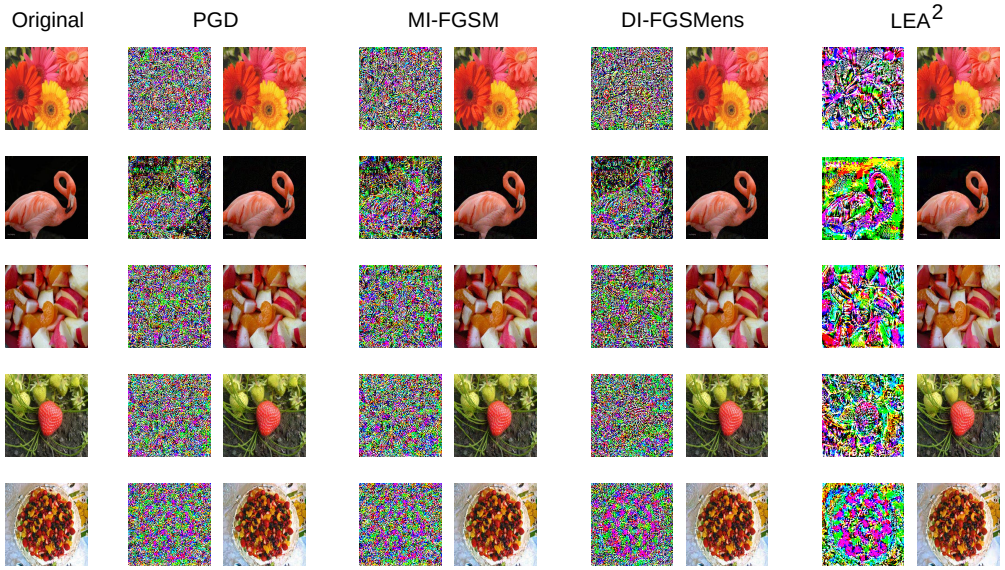


Figure 4. Adversarial examples generated by different attacks approaches on ImageNet-30. The maximum perturbation for all attack methods is $\epsilon = 8/255$.

to test the transferability of the various adversarial attacks on the advanced defended models as shown in Table 4. We can see that our attack LEA^2 achieves better transferability under different defended models compared with extensive baselines and state-of-the-art attacks.

C.5. Perception Study

An important characteristic of adversarial examples is that they are invisible to humans. In order to further confirm that adversarial examples generated by LEA^2 are not easy to be perceived by humans, we use conventional average l_2 distortion and structural similarity index SSIM to evaluate the imperceptibility of LEA^2 and compare it with the advanced white-box attacks and black-box attacks, as shown in Figure 3.

In order to comprehensively analyze the imperceptibility of adversarial examples, for first-order attacks based on a single model, we explored the l_2 norm and SSIM of adversarial examples generated by standard trained ResNet18 and PGD-WideResNet respectively. Then the substitute models used for ensemble attacks MI-FGSMens, DI-FGSMens, and LEA^2 are the same as those described in Section 5.1. l_2 norm is used to measure the move distance of an adversarial example from its original example, the smaller l_2, the lower the distortion rate of adversarial examples. The SSIM is used to measure the similarity between adversarial examples and original images, the larger SSIM means that adversarial examples are more similar to original examples.

As shown in Figure 3, the first and third columns show the l_2 distance between adversarial examples and original examples. As can be observed, adversarial examples generated based on the robust model generally have a higher distortion rate than those generated based on the standard model, which is consistent with the analysis in Section 4.1.

Adversarial perturbations generated based on the robust model are mainly added to low-frequency domains, whereas those generated based on the standard model are primarily located on high-frequency domains, and the changes in the high-frequency domains are more invisible to humans. The l_2 distance between the adversarial examples crafted by LEA^2 and their original images is below the mean value of l_2 of the first six black-box and white-box attacks, that is to say, LEA^2 does not produce obvious distortion of adversarial examples. The second and fourth columns are the structural similarity (SSIM) between various adversarial examples and original examples. Our method LEA^2 has the highest SSIM, which means that our adversarial examples are similar to the original examples. In addition, Figure 4 shows the adversarial examples generated by the white-box attacks, black-box attacks based on a single model, the ensemble attack, and our method LEA^2 on ImageNet-30.

References

[1] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018.

[2] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4312–4321. Computer Vision Foundation / IEEE, 2019.

[3] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *CoRR*, abs/1608.00853, 2016.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[5] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 1127–1137. AUAI Press, 2019.

[6] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[9] Kamisetty Ramamohan Rao and Patrick C. Yip. *Discrete Cosine Transform - Algorithms, Advantages, Applications*. 1990.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[11] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120, 2017.

[12] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147. JMLR.org, 2013.

[13] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[14] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[15] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2730–2739. Computer Vision Foundation / IEEE, 2019.

[16] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[17] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.