Supplementary Materials for Paper “Defending against Universal Adversarial Patches by Clipping Feature Norms”

Cheng Yu[1,3], Jiansheng Chen[1,2,3]*, Youze Xue[1], Yuyang Liu[1], Weitao Wan[1], Jiayu Bao[1], and Huimin Ma[3]
[1]Department of Electronic Engineering, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology, China
[3]University of Science and Technology Beijing, China

1. Analysis of the Decay Rate of $\|\mathbf{f}^*_{i,j}\|$

According to the analysis in Section 3.2.2 as well as the translation invariance of the Hadamard product, $\|\mathbf{f}^*_{i,j}\|$ can be formulated as Eq. 1, where ERF is a 2D Gaussian function.

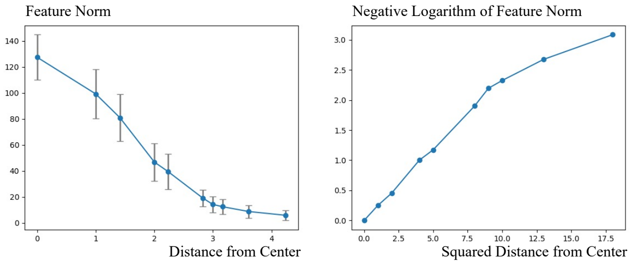
$$\begin{aligned}\|\mathbf{f}^*_{i,j}\| &\propto \sum tr_{i,j}(\text{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}}) \\ &= \sum (\text{ERF}) \odot tr_{-i,-j}(\hat{\mathbf{x}} - \tilde{\mathbf{x}}) \end{aligned} \tag{1}$$

Similar to Section 3.2.3, we consider one spatial dimension and one channel for ERF and $(\hat{\mathbf{x}} - \tilde{\mathbf{x}})$ without loss of generality, because any 2D Gaussian function with the covariance matrix of rank r can be regarded as a linear combination of r multiplications of two 1D Gaussian functions. As such, denote σ as the variance of the ERF. The lower and upper boundary for the mask of the patch $\mathbf{M}$ is denoted as $A - B$ and $A + B$ respectively, where A is the center of the patch. Then, ERF and $(\hat{\mathbf{x}} - \tilde{\mathbf{x}})$ can be formulated as Eq. 2.

$$\begin{aligned}\text{ERF} &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{i^2}{2\sigma^2}) \\ (\hat{\mathbf{x}} - \tilde{\mathbf{x}}) &= \sum_{m=-B}^{B} x^*_{A+m}\delta(i - A - m) \end{aligned} \tag{2}$$

As such, we can get $\|\mathbf{f}^*_i\|$ as Eq. 3. It can be observed that $\|\mathbf{f}^*_i\|$ is spatially distributed as a weighted sum of Gaussian functions of the same variance σ but different centers $A+m$.

$$\begin{aligned}\|\mathbf{f}^*_i\| &= \sum (\text{ERF}) \odot tr_{-i}(\hat{\mathbf{x}} - \tilde{\mathbf{x}}) \\ &\propto \sum_{m=-B}^{B} x^*_{A+m}\exp(-\frac{(i - A - m)^2}{2\sigma^2}) \end{aligned} \tag{3}$$



(a) (b)

Figure 1. **(a) Curve of the mean value and standard deviation of $\|\mathbf{f}^*_{i,j}\|$ with the relative location $\|(i,j) - \mathbf{A}\|$ of the adversarial patch**, which is the result of normalizing Fig. 2 (b) of our paper to the input size. **(b) Curve of $-\log\|\mathbf{f}^*_{i,j}\|$ with $\|(i,j) - \mathbf{A}\|^2$**, which is an illustration of DR.

Then we compare the decay rate of $\|\mathbf{f}^*_i\|$ from center A with the Gaussian functions. Formally, we calculate the decay rate as Eq. 4. Thus, the corresponding infinite signal of $\|\mathbf{f}^*_i\|$ has the same squared exponential decay rate as the Gaussian functions when $|i - A|/B \to +\infty$. This means that if the size of the image is far larger than the size of the patch, $\|\mathbf{f}^*_i\|$ decays like Gaussian in the region of the image.

$$\begin{aligned}\text{DR} = &\lim_{\frac{|i-A|}{B}\to+\infty} \frac{\log\|\mathbf{f}^*_i\|}{(i-A)^2} \\ = &\lim_{\frac{|i-A|}{B}\to+\infty} \frac{1}{(i-A)^2}(\log(\sum_{m=-B}^{B} x^*_{A+m}\exp(\\ &-\frac{(i-A)^2 + m^2 - 2m(i-A)}{2\sigma^2})) + Const) \\ = &-\frac{1}{2\sigma^2} + \lim_{\frac{|i-A|}{B}\to+\infty} \frac{1}{(i-A)^2}\log(\sum_{m=-B}^{B} x^*_{A+m} \\ &\exp(\frac{m^2}{2\sigma^2})\exp(\frac{m(i-A)}{\sigma^2})) = -\frac{1}{2\sigma^2} \end{aligned} \tag{4}$$

We extend the experiments in Section 3.2.2 to show the decay rate as in Fig. 1. It can be observed that in a con-

---

*Corresponding author.

siderable range of distance, $-\log \|\mathbf{f}^*_{i,j}\|$ and $\|(i,j)-\mathbf{A}\|^2$ is approximately linear, validating that $\|\mathbf{f}^*_{i,j}\|$ has a squared exponential decay rate. But the decay rate slows down when the distance becomes quite large. It is probably because the $\|\mathbf{f}^*_{i,j}\|$ becomes too small, making its value vulnerable to accidental errors.

2. Detailed Deduction to $o^{-1}[i]$

According to the definition of the z-transformation [2], the system function $O^{-1}(z)$ for the inverse system can be formulated as Eq. 5, where S is the size of the convolution kernel, $-B = \text{argmin}_m w[m] \neq 0$ is the lower boundary of the convolution kernel in space, and $\{v_s\}_{s=1}^{S}$ are the poles of $O^{-1}(z)$ which is the set of real numbers and conjugate complex number pairs.

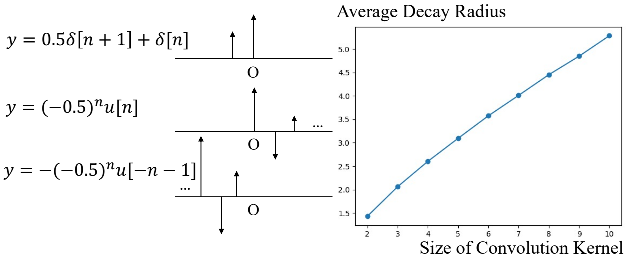
$$O^{-1}(z) = \frac{1}{O(z)} = \frac{1}{\sum_m w[m]z^{-m}} = \frac{\frac{1}{w[-B]}z^{-B}}{\prod_{s=1}^{S}(1-v_s z^{-1})}, \tag{5}$$

We consider that there are no high order poles in the following deduction which generally holds, and discuss the high order poles afterward. Since $O^{-1}(z)$ is a rational proper fraction of z^{-1}, it can be expanded in a partial fraction expansion as is shown in Eq. 6.

$$\begin{aligned} O^{-1}(z) &= \sum_{s=1}^{S} q_s \frac{1}{1-v_s z^{-1}} = \sum_{s=1}^{S} q_s O_s(z), \\ \text{where } q_s &= O^{-1}(z)(1-v_s z^{-1})|_{z=v_s}. \end{aligned} \tag{6}$$

So $o^{-1}[i]$ can be regarded as the linear combination of unit impulse responses of system functions with each pole. For infinite bilateral signals, there are two possibilities for the unit impulse response of system function $O_s(z) = 1/(1-v_s z^{-1})$. One is right-sided sequence $o_s[i] = {v_s}^i u[i]$, while the other is left-sided sequence $o_s[i] = -{v_s}^i u[-i-1]$. The difference of them is that they have different region of convergence (ROC) in z-plane, with $|z| > |v_s|$ for the right-sided sequence and $|z| < |v_s|$ for the left-sided sequence.

Notice that not both of the possible unit impulse responses $o_s[i]$ are suitable for being the part of the inverse system of a convolution layer. Obviously, the output of the inverse system is the input features of the convolution layer, so the system should be Bounded-In-Bounded-Out (BIBO) stable, which requires the ROC including the unit circle. Seen from another perspective, the actual feature map is finite, so when $o_s[i]$ is divergent, it cannot satisfy the boundary conditions. Thus, for $|v_s| \leq 1$, the right-sided sequence is chosen, while for $|v_s| > 1$, we choose the left-sided sequence, as is shown in Fig.2 (a). As a result, each BIBO stable $o_s[i]$ has a unilateral exponential decay, and $o^{-1}[i]$ is their linear combination as is in Eq. 7. For poles with order j, the possible unit impulse responses are the multiplication of the possible unit impulse responses of the first order pole and a polynomial with order $j-1$. So the BIBO stable $o_s[i]$ has the same unilateral exponential decay as that of first order poles.



(a) (b)

Figure 2. **(a) An example of the unit impulse response of a 1D convolution system as well as its possible inverse systems.** From top to bottom are the original convolution system, the BIBO stable inverse system, and the unstable inverse system. **(b) The curve of the estimated average decay radius of the inverse system with the size of the original 1D convolution kernel.**

$$\begin{aligned} o^{-1}[i] &= \sum_{s=1}^{S} q_s o_s[i], \\ o_s[i] &= \begin{cases} {v_s}^i u[i] & |v_s| \leq 1 \\ -{v_s}^i u[-i-1] & |v_s| > 1 \end{cases} \\ \text{s.t. } O^{-1}(z) &= \frac{1}{\sum_m w[m]z^{-m}} = \sum_{s=1}^{S} q_s \frac{1}{1-v_s z^{-1}} \end{aligned} \tag{7}$$

We further study the decay radius with the size of the convolution layer S using the Monte Carlo algorithm, where each $w[m]$ is randomly chosen from a standard normal distribution. Fig.2 (b) illustrates the results. It can be observed that the estimated decay radius is approximately linear with the size of convolution, which means that the inverse system of a convolution layer has a similar locality to the original layer.

3. Comparison between Gray-box and White-box Attack

In our paper, all the experimental results are reported under the pure **white-box** adversarial attack, which takes the defending approaches into consideration when training the adversarial patch. But the experiments in the original papers of some previous works *e.g.* DW and LGS are tested under the gray-box attack, where the adversarial patch is generated to attack the original model. We have analyzed why we choose the pure white-box attack in our paper, and here we compare the performance of different defending methods under gray-box and white-box LaVAN attacks on ImageNet in Table 1.

Table 1. Clean and adversarial Acc for different defending methods evaluated on different target models gray-box and white-box LaVAN attacks on ImageNet. The best results are marked as **bold**.

	ResNet-50			Inception-V3			MobileNet-V2		
Defense	Clean	Gray	White	Clean	Gray	White	Clean	Gray	White
PatchGuard Window=4	67.0%	34.1%	31.6%	74.8%	28.8%	23.3%	63.0%	31.5%	27.4%
DW	42.4%	63.8%	32.7%	35.6%	67.7%	30.2%	38.0%	50.4%	23.1%
LGS	69.8%	58.7%	0.6%	**75.0%**	**70.5%**	1.7%	65.3%	55.6%	11.5%
Ours $\alpha = 1.0$	72.4%	65.0%	58.3%	71.6%	67.5%	58.8%	64.1%	56.2%	**52.0%**
Ours $\alpha = 1.1$	**73.3%**	**66.7%**	**59.5%**	74.3%	69.4%	**59.0%**	**65.5%**	**56.7%**	48.9%

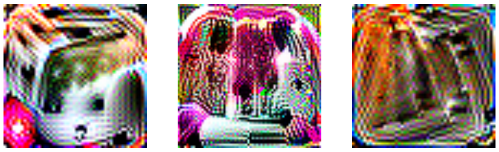


Figure 3. **Adversarial patches targeting ImageNet class 'toaster'.** From left to right are patches generated by LaVAN on ResNet-50 without defense, Inception-V3 using LGS and MobileNet-V2 with FNC respectively.

It can be observed in Table 1 that the performance of our method as well as PatchGuard keeps stable under gray-box and white-box attack. The difference between our method and PatchGuard is that our method performs significantly better than PatchGuard in both situations. For comparison, the performance of the detection-based DW and LGS drops dramatically from the gray-box attack to white-box attack, revealing that they only focus on the difference between the patch generated by one specific kind of attack and the benign regions, which is not essential for the success of the attack. Similar results have also been reached in previous researches [1]. Interestingly, the adversarial Acc of DW under gray-box attack is even significantly higher than the clean Acc. It is because the adversarial patch generated by gray-box attack is more salient than the benign salient regions, leading to the latter not being detected and thus improving the accuracy. However, the difference in salience is not essential for distinguishing the patch and the benign salient regions, so DW is far less effective under white-box attack.

We then elaborate the attack method in testing the effectiveness of different defending methods as a supplement to the experimental setup in our paper. The original Adversarial Patch (AdvP) as well as the LaVAN method are used to generate adversarial examples and lead to similar results. The patches are applied on random location of the image and takes up 5% of total image pixels on ImageNet and 10% on CIFAR10. The target class is $toaster$ for ImageNet and dog for CIFAR10 without instruction. Fig. 3 shows some examples of the generated patch on ImageNet.

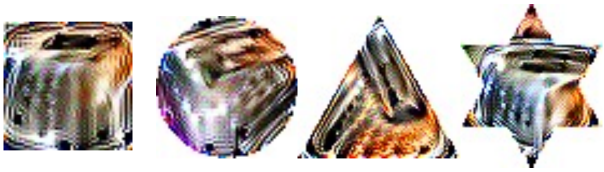


Figure 4. **Adversarial patches of different shapes targeting ImageNet class 'toaster'.**

4. Patches of Other Shapes

We have verified the effectiveness of our method against rectangle patches of different aspect ratios in Section 4, then we generate adversarial patches of other shapes and test the performance of our method against them. Circular patch, triangular patch and star-shaped patch are adopted as is shown in Fig. 4. LaVAN attack with the 5% patch for ResNet-50 on ImageNet is employed here.

Table 2. Adversarial Accs for patches of different shapes.

Shape	Square	Circle	Triangle	Star
Acc	58.3%	59.2%	62.4%	57.3%

As is shown in Table 2, the adversarial Acc keeps stable across different shapes of adversarial patches with a variance of 5.1%, indicating the robustness of our method against different shapes of the patch.

5. More Visualization Results of Feature Norm Maps

We provide more visualization results of feature norm maps here to illustrate the impact on feature maps of our method. Firstly, we compare the norm of feature vectors in intermediate feature maps for adversarial examples on models with/without FNC. We choose the same clean images with adversarial patches on the same locations for a better comparison. The adversarial patches are generated by the white-box LaVAN attack targeting ResNet-50 with/without FNC respectively on ImageNet.

The results for ResNet-50 with/without FNC are illustrated in Fig. 5 and Fig. 6 respectively, where the feature norm maps on the same layer of different models are normalized to the same scale. It can be observed from Fig. 5

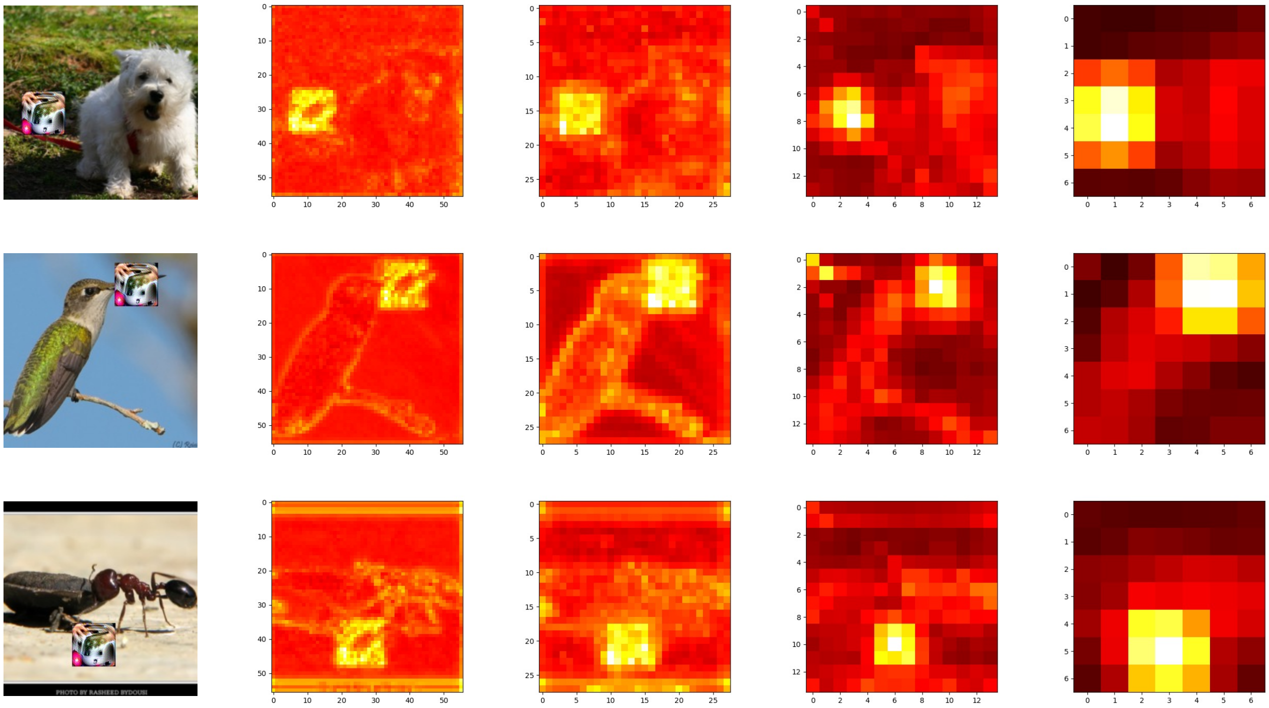


Figure 5. **Feature norm maps of ResNet-50 without FNC.** From left to right are the input images, feature norm maps on Conv2-3, Conv3-4, Conv4-6 and Conv5-3.

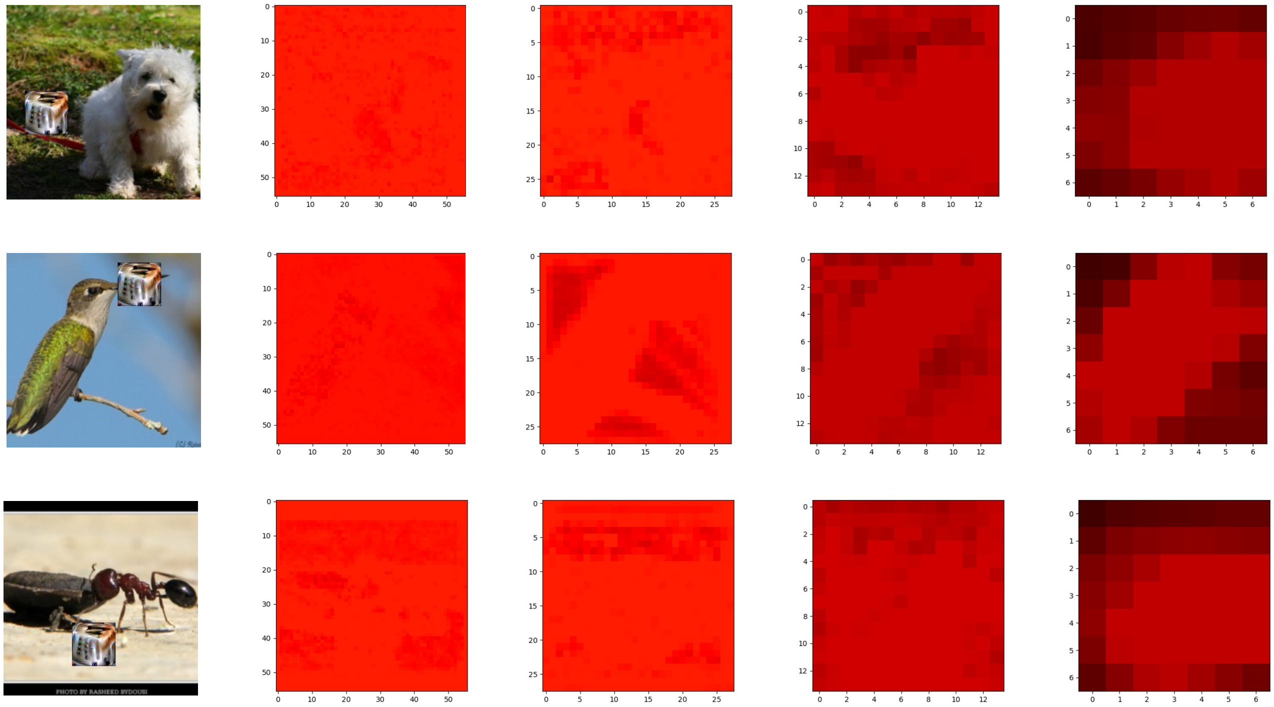


Figure 6. **Feature norm maps of ResNet-50 with FNC.** From left to right are the input images, feature norm maps on Conv2-3, Conv3-4, Conv4-6 and Conv5-3 after FNC. The suppression of feature norms by FNC is shown compared to Fig. 5.

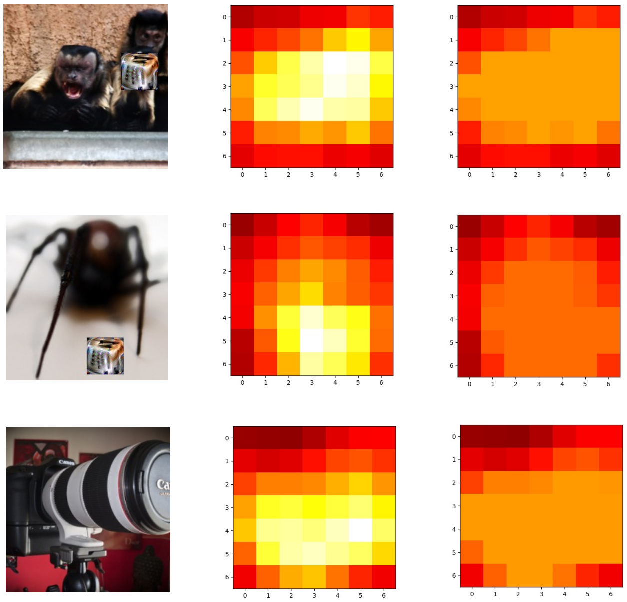


Figure 7. **Visualizations of the FFM norm map before/after FNC for clean and adversarial images on ResNet-50 with FNC.** The input image in first two rows are adversarial images, while the input of the third row is a clean image.

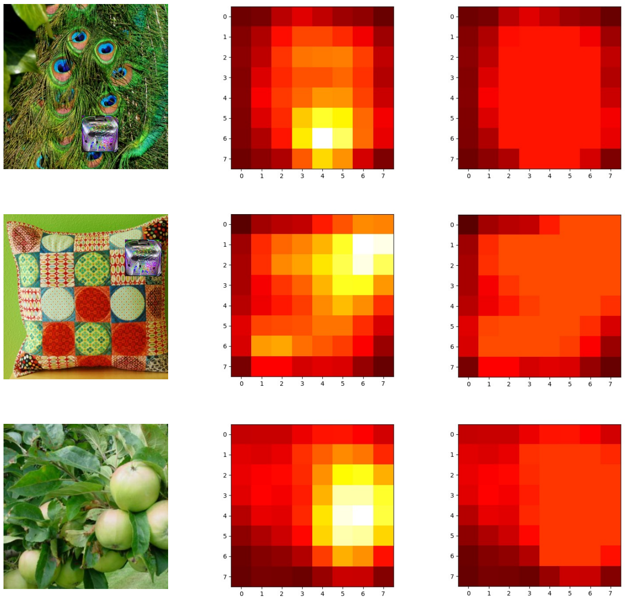


Figure 8. **Visualizations of the FFM norm map before/after FNC for clean and adversarial images on Inception-V3 with FNC.** The input image in first two rows are adversarial images, while the input of the third row is a clean image.

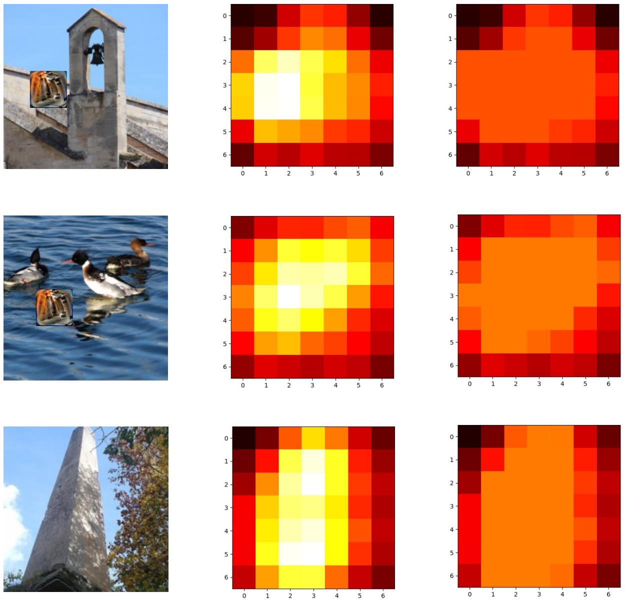


Figure 9. **Visualizations of the FFM norm map before/after FNC for clean and adversarial images on MobileNet-V2 with FNC.** The input image in first two rows are adversarial images, while the input of the third row is a clean image.

that the model without FNC focuses more on the features of the patch since they propagate on very shallow layers. The relative norm of feature vectors on the corresponding location of the patch becomes larger and larger in forward propagation, and eventually dominates in classification. However in the model with FNC, as is shown in Fig. 6, the norm of the feature vectors on different locations in shallow layers are more similar to each other, and the model focuses more on the benign object in deep layers. As such, the impact of the patch on feature maps is suppressed, leading to correct classification results.

Furthermore, we extend the visualizations of the FFM norm map before/after FNC for clean and adversarial images on model with FNC to different CNN architectures. Fig. 7, Fig. 8 and Fig. 9 are the results on ResNet-50, Inception-V3 and MobileNet-V2 respectively. It can be observed from the FFM norm maps that FNCs successfully suppress the large norm feature vectors in both clean and adversarial images and decrease the variance of the norm of feature vectors, as is analyzed in Section 4.

References

[1] Ping Yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. In *International Conference on Learning Representations*, 2020.

[2] Alan V Oppenheim, John R Buck, and Ronald W Schafer. *Discrete-time signal processing. Vol. 2*. Upper Saddle River, NJ: Prentice Hall, 2001.