Supplementary material for

How secure are the adversarial examples themselves?

[1]Hui Zeng, [1]Kang Deng, [2]Biwei Chen, and [1]Anjie Peng

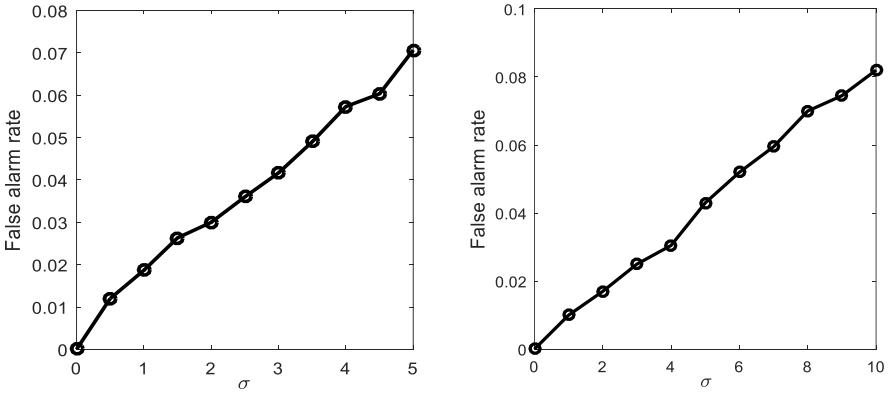
[1]School of Computer Science and Technology, Southwest University of Science and Technology, 621010, China
[2]Department of Business and Economics, Houghton College, NY 14744, USA
penganjie200012@163.com

Due to the page limitation, we provide more detail experimental results in this document. Following are the description:

In Fig.1, we show the false alarm rate of $\delta^1()$ in detecting different attacks as a function of σ. In our implementation, for both BIM and PGD, their attack target is a pre-trained Resnet18 model [1]. Hence, they share the same curve. The attack target of for C&W is a pre-trained Inception v3 model [2].

In Fig. 2, we show the ROC performance of the two single tests on three attacks.

In Fig. 3, we show the P_d matrixes when $P_{fa} = 0.04$ for three attacks.



(a) (b)

Fig.1. The false alarm rate of the noise addition-then-denoising detection as a function of σ. (a) In detecting the BIM/PGD attack, for which the target CNN model is ResNet18 [1], (b) in detecting the C&W L_2 attack, for which the target CNN model is Inception v3 [2].

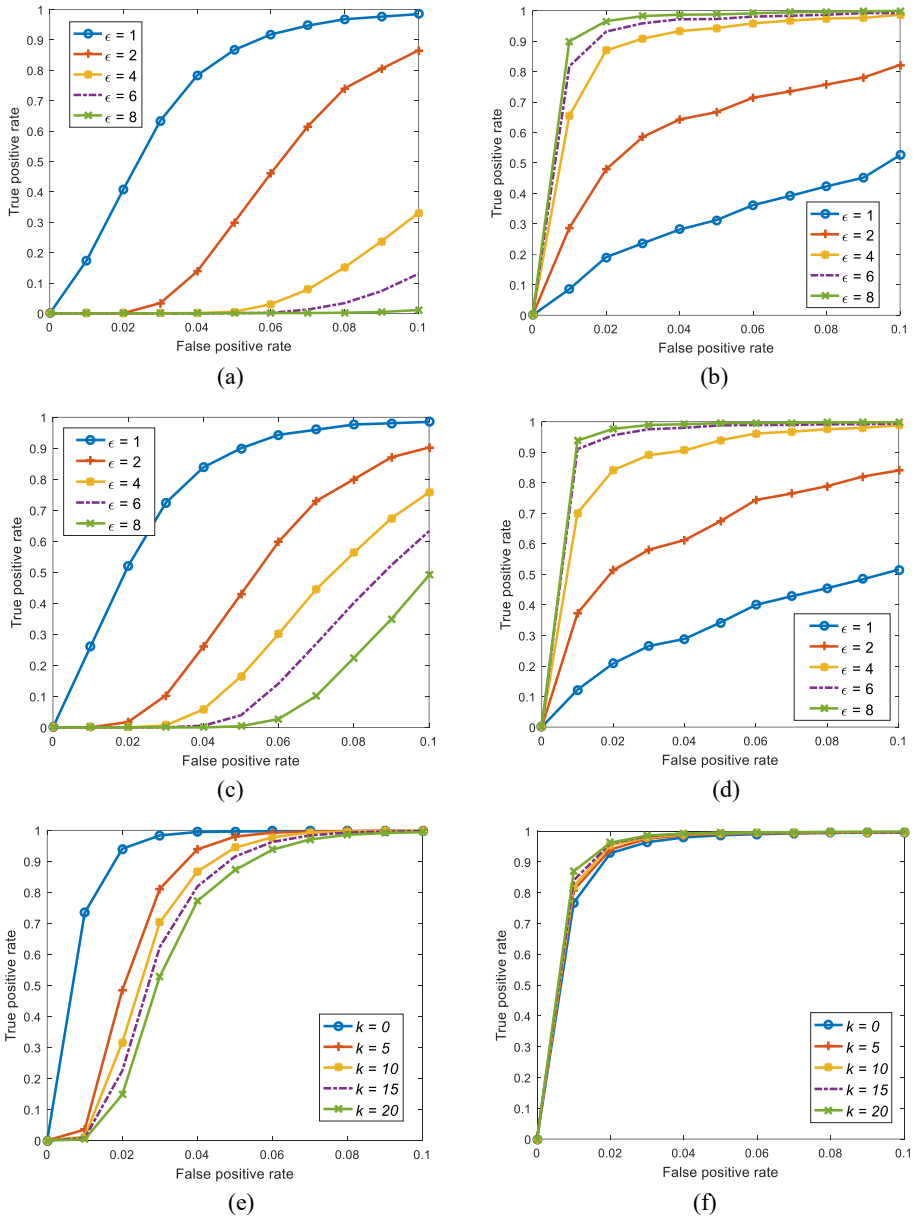


Fig.2. The ROC performance of the two single tests on different attacks. (a) BIM attack, noise addition-then-denoising test, (b) BIM attack, SRM based test, (c) PGD attack, noise addition-then-denoising test, (d) PGD attack, SRM based test, (e) C&W L_2 attack, noise addition-then-denoising test, (f) C&W L_2 attack, SRM based test,

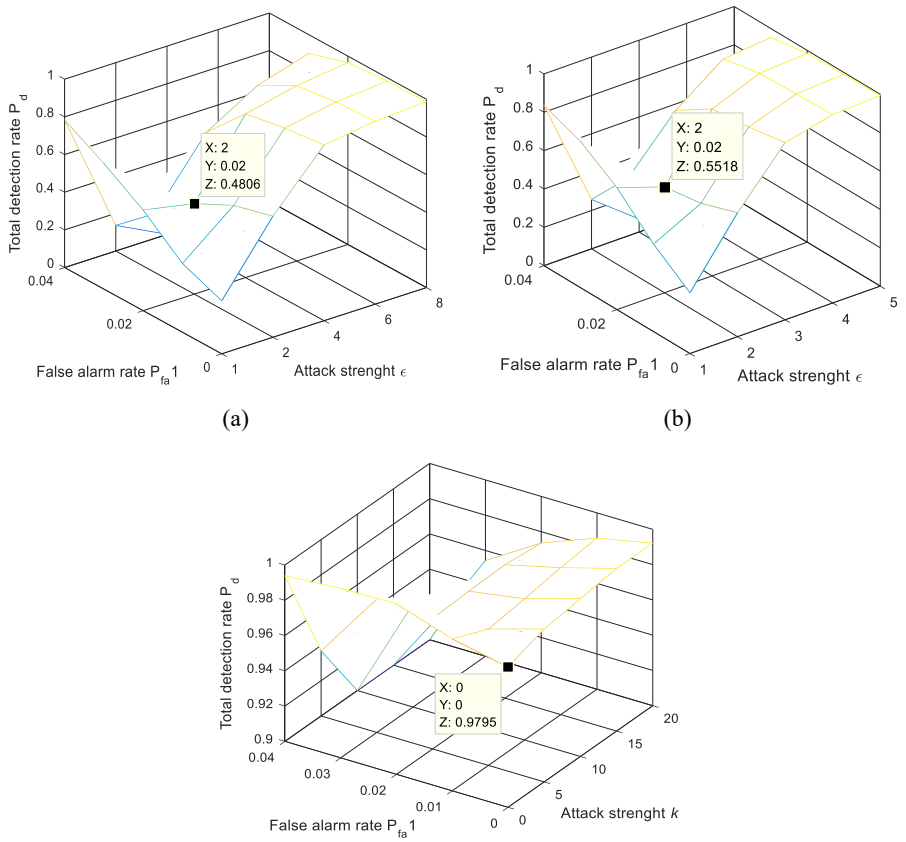


Fig.3. P_d matrix when $P_{fa} = 0.04$. (a) BIM attack, (b) PGD attack, (b) C&W L_2 attack.

References

1. He, K., Zhang, X., Ren, S., et al: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
2. Szegedy, C., Vanhoucke, V., Ioffe, S., et al: Rethinking the Inception architecture for computer vision. arXiv preprint arXiv: 1512.00567 (2015).