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]Center for Data Science Analysis, Houghton College, NY 14744, USA
penganjie200012@163.com

Due to the page limitation of our submission to ICASSP2022, we provide more details about experiments in this document.

Experiment Settings

Our experiments are conducted on a subset of the ImageNet validation dataset. This subset is composed of 6000 images, which belong to 1000 classes. We split these 6000 images into two halves. The first 3000 images are used for training an ensemble classifier, and to obtain the relationship of σ and P_{fa}^{1} in $\delta^{1}()$, and the relationship of t and P_{fa}^{2} in $\delta^{2}()$.Taking the BIM attack for example, we add Gaussian noise $N(0, \sigma^{2})$ to 3000 benign images and then denoise them with the same parameter σ. P_{fa}^{1} is calculated as the probability that the classification result has been changed after noise addition-then-denoising. The relationship between σ and P_{fa}^{1}can be obtained by varying σ. Similarly, we use 2000 (benign, adversarial) image pairs to train an ensemble classifier and use the trained classifier to classify the remaining 1000 benign images. Here the adversarial images are generated with mixed attack strengths, and the number of sub-classifiers is set as *N*=51. By doing so, the relationship between t and P_{fa}^{2} can be obtained.[1]

The remaining 3000 images are used for testing. A successful attack is declared when $F(\boldsymbol{I}') = y_t$. To make the attack more practical, $\boldsymbol{I}'$ is saved in PNG format before checking its attacking efficiency. The two-step test is then performed on the successfully attacked images. For the constrained perturbation attacks, $\epsilon \in \{1, 2, 4, 6, 8\}$. For the optimized perturbation attacks, $k \in \{0, 5, 10, 15, 20\}$. Preliminary experiments suggest that Nash equilibrium will not exist in the region of $\epsilon > 8$ or $k > 20$. The strategy of

[1] Some of previous works [1, 2] excluded the images that cannot be correctly predicted by the CNN classifier, i.e.,$F(\boldsymbol{I}) \neq y_{true}$, in calculating false alarm rate. In our experiments, we did not exclude such images based on the following considerations: 1. the spatial instability of the images that can be correctly classified and that cannot be correctly classified is statistically different. The former is usually stronger than the latter. 2. We focus on the targeted attack, which means all images can be used as a cover in attacking. Note in untargeted attacks, only images that are correctly classified can be used as a cover.

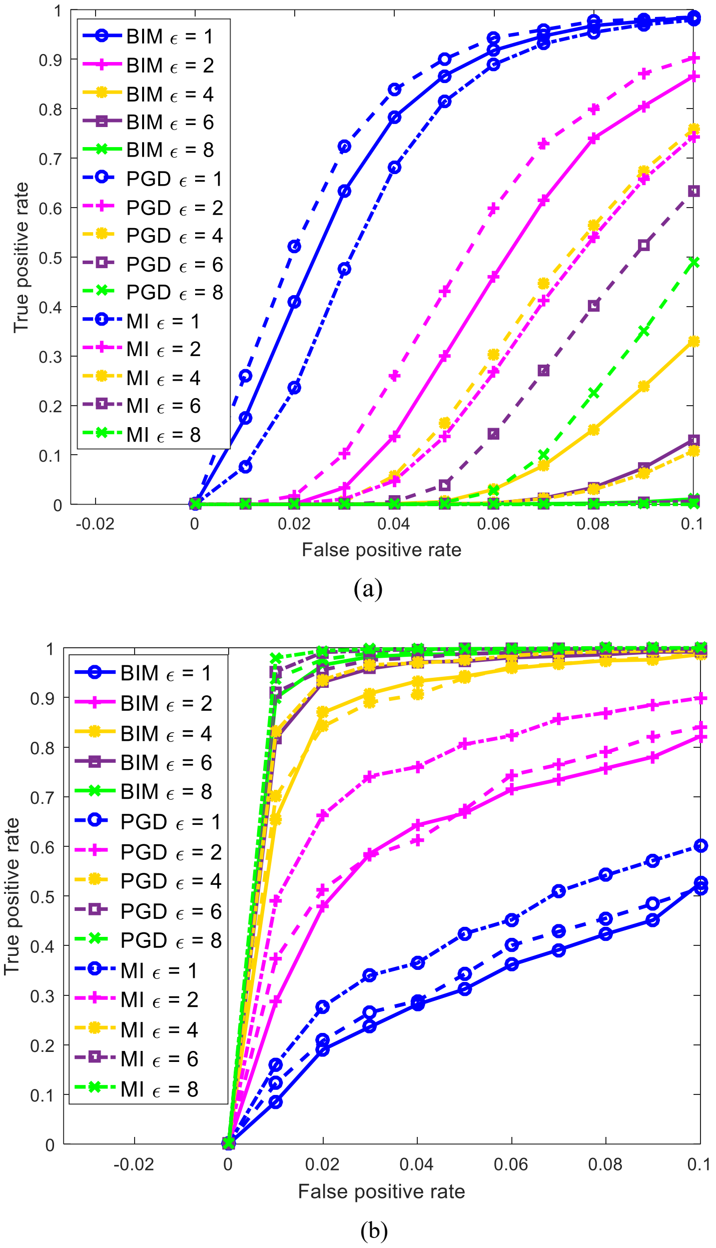
the investigator is $P_{fa}^1 \in \{0{:}\,0.01{:}\,P_{fa}\}$. Since the detection performance in the low P_{fa} area is more critical in practice, the upper bound of P_{fa} is set as 0.1.

Detection performance of each single test

To supplement Figure 4 of the paper, we show the ROC performance of the two single tests on three perturbation-constrained attacks in Fig. 1. The solid lines, dotted lines, and dot-dash lines denote ROC curves of BIM attack, PGD attack, and MI attack, respectively.

Adversarial example security

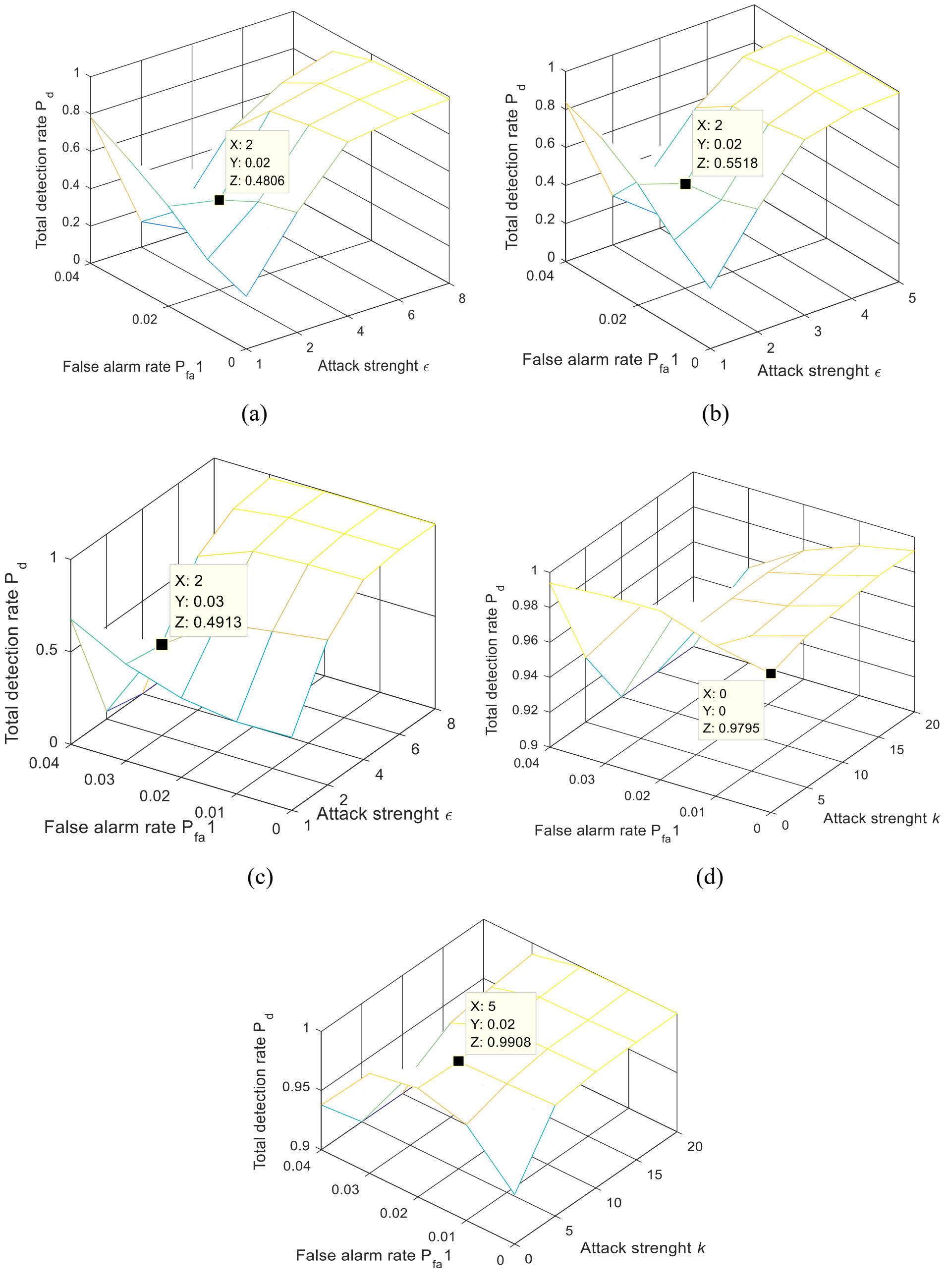
To supplement Figure 6 of the paper, we show the P_d matrices when $P_{fa} = 0.04$ for various attacks in Fig. 2.



(a)

(b)

Fig.1. ROC performance of the two single tests on perturbation-constrained attacks. (a) Noise addition-then-denoising test, (b) SRM based test.



(a)

(b)

(c)

(d)

(e)

Fig.2. P_d matrix and the corresponding Nash equilibrium when $P_{fa} = 0.04$. (a) BIM attack, (b) PGD attack, (c) MI attack, (d) C&W attack, (e) ST attack.

References

1. B. Liang, H. Li, M. Su, et al., “Detecting adversarial image examples in deep neural networks with adaptive noise reduction,” IEEE Transactions on Dependable and Secure Computing, doi: 10.1109/TDSC.2018.2874243, 2018
2. K. Deng, A. Peng, H. Zeng, “Detecting C&W adversarial images based on noise addition-then-denoising,” To appear in International conference on image processing, 2021