HOW SECURE ARE THE ADVERSARIAL EXAMPLES THEMSELVES?

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

[1]School of Computer Sci. and Tech., Southwest University of Sci. and Tech., 621010, China
[2]Center for Data Science Analysis, Houghton College, NY 14744, USA

ABSTRACT

Existing adversarial example generation algorithms mainly consider the success rate of spoofing target model, but pay little attention to its own security. In this paper, we propose the concept of adversarial example security as *how unlikely themselves can be detected*. A two-step test is proposed to deal with the adversarial attacks of different strengths. Game theory is introduced to model the interplay between the attacker and the investigator. By solving Nash equilibrium, the optimal strategies of both parties are obtained, and the security of the attacks is evaluated. Five typical attacks are compared on the ImageNet. The results show that a rational attacker tends to use a relatively weak strength. By comparing the ROC curves under Nash equilibrium, it is observed that the constrained perturbation attacks are more secure than the optimized perturbation attacks in face of the two-step test. The proposed framework can be used to evaluate the security of various potential attacks and further the research of adversarial example generation/detection.

Index Terms— adversarial examples, adversarial example detection, two-step test, game theory, Nash equilibrium

1. INTRODUCTION

The last decade witnessed the rapid progress of convolutional neural networks (CNN) on vision related fields [1, 2]. Recently, some farsighted researchers began to study the security of CNNs, e.g., adversarial example (AE) generation and detection. Szegedy proposed the concept of AE as shown in Fig. 1, which is generated by adding visually imperceptible perturbation to a benign example yet can force a CNN model producing erroneous output [3]. According to the perturbation limitation in generating AEs, existing algorithms can be divided into two categories: constrained perturbation and optimized perturbation [4]. The objective of the former is to maximize the adversarial loss under given perturbation restrictions, usually measured with L_p distance. The representative algorithms of this class are fast gradient sign method (FGSM) [5], randomized FGSM [6], projected gradient descent (PGD) [7], Basic iterative method (BIM) [8], Momentum iterative FGSM (MI) [9], to name a few. On the other hand, the optimized perturbation attack is to minimize the perturbation under some adversarial criteria. L-BFGS [3], DeepFool [10], Carlini and Wagner (C&W) attack [11],

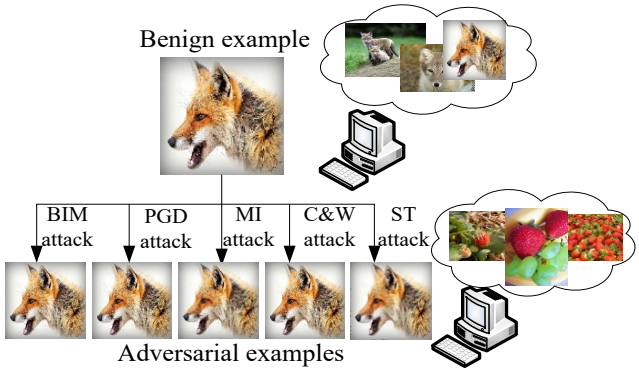

Fig.1. AE. (top) the original image, (bottom) AEs. All these AEs are wrongly classified as 'strawberry' by their target CNN models.

Decoupling direction and norm (DDN) attack [12], and Spatially transformed (ST) attack [13] can be categorized as such class.

In the meantime, there are also studies on how to defend the CNN models from being fooled by AEs. These defenses mainly fall into two categories. The first one is enhancing the robustness of the CNNs, which is usually achieved by modifying the network architecture or the training process [5, 14, and 15]. The other way is detecting the probe example before input into the CNN model. Hendrycks [16] and Li [17] found the discrepancy between benign examples and AEs in their principal components. Lu analyzed the difference between AEs and benign examples in response of the last ReLU layer [18]. In [19], Liu reformulated AE detection as a steganalysis problem and thus proposed a detection method based on spatial rich model [20]. Peng utilized the dependencies among three color channels in detecting AEs [21]. Liang [22], Deng [23], and Xu [24] took advantage of the spatial instability nature of AEs to detect them. In this study, we focus on the second category defenses.

Emerging detection methods suggest that an AE that can fool a target CNN model is insufficient. The ability to circumvent standard detection algorithms is also crucial. However, existing attacks pay little attention to whether themselves are easily detectable, i.e., ignore their own security. Here we define the security of AEs as *how indistinguishable they are from benign examples under a specific detection scheme*. On the other hand, existing detection algorithms usually focus on specific attacks with

The work was supported by NSFC (No. 61702429), Doctoral Research Fund of SWUST (No. 18zx7163), Sichuan Science and Technology Program (No. 22ZDYF3644).

specific settings, whereas attackers may adjust their strategy to evade detection in practice. In this paper, we model the interplay between AE generation and detection as a game [25, 26]. By solving Nash equilibrium (NE) of the game, we are better informed of what strategy will be adopted by both sides, given their awareness of the existence of each other. Our contributions are summarized as follows:

1) A two-step test is proposed to deal with attacks of different strengths. While the components of this test are not novel techniques that we propose here, the framework of combining them to complementarily detect a farsighted attacker who can adjust his strength is new. The necessity of such a combination is justified both in theory and experiments.

2) A zero-sum game is used to model the interplay between AE generation and detection. NEs are solved for five typical attacks to evaluate their security under the two-step test.

3) Unlike existing works that either put themselves in the position of an attacker or a defender, we take the perspective of a third party. Such perspective enables us to make a fair and quantitative evaluation of the security of the state-of-the-art attacks, which is unique in the literature.

2. BACKGROUND

This section briefly reviews the AE generation and detection methods involved in this study.

2.1. Adversarial example generation

Adversarial attacks have two modes: targeted and untargeted. A targeted attack forces a CNN model to classify the generated image as a given label, i.e. $F(\boldsymbol{I}') = y_t$, where $\boldsymbol{I}'$is the adversarial image, $F()$ is the CNN model and y_t is the target label (the 'strawberry' in Fig. 1). The untargeted attack only misleads the CNN model making a wrong classification. Since the former poses an even greater threat to the CNN models, we focus on the targeted attack in this study.

We begin with the BIM attack [8], which is developed from the FGSM attack. It applies FGSM iteratively as follows:

$$\boldsymbol{I}'_{N+1} = Clip_{\boldsymbol{I},\epsilon}\{\boldsymbol{I}'_N - \alpha sign(\nabla_{\boldsymbol{I}'_N} J(\boldsymbol{I}'_N, y_t))\} \quad (1)$$

where $\nabla_{\boldsymbol{I}'_N} J()$ denotes the gradient of the loss function $J()$ with respect to $\boldsymbol{I}'_N$. The accumulated perturbation for each pixel is restricted to $[-\epsilon, \epsilon]$ by $Clip_{\boldsymbol{I},\epsilon}\{\}$. Unlike BIM, which begins with the original image, PGD begins with a noisy version of the original image [7]. To enhance the transferability of the generated AEs, Dong integrated a momentum term into the iterative process in the MI attack [9]:

$$g_{N+1} = \mu \cdot g_N + \nabla_{\boldsymbol{I}'_N} J(\boldsymbol{I}'_N, y_t) \quad (2)$$

where μ is the decay factor, and the MI attack reduces to the BIM attack when $\mu = 0$.

C&W attack generates AEs by solving the following optimization problem [11]:

$$\begin{aligned} &\text{minimize } \left|\left|\boldsymbol{\delta}\right|\right| + c \cdot f(\boldsymbol{I} + \boldsymbol{\delta}) \\ &s.t.\ \boldsymbol{I}' = \boldsymbol{I} + \boldsymbol{\delta} \in [0,1]^n \end{aligned} \quad (3)$$

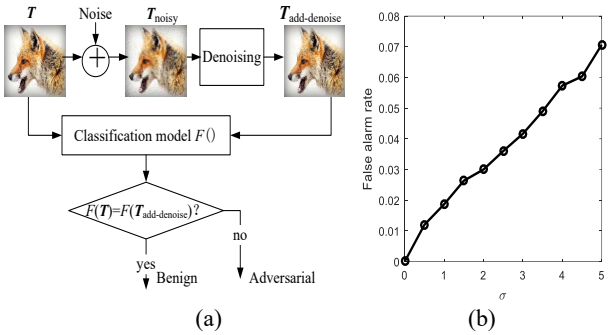

Fig.2. The noise addition-then-denoising detection [23]. (a) the diagram, (b) the false alarm rate as a function of σ on the ImageNet.

where c is used to balance fidelity loss and adversarial loss. $f()$ is defined as

$$f(\boldsymbol{x}) = \max(-k, \max_{i \neq y_t}\{Z(\boldsymbol{x})_i\} - Z(\boldsymbol{x})_{y_t}) \quad (4)$$

where $Z()$ is the output vector before the softmax layer of the CNN model, and k is used for controlling the attack confidence. According to the distance metric used in (3), there are three versions of the C&W attack, C&W L_0, C&W L_2 and C&W L_∞, among which C&W L_2 is reported to have the best attack performance and the fastest speed, and thus is chosen as our research target. Unlike all the above methods manipulating pixel values, ST attack introduces perturbation to the pixel position to achieve better spatial stability [13].

2.2. Detection algorithm

Existing detection methods are based on the following two assumptions about AEs. 1) Spatial instability. 2) Adversarial perturbation destroys local correlation. Starting from the first assumption, we proposed a noise addition-then-denoising method to detect AE [23], as shown in Fig. 2(a). For a probe image $\boldsymbol{T}$, a noisy version $\boldsymbol{T}_{noisy}$ is generated by adding Gaussian noise $N(0, \sigma^2)$. Then, $\boldsymbol{T}_{noisy}$ is denoised with a denoising filter, e.g., the FFDNet [27]. $\boldsymbol{T}$ is identified as adversarial if $F(\boldsymbol{T}) \neq F(\boldsymbol{T}_{add-denoise})$. The false alarm rate ($P_{fa}$) of this method increases with σ. This is because even a benign image cannot keep its classification label when σ is large. Fig. 2(b) shows P_{fa} of this method as a function of σ on the ImageNet [28]. A typical detection method based on the second assumption is the SRM based method [19]. Specifically, a 34671-D SRM feature set is extracted from an image and fed into an ensemble classifier [29] to identify AEs.

3. GAME THEORETIC EVALUATION

We use attack strength r to represent the parameter ϵ or k of the attacks in a unified way. A stronger r makes $\boldsymbol{I}'$ being classified as y_t with higher confidence. However, a larger r also means $\boldsymbol{I}'$ has changed more from $\boldsymbol{I}$, which increases the risk of $\boldsymbol{I}'$ being detected by SRM based test. Hence, a rational attacker will make a tradeoff in choosing r between escaping these two tests. Similarly, the investigator has to make a tradeoff in allocating the attention of the two tests.

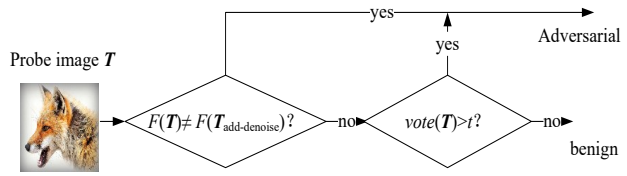

Fig.3. The diagram of the two-step test

3.1. Two-step test

We define a two-step test employed by the investigator as Fig. 3, where a probe image $\boldsymbol{T}$ is first detected with the noise addition-then-denoising test ($\delta^1()$), whose false alarm rate P_{fa}^1 can be controlled by the standard deviation σ.

$\delta^1(\boldsymbol{T}) = H_0^1$: $\boldsymbol{T}$ is benign. The subindex "0" denotes the null hypothesis, and the superindex "1" denotes the first step.

$\delta^1(\boldsymbol{T}) = H_1^1$: $\boldsymbol{T}$ is adversarial, whose acceptance region is:

$$\boldsymbol{T}: F(\boldsymbol{T}) \neq F(\boldsymbol{T}_{add-denoise}) \quad (5)$$

For the images accepted in H_0^1, the investigator performs the SRM based test ($\delta^2()$). Here, we classify whether $\boldsymbol{T}$ is benign or adversarial by comparing the voting result of the ensemble classifier with a predefined threshold t, which is determined by an allowed false alarm rate P_{fa}^2:

$\delta^2(\boldsymbol{T}) = H_0^2$: $\boldsymbol{T}$ is benign.

$\delta^2(\boldsymbol{T}) = H_1^2$: $\boldsymbol{T}$ is adversarial. The acceptance region of H_1^2 is $vote(\boldsymbol{T}) > t$.

The total detection rate of the two-step test is

$$P_d = P(\delta^1(\boldsymbol{T}) = H_1^1 \cup \delta^2(\boldsymbol{T}) = H_1^2 | \boldsymbol{T} \; is \; adversarial) \quad (6)$$

The total false alarm rate is the probability that a benign example is misclassified as adversarial.

$$\begin{aligned} P_{fa} &= P(\delta^1(\boldsymbol{T}) = H_1^1 \cup \delta^2(\boldsymbol{T}) = H_1^2 | \boldsymbol{T} \text{ is benign }) \\ &= P(\delta^1(\boldsymbol{T}) = H_1^1 | \boldsymbol{T} \text{ is benign }) + P(\delta^2(\boldsymbol{T}) = H_1^2 | \boldsymbol{T} \text{ is benign}) \\ &\quad - P(\delta^1(\boldsymbol{T}) = H_1^1 \cap \delta^2(\boldsymbol{T}) = H_1^2 | \boldsymbol{T} \text{ is benign }) \\ &\approx P_{fa}^1 + P_{fa}^2 \end{aligned} \quad (7)$$

The last approximation holds because the chance of a benign image introducing false alarms in both tests is small.

3.2. Adversarial example detection game

Since the objectives of attacker and investigator are strictly competitive, we model this interplay as a game:

Adversarial-detection$(S_I, S_A, \boldsymbol{U})$ *game* is a zero sum, complete information game played by the investigator and the attacker, featured by the following strategies and payoff:

1) S_I: The investigator's strategy space, i.e., P_{fa}^1 that can be allocated to $\delta^1()$.

2) S_A: The attacker's strategy space, i.e., the attacking strength r in generating AEs.

3) $\boldsymbol{U}$: The payoff matrix, which is defined as the total detection rate of the two-step test

$$U(P_{fa}^1, r) = P_d(P_{fa}^1, r) \quad (8)$$

To make the following study tractable, we limit the strategy spaces of the investigator and the attacker to P_{fa}^1 and r, respectively. Both players may have more flexibility in practice. For example, the investigator can choose different detection methods as alternatives to $\delta^1()$ or $\delta^2()$. She may also adopt different strategies in combining the detection results. To ensure the existence of pure-strategy NE, We assume that the investigator chooses her strategy first and allows the attacker to respond. Under such assumption, the NE of the Adversarial-detection $(S_I, S_A, \boldsymbol{U})$ game can be obtained from

$$(P_{fa}^{1*}, r^*) = arg \max_{P_{fa}^1} \min_r U(P_{fa}^1, r) \quad (9)$$

For every given P_{fa}, a P_d matrix is obtained by varying P_{fa}^1 and r, and the corresponding NE can be solved using (9). By examining the relationship between $P_d(P_{fa}^{1*}, r^*)$ and P_{fa}, a receiver operating characteristic (ROC) curve under NE can be obtained and is called NEROC [30].

4. EXPERIMENTAL RESULTS

Experiments are performed to evaluate the security of five attacks: BIM, PGD, MI, C&W, and ST, implemented with the advertorch toolbox [31]. The target model is a pre-trained ResNet18 model [1].

4.1. Experiment Settings

Six thousand images from the ImageNet validation dataset are used in our experiments. We split them into two halves. The first 3000 images are used for training an ensemble classifier and for obtaining the relationship of σ and P_{fa}^1 in $\delta^1()$, and the relationship of t and P_{fa}^2 in $\delta^2()$. A successful attack is declared when $F(\boldsymbol{I}') = y_t$. $\boldsymbol{I}'$ is saved as PNG format before checking its attacking efficiency. The two-step test is only performed on the successfully attacked images. For the constrained perturbation attacks, $\epsilon \in \{1, 2, 4, 6, 8\}$. For the optimized perturbation attacks, $k \in \{0, 5, 10, 15, 20\}$. The strategy of 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. More detailed settings and results are provided in the supplementary material: *github.com/zengh5/adversarial-example-security*.

4.2. Detection performance of each single test

We first evaluate the performance of two single tests on the constrained perturbation attacks. Fig. 4(a) shows the ROC curves of $\delta^1()$ on the BIM attack and MI attack with different ϵs. Both attacks are more and more difficult to detect with the increase of ϵ as expected. Fig. 4(b) shows the ROC curves of $\delta^2()$. Contrary to Fig. 4(a), both attacks are getting easier to detect in this test with the increase of ϵ. Fig. 4(a) and (b) indicate a strong complementarity between $\delta^1()$ and $\delta^2()$, which forces the attacker to make a tradeoff in choosing ϵ.

The detection performance on optimized perturbation attacks is shown in Fig. 5. Taking a closer look at Fig. 5(a), we observe that the spatial stability of the ST attack is much better than that of the C&W attack (compare the two pink

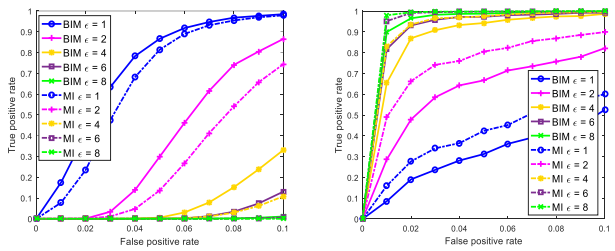

Fig.4. ROC performance of the two single tests on the BIM and MI attacks. (a) Noise addition-then-denoising test, (b) SRM based test.

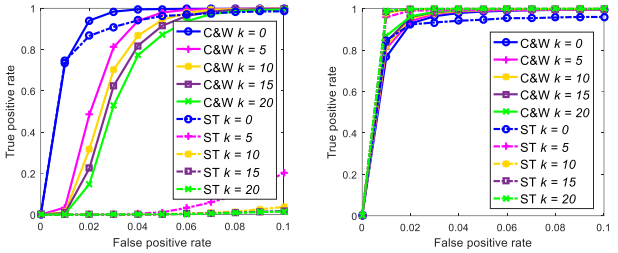

Fig.5. ROC performance of the two single tests on the C&W and ST attacks. (a) Noise addition-then-denoising test, (b) SRM based test.

lines for instance), which is in line with the observation of [13]. Important findings can be observed by comparing Fig. 5(b) with Fig. 4(b). In Fig. 4(b), the performance of $\delta^2()$ is strongly related with ϵ. However, it is not very sensitive to k in Fig. 5(b), which means an attacker who adopts optimized perturbation attacks is difficult to evade detection by adjusting attack strength. Such difference is more evident by examining the results under NE in the next.

4.3. Why is the two-step test needed?

From the results of the last section, we note that neither of the two tests is individually competent for detecting a farsighted attacker who is flexible with his attacking strength. For example, the BIM attacker can always choose a sufficiently large ϵ to escape detection if only $\delta^1()$ is utilized by the investigator, and choose a strength as weak as possible if only $\delta^2()$ is adopted by the investigator. On the contrary, as can be seen in the following subsection, if the investigator

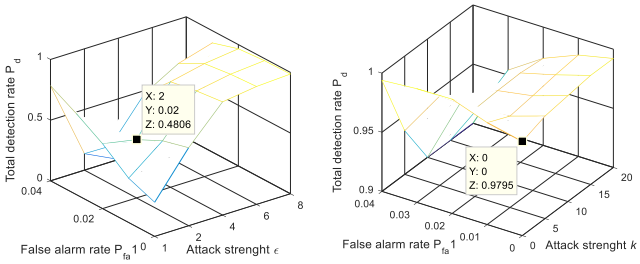

Fig.6. P_d matrix at $P_{fa} = 0.04$. (a) BIM attack, (b) C&W attack.

combine $\delta^1()$ and $\delta^2()$ together, a much higher detection rate can be achieved *no matter how the attacker adjusts his attacking strength*.

Another interesting question is *which attack is more secure*, e.g., BIM *vs* MI? The answer is not straightforward because MI attack shows better security in $\delta^1()$, whereas BIM attack can better circumvent $\delta^2()$. Game theory analysis based on the two-step test can answer this question, as will be shown in the next.

4.4. Adversarial example security

We then evaluate the security of different attacks under the two-step test. Fig. 6(a) shows the total detection rate P_d at $P_{fa} = 0.04$ when the attacker adopt the BIM attack. According to (9), the NE point is $(P_{fa}^{1*} = 0.02, \epsilon^* = 2)$. In this NE, the total detection rate $P_d = 48.06\%$. Fig. 6(b) shows the P_d matrix for the C&W attack at $P_{fa} = 0.04$, where the NE point is $(0, 0)$. This means the investigator only performs $\delta^2()$, and the attacker responds with the weakest attack strength. The NEROC curves for different attacks are presented in Fig. 7. It is observed that the optimized perturbation attacks have a higher probability of being detected than the constrained perturbation attacks. This is because the attacker can efficiently escape detection by adjusting attacking strength in constrained perturbation attacks. However, the security gain by adjusting strength is limited for the optimized perturbation attacks.

5. CONCLUSION

In this paper, the security of AEs is defined as how unlikely themselves can be detected. We model the interplay between the investigator and the attacker with game theory. By examining the detection performance of the two-step test under NE, we have the following findings: 1) for all the studied AE algorithms, a rational attacker always tends to adopt a weak strength. 2) The attacks that show better spatial stability, e.g., ST vs C&W, may destroy the inherent local correlation within natural images more seriously and result in lower overall security. 3) The constrained perturbation attacks are more secure than the optimized perturbation attacks in face of the two-step test.

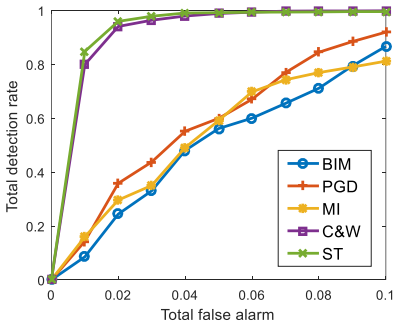

Fig.7. Nash equilibrium ROC of the adversarial-detection game when the attacker utilizing different attack algorithms.

12. REFERENCES

[1] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[2] K. Bourzac, "Bringing big neural networks to self-driving cars, smart phones, and drones," IEEE Spectrum, pp. 13–29, 2016.

[3] C. Szegedy, W. Zaremba, I. Sutskever, et al., "Intriguing properties of neural networks," Proceedings of International Conference on Learning Representations 2014, arxiv: 1312.6199.

[4] X. Yuan, P. He, Q. Zhu, et al., "Adversarial Examples: Attacks and Defenses for Deep Learning," IEEE Transactions on Neural Networks and Learning Systems, 30(9): 2805–2824, 2019.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv: 1412.6572, 2014.

[6] F. Tramèr, A. Kurakin, N. Papernot, et al., "Ensemble adversarial training: Attacks and defenses," proceedings of International Conference on Learning Representations, 2018. arXiv: 1705.07204.

[7] A. Madry, A. Makelov, L. Schmidt, et al., "Towards deep learning models resistant to adversarial attacks," proceedings of International Conference on Learning Representations, 2017. arXiv: 1706.06083.

[8] A. Kurakin, I. Goodfellow, S. Bengio, "Adversarial examples in the physical world," Proceedings of International Conference on Learning Representations, 2016. arXiv:1607.02533.

[9] Y. Dong, F. Liao, T. Pang, et al., "Boosting adversarial attacks with momentum," Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.

[10] S. Moosavidezfooli, A. Fawziand, P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," In Proceedings of Computer Vision and Pattern Recognition, pp. 2574–2582, 2015.

[11] N. Carlini, D. Wagner, "Towards evaluating the robustness of neural networks" IEEE Symposium on Security and Privacy, pp. 39–57, 2017.

[12] J. Rony, L. G. Hafemann, L. S. Oliveira, et al., "Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4317–4325.

[13] C. Xiao, J. Y. Zhu, B. Li, et al., "Spatially transformed adversarial examples," International conference on learning representations, 2018.

[14] N. Papernot, "Distillation as a defense to adversarial perturbations against deep neural networks," IEEE Symposium on Security and Privacy, pp. 582–597, 2016.

[15] N. Papernot, P. McDaniel, I. J. Goodfellow, et al., "Practical black-box attacks against machine learning" In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519, 2017.

[16] D. Hendrycks, K. Gimpel, "Early methods for detecting adversarial images" arXiv:1608. 00530, 2016.

[17] X. Li, F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," Proceedings of IEEE International Conference on Computer Vision, pp. 5775–5783, 2017.

[18] J. Lu, T. Issaranon, D. Forsyth, "Safetynet: Detecting and rejecting adversarial examples robustly," Proceedings of IEEE International Conference on Computer Vision, pp. 446–454, 2017.

[19] J. Liu, W. Zhang, Y. Zhang, et al., "Detection Based Defense Against Adversarial Examples From the Steganalysis Point of View," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4820–4829

[20] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," IEEE Trans. Info. Forensics and Security, 7(3): 868–882, 2012.

[21] A. Peng, K. Deng, J. Zhang, et al., "Gradient-based adversarial image forensics," the 27th International Conference on Neural Information Processing, pp. 417–428, 2020.

[22] 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, 18(1): 72–85, 2018.

[23] K. Deng, A. Peng, H. Zeng, "Detecting C&W adversarial images based on noise addition-then-denoising," Int. conf. image processing 2021, pp. 3607–3611

[24] W. Xu, Y. David, J. Yan, "Feature squeezing: detecting adversarial examples in deep neural networks," Network and Distributed System Security Symposium, arXiv: 1704.01155, 2017.

[25] D. Fudenberg, J. Tirole, "Game theory," Cambridge, MA, USA: MIT Press, 1991.

[26] M. Barni and B. Tondi, "The Source Identification Game: An Information-Theoretic Perspective," IEEE Trans. Info. Forensics and Security, 8(3): 450–463, 2013

[27] K. Zhang, W. Zuo, L. Zhang, "FFDNet: Toward a Fast and Flexible Solution for CNN based Image Denoising," IEEE Trans. Image Processing, 27(9): 4608–4622, 2018.

[28] O. Russakovsky, J. Deng, H. Su, et al., "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, 115(3): 211–252, 2015.

[29] J. Kodovsky, J. Fridrich, V. Holub, "Ensemble classifiers for steganalysis of digital media," IEEE Trans. Info. Forensics and Security, 7(2): 432–444, 2012.

[30] M. C. Stamm, W. S. Lin, K. J. R. Liu, "Temporal forensics and anti-forensics for motion compensated video," IEEE Trans. Info. Forensics and Security, 7(4): 1315–1329, 2012.

[31] G. W. Ding, L. Wang, X. Jin, "advertorch v0.1: An Adversarial Robustness Toolbox based on PyTorch," 2019. https://arxiv.org/abs/1902.07623