ADVERSARIAL EXAMPLE DETECTION BAYESIAN GAME

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

[1]Southwest University of Science and Technology, [2]Beijing Normal University

ABSTRACT

Despite the increasing attack ability and transferability of adversarial examples (AE), their security, i.e., how unlikely they can be detected, has been ignored more or less. Without the ability to circumvent popular detectors, the chance that an AE successfully fools a deep neural network is slim. This paper gives a game theory analysis of the interplay between an AE attacker and an AE detection investigator. Taking the perspective of a third party, we introduce a game theory model to evaluate the ultimate performance when both the attacker and the investigator are aware of each other. Further, a Bayesian game is adopted to address the information asymmetry in practice. Solving the mixed-strategy Nash equilibrium of the game, both parties' optimal strategies are obtained, and the security of AEs can be evaluated. We evaluate four popular attacks under a two-step test on ImageNet. The results may throw light on how a farsighted attacker or investigator will act in this adversarial environment. Our code is available at: *https://github.com/zengh5/AED_BGame*.

Index Terms— adversarial examples, adversarial example detection, game theory, Bayesian game, mixed-strategy Nash equilibrium

1. INTRODUCTION

The discovery of adversarial examples (AE) [1] has raised security concerns in deploying convolutional neural network (CNN)-based applications [2, 3]. Existing works on AE are either from an attack or defense perspective [4]. Most advanced attacks are devoted to enhancing the transferability [5] or robustness (to standard image processing) [6] of AEs. State-of-the-art defenses mainly fall into two categories. The first one is enhancing the robustness of the CNNs by modifying the network architecture or the training process [7-9]. The other way is detecting and rejecting potential AEs before inputting them into the CNN model [10].

While the attacker may circumvent the first type of defense by increasing the attack ability or mounting a secondary attack, how to avoid being detected is widely ignored. [11] points out that a successful AE should be able to circumvent standard detectors and defines the security of AEs as how indistinguishable they are from benign examples. Under the assumption that both the AE attacker and the AE detection investigator have complete information about each other, the security of several attacks is evaluated with game theory [12, 13]. However, the complete information assumption may not hold in practice, e.g., the investigator does not know the attack method used by the attacker. In our study, we resort to the Bayesian game to address the information asymmetry between the attacker and the investigator. The detailed solution to the Bayesian game is given, and the Nash equilibrium ROCs for four widely used attacks are obtained on the ImageNet. We are informed how the information asymmetry will affect the AE's security by comparing the Bayesian game with its complete information counterpart. Our contributions are summarized as follows:

1) For the first time, the information asymmetry between the AE attacker and investigator is modeled with a Bayesian game, which removes the complete information assumption in [11].

2) Solving the games with mixed-strategy Nash equilibrium removes the sequential-move assumption in [11].

2. BACKGROUND

2.1. Adversarial example generation

Since the targeted attack is known to be able to raise more security concerns than its untargeted counterpart, we focus on the targeted attack in this study, in which the attacker forces a CNN model to classify the generated image as a given label y_t, i.e., $F(\boldsymbol{I}') = y_t$, where $\boldsymbol{I}'$ is the adversarial image, and $F()$ is the classification model.

IFGSM [14] and its variants. IFGSM perturbs a benign image $\boldsymbol{I}$ iteratively with step size α:

$$\boldsymbol{I}'_0 = \boldsymbol{I}$$
$$\boldsymbol{I}'_{N+1} = Clip_{\boldsymbol{I},\epsilon}\{\boldsymbol{I}'_N - \alpha sign(\nabla_{\boldsymbol{I}'_N} J(\boldsymbol{I}'_N, y_t))\} \quad (1)$$

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 $Clip_{\boldsymbol{I},\epsilon}\{\}$. To enhance the transferability of the generated AEs, Dong et al. integrate a momentum term into the iterative process in the MI attack [5]:

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

where μ is the decay factor, and the MI attack reduces to IFGSM when $\mu = 0$. For the same purpose, other enhanced variants of IFGSM compute the gradients with respect to the randomly-transformed or translated input [15, 16].

C&W attack [17]. C&W attack generates AEs by solving the following optimization problem:

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

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

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

This work was supported by the network emergency management research special topic (no. WLYJGL2023ZD003), the Opening Project of Guangdong Province Key Laboratory of Information Security Technology (no. 2020B1212060078), Sichuan Science and Technology Program (no. 2022YFG0321).

where $Z()$ is the output logits of the CNN model, and k is used for controlling the attack confidence.

ST attack [6]. Unlike all the above methods manipulating pixel values, ST attack introduces perturbation to the pixel position by minimizing the following objective function:

$$f(\boldsymbol{x}) + \tau \cdot f_{flow}(\boldsymbol{x}) \quad (5)$$

where $f()$ is the same as the adversarial loss in (4), $f_{flow}()$ is the spatial transformation perturbation term, and τ is used to balance these two losses. ST attack is known for its better spatial stability.

2.2. Adversarial Example Detector

Existing detectors can be divided into two categories according to the assumption about AEs they are based on. The first-type detectors utilize the spatial instability of AEs, i.e., a slight disturbance to $\boldsymbol{I}'$ may change its classification label. 2) The second-type detectors believe that adversarial perturbation destroys certain characteristics of natural images, e.g., local correlation.

Based on the first assumption, an Adaptive Noise Reduction-based detector (ANR) is proposed in [18]. A probe image first undergoes different noise reduction processings. It is then identified as adversarial if its label is inconsistent before/after denoising. Similarly, in [19], various squeezing methods, e.g., bit-depth reduction and non-local filter, are applied for a probe image. Then, the L_1 distances between a classifier's prediction on the probe image and its squeezed versions are calculated. Finally, the maximum value of these distances is taken as the metric to determine whether the probe image is benign or adversarial. Recently, a novel image processing method called noise addition-then-denoising (AddDe) is proposed to reveal potential AEs [20]. For a probe image $\boldsymbol{T}$, an AddDe-processed version $\boldsymbol{T}_{add-denoise}$ is generated by first adding Gaussian noise $N(0, \sigma^2)$ then denoising. $\boldsymbol{T}$ is identified as adversarial if $F(\boldsymbol{T}) \neq F(\boldsymbol{T}_{add-denoise})$. The false alarm rate of this method increases with σ. This is because even a benign image cannot keep its classification label when σ is large. [10] reveals potential AEs by examining the difference in adversarial gradient directions (AGDs) before and after a random perturbation.

Under the second assumption, [21] declares that AEs are out-of-distribution samples in the representation space of CNNs. [22] reformulates the AE detection as a steganalysis problem and proposes a spatial rich model (SRM) [23]-based detector. Specifically, a 34671-D feature set is extracted from an image and fed into an ensemble classifier [24] to identify AEs. Most second-type detectors are learning-based. Thus, they must train a dedicated model for each attack, even each strength. On the other hand, the first-type detectors tend to be attack-agnostic (except [10]).

3. ADVERSARIAL EXAMPLE DETECTION GAME

3.1. Interplay between attacker and investigator

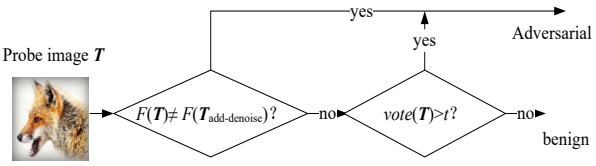

Fig. 1. The diagram of the two-step test.

Both types of detectors reviewed in the last section have their own merits and limitations. Detectors of the first type perform well against attacks of weak strength but are incompetent to stronger attacks. By contrast, the second-type detectors are good at detecting high-intensity attacks but poor at detecting weak attacks. Here we use attack strength r to represent the parameter ϵ in (1) or k in (4) in a unified way.

A farsighted attacker thus can avoid being detected by adjusting r once he knows the type of detector. In response to such farsighted attackers, we assume the investigator adopts the two-step test proposed in [11] for detecting AEs. The diagram of the two-step test is displayed in Fig. 1. A probe image $\boldsymbol{T}$ is first examined with $\delta^1()$, which is a first-type detector. If $\boldsymbol{T}$ passes $\delta^1()$, the investigator runs a second-type detector $\delta^2()$ on it. Only if it is considered benign in both steps will it be declared benign by the two-step test. In this study, we assume the investigator adopts the AddDe-based detector to perform $\delta^1()$, and the SRM-based detector to perform $\delta^2()$. However, it is necessary to point out that other detectors of the corresponding type can replace them.

As analyzed above, steps one and two of the two-step test are characterized by their complementarity. If the attacker wants to circumvent $\delta^1()$ by increasing attack strength, his risk of being detected by $\delta^2()$ will increase. Therefore, a rational attacker will make a tradeoff in choosing r, between circumventing $\delta^1()$ and $\delta^2()$. Similarly, the investigator needs to make a tradeoff in allocating attention between $\delta^1()$ and $\delta^2()$. In [11], such attention allocation is quantified as allocating allowed false alarm rate between $\delta^1()$ and $\delta^2()$. Such interplay between the attacker and the investigator is typically a game.

3.2. Game model

To make the following study tractable, we limit the strategy spaces of the investigator and the attacker to the false alarm rate of $\delta^1()$ (P_{fa}^1) and attacking strength r, respectively. Here we begin with a complete information case. That is, both sides have complete information about the game, e.g., they know the payoff matrix of the other.

Definition 1 *(Complete Information Game): AE-detection game* $(S_I, S_A, \boldsymbol{U})$ 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

$$U(P_{fa}^{1}, r) = P_d(P_{fa}^{1}, r) \quad (6)$$

Next, we relax the complete information assumption of *AE-detection*$_a(S_I, S_A, \boldsymbol{U})$ *game* and allow the attack method to be the attacker's private information. Specifically, a Bayesian game is used to model this potential information asymmetry. This is a more practical scenario since the investigator never knows the exact method adopted by the attacker.

Definition 2 (Bayesian Game): AE-detection$_b$ $(S_I, S_A, \Omega, \boldsymbol{p}, \boldsymbol{U})$ *game* is a zero-sum, incomplete information game defined as the *AE-detection*$_a(S_I, S_A, \boldsymbol{U})$ with the difference that:

1) Ω: The set of attack methods. $\Omega \in \{A_1, A_2, \cdots, A_N\}$, where N is the number of potential attacks, A_l corresponds to the case the attacker adopts the l^{th} attack method.

2) $\boldsymbol{p}$: The prior belief about the probability measure of Ω. $\boldsymbol{p} = [p_1, p_2, \cdots, p_N]$ and $\sum_l p_l = 1$.

Note $\boldsymbol{p}$ is the common knowledge of both sides. If $p_l = 1$ for any $l \in \{1,2,\cdots,N\}$, the Bayesian game reduces to a complete information game.

3.2. Game solution

The most commonly used solution concept in game theory is Nash equilibrium, a profile of strategies such that each player's strategy is an optimal response to other players' strategies. The games defined above are finite strategic games that may not have pure-strategy Nash equilibrium. Hence, we resort to the mixed-strategy Nash equilibrium, which is determined to exist in finite strategic games. The attacker's mixed-strategy $\boldsymbol{r} = [y_1, y_2, \ldots, y_n]$ is a probability distribution over different rs, and the investigator's mixed-strategy $\boldsymbol{P}_{fa}^{1} = [x_1, x_2, \ldots, x_m]$ is a probability distribution over different P_{fa}^{1}s.

To solve the *AE-detection*$_a(S_I, S_A, \boldsymbol{U})$ game, we formulate it as a linear optimization problem [25], i.e., maximizing v, which is subject to

$$\begin{cases} x_i \geq 0 & i = 1,2,\ldots m \\ \sum_i x_i = 1 & \\ \sum_i U_{ij} x_i - v \geq 0 & j = 1,2,\ldots,n \end{cases} \quad (7)$$

where $U_{ij} = P_d(P_{fai}^{1}, r_j)$ is the total detection rate when the investigator adopts P_{fai}^{1} and the attacker adopts r_j. Solving the optimization over $m+1$ parameters $(v, x_1, x_2, \ldots, x_m)$, we can obtain the solution $v*$ to the *AE-detection*$_a(S_I, S_A, \boldsymbol{U})$ game and the Nash equilibrium strategy $\boldsymbol{P}_{fa}^{1*}$ for the investigator. The attacker's Nash equilibrium strategy $\boldsymbol{r}^*$ can be obtained by solving a dual problem of (7).

To solve the Bayesian game, we replace the payoff matrix $\boldsymbol{U}$ in (7) with its expected version $\boldsymbol{UE}$:

$$UE_{ij} = \sum_l p_l U_{ij,l} \quad (8)$$

where $\boldsymbol{U}_l$ is the payoff matrix when the attacker adopts the l^{th} attack method. Then the investigator's strategy $\boldsymbol{P}_{fa}^{1*}$ can be computed as done in the complete information game. Finally, the attacker's strategy, which is type-contingent, can be calculated as:

$$r_l^* = arg\min_r U_l(\boldsymbol{P}_{fa}^{1*}, r) \quad (9)$$

By examining the relationship between $P_d(\boldsymbol{P}_{fa}^{1*}, \boldsymbol{r}^*)$ and P_{fa}, a receiver operating characteristic (ROC) curve called Nash equilibrium ROC (NEROC) [26] can be obtained, which will be used to evaluate the security of AEs next. In the supplementary file: *AED_BGame/supp.pdf*, we also provide the results when one side deviates from the NE strategy.

4. EXPERIMENTAL RESULTS

In this section, we leverage both *AE-detection*$_a(S_I, S_A, \boldsymbol{U})$ game and *AE-detection*$_b(S_I, S_A, \Omega, \boldsymbol{p}, \boldsymbol{U})$ game to evaluate the security of four widely used attacks: IFGSM, MI, C&W, and ST. The target model is a pre-trained ResNet18 model [2].

4.1. Experiment Settings

Our experiments use ten thousand images from the ImageNet validation dataset [27]. Among them, 7000 images are used for training and the remaining for testing. A successful attack is declared when $F(\boldsymbol{I'})$ equals a randomly assigned y_t. The two-step test is only performed on the successfully attacked images. For IFGSM and MI attacks, the attack strength $\epsilon \in \{1, 2, 4, 6, 8\}$. For C&W and ST attacks, $k \in \{0, 5, 10, 15, 20\}$. Preliminary experiments suggest that Nash equilibrium will not exist in the region of $\epsilon > 8$ or $k > 20$. Considering the fact that the investigator is unsure of the attack method mounted by the attacker, the ensemble classifier used for $\delta^2()$ is trained with different attacks and mixed attack strengths. The investigator's strategy $P_{fa}^{1} \in \{0{:}\,0.01{:}\,P_{fa}\}$. 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.

4.2. Complementarity of the two-step test

The subsection elaborates the complementarity between $\delta^1()$ and $\delta^2()$ in AE detection. Fig. 2(a) shows the ROC curves of $\delta^1()$ on the C&W attack and ST attack with different ks. Both attacks are increasingly challenging to detect with the increase of k as expected. Fig. 2(b) shows the ROC curves of $\delta^2()$. Contrary to Fig. 2(a), both attacks are getting easier to detect with the increase of k. The complementarity between $\delta^1()$ and $\delta^2()$ suggests that a

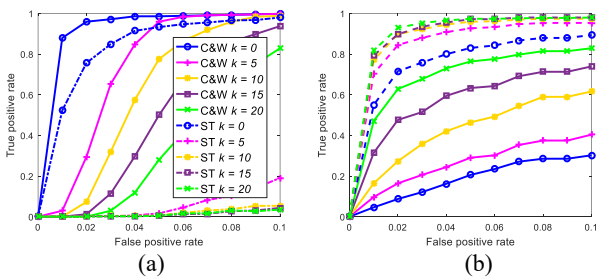

Fig. 2. ROC performance of the two single tests on the C&W and ST attacks. (a) Noise addition-then-denoising test, (b) SRM-based test. Fig. 2(a) and (b) share a legend for better visualization.

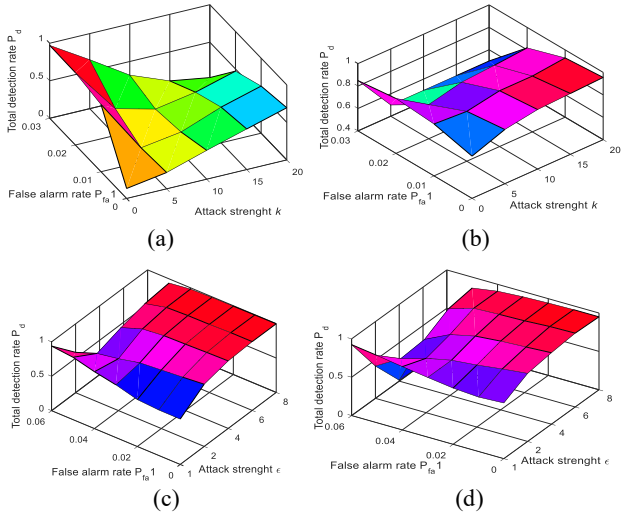

Fig. 3. P_d matrix of the two-step test. (a) C&W attack, $P_{fa} = 0.03$, (b) ST attack, $P_{fa} = 0.03$, (c) IFGSM, $P_{fa} = 0.06$, (d) MI, $P_{fa} = 0.06$.

farsighted attacker should not adopt a too-strong or too-weak strength under the two-step test.

4.3. Complete information game

We then evaluate the security of different attacks under the two-step test. Fig. 3(a) shows the total detection rate P_d at $P_{fa} = 0.03$ when the attacker adopts the C&W attack, no point in which is both the lowest in the k-direction and the highest in the P_{fa}^1 -direction, i.e., pure-strategy Nash equilibrium does not exist. Hence, we find mixed-strategy Nash equilibrium here as the solution to the game. In the mixed-strategy Nash equilibrium, the attacker chooses k = [5, 10] with a probability combination of [0.13, 0.87], and the investigator chooses P_{fa}^1 = [0, 0.03] with a probability combination of [0.67, 0.33]. The total detection rate under Nash equilibrium here is 0.36. Fig. 3(b) shows P_d at $P_{fa} = 0.03$ for the ST attack. The Nash equilibrium is that the attacker chooses k = [0, 5] with a probability combination of [0.31, 0.69], and the investigator chooses P_{fa}^1 = [0, 0.01] with a probability combination of [0.15, 0.85]. The total detection rate under this Nash equilibrium is 0.85. Fig. 3(c) and (d) plot P_d at $P_{fa} = 0.06$ when the attacker adopts the IFGSM and MI, respectively. The total detection rates under Nash equilibrium are 0.80 in Fig. 3(c) and 0.78 in Fig. 3(d). An instructive observation is that the attacker tends to adopt relatively weak strength under Nash equilibrium. In all of our experiments, the supports of $\boldsymbol{r}^*$(the strategies with nonzero probability) are always in $k \leq 10$ or $\epsilon \leq 4$.

The NEROC curves of the *AE-detection*$_d(S_I, S_A, \boldsymbol{U})$ game are presented as dashed lines in Fig. 4. By comparing the total detection rate under Nash equilibrium, we observe that the security of the C&W attack is much stronger than that of the ST attack. Such comparison is not straightforward by examining two tests separately. Note Fig. 2 suggests that the ST attack is more secure under $\delta^1()$, whereas the C&W attack is more secure under $\delta^2()$.

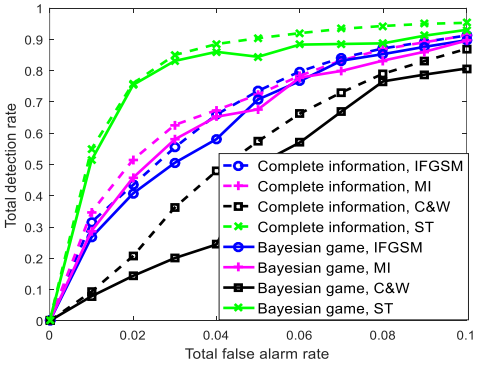

Fig. 4. Nash equilibrium ROC of the adversarial-detection games.

4.4. Bayesian game

Finally, we examine the outcome of the Bayesian game. We assume the prior belief $\boldsymbol{p}$ = [0.25, 0.25, 0.25, 0.25] for simplicity, i.e., the investigator believes that the attacker has an equal chance to mount different attacks. The NEROC curves of the Bayesian game are presented as solid lines in Fig. 4. The total detection rate of the Bayesian game has a non-negligible decline compared with its complete information counterpart, especially for C&W and ST attacks, which means that the Bayesian game is favorable for the attacker. Take the $P_{fa} = 0.03$ for example, the total detection rate of the C&W attack in the Bayesian game is 0.20, 0.16 lower than in the *AE-detection*$_d(S_I, S_A, \boldsymbol{U})$ game. The result is intuitive because the attacker has an information advantage over the investigator in the Bayesian game.

An interesting observation of the Bayesian game is that its total detection rate does not necessarily increase monotonically with P_{fa}. Take the ST attack for example, the detection rate at $P_{fa} = 0.05$ is even lower than that of $P_{fa} = 0.04$. This is because the investigator deviates further from the optimal strategy due to information disadvantage at $P_{fa} = 0.05$.

5. CONCLUSION

This paper models the interplay between an AE attacker and an investigator with games. We study the optimal strategies for both sides under complete and incomplete information scenarios. To the best of our knowledge, it is the first time that the Bayesian game has been introduced to model information asymmetry in AE detection. We solve the Bayesian game with the mixed-strategy Nash equilibrium. By comparing the Nash equilibrium ROC curves of the Bayesian game and its complete information counterpart, we can realize to what extent information asymmetry affects AE's security.

6. REFERENCES

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

[2] 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.

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

[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] 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.

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

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," International conference on learning representations, 2015.

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

[9] 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.

[10] Y. Wu, S. S. Arora, Y. Wu, et al., "Beating attackers at their own games: Adversarial example detection using adversarial gradient directions," the 35th AAAI Conference on Artificial Intelligence, 2021.

[11] H. Zeng, K. Deng, B. Chen, et al., "How secure are the adversarial examples themselves?" 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2879–2883.

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

[13] M. Barni and B. Tondi, "The source identification game: An information-theoretic perspective," IEEE Trans. Info. Forensics and Security, 8(3): 450–463, 2013.

[14] A. Kurakin, I. Goodfellow, S. Bengio, "Adversarial examples in the physical world," International Conference on Learning Representations, 2017.

[15] C. Xie, Z. Zhang, Y. Zhou, et al., "Improving transferability of adversarial examples with input diversity," In CVPR2019, 2730–2739.

[16] Y. Dong, T. Pang, H. Su, et al., "Evading defenses to transferable adversarial examples by translation-invariant attacks," In CVPR2019, 4312–4321.

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

[18] 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.

[19] 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.

[20] 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.

[21] K. Lee, K. Lee, H. Lee, et al., "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), pp. 7167–7177.

[22] 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

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

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

[25] J. Hespanha, "Noncooperative game theory: an introduction for engineers and computer scientists", Princeton University Press, 2017.

[26] 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.

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