Journal of Electronic Research and Application, 2026, Volume 10, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Survey on Artificial Intelligence Systems Robustness: Adversarial Attacks and Defenses

Wei Zheng

Sichuan Forestry and Grassland Administration, Chengdu 610081, China

Abstract: Artificial intelligence systems have achieved widespread applications across many fields such as image classification, speech recognition, and game playing. However, as their decision-making logic is primarily learned from data, their outputs are highly sensitive to data anomalies and are particularly vulnerable to adversarial perturbations. This paper conducts a comprehensive survey on the robustness of artificial intelligence systems, reviewing classical adversarial attack and defense methods, and summarizing future development trends. We hope this work can provide valuable insights for research on the robustness of artificial intelligence systems and support the development of trustworthy artificial intelligence.

Keywords: Artificial intelligence; Adversarial attacks; Adversarial defense

Online publication: February 27, 2026

1. Introduction

Artificial intelligence (AI) technologies such as deep learning, have been widely applied in many fields such as autonomous driving and healthcare, significantly enhancing productivity [1,2]. They are poised to exert even more extensive and profound impacts on production methods and lifestyles across various industries in the future. However, unlike traditional software that acquires decision-making logic through explicit programming, the decision-making logic of AI systems is primarily learned by pre-defined model structures from training data. Developers can only indirectly influence the system's decision-making logic by modifying training data, features, and architectural details of the model (e.g., the number of layers) [3]. Consequently, AI systems are highly sensitive to data anomalies, vulnerable to biased data, and face significant challenges in robustness.

A study by Szegedy et al. revealed that deep learning models can produce completely erroneous outputs when subjected to deliberate, minimal perturbations in their inputs, changes that are nearly imperceptible to humans [4]. In light of this, the academic community sparked a research boom on the robustness of artificial intelligence, resulting in a variety of robustness analysis methods, techniques, and tools. While several studies have conducted surveys on the robustness of AI systems—for instance, Hamon et al. extended the discussion from technical aspects to policy recommendations, addressing robustness and explainable challenges in real-world AI applications [5]. Javed systematically examined robustness issues in deep learning systems for medical diagnostics; and Tocchetti provided a systematic analysis of research progress, challenges, and future directions in AI robustness from a human-centered perspective [6,7]. However, those studies remain a lack of systematic investigation into AI robustness specifically from the angles of adversarial attacks and defenses. For example, Wang et al. focused solely on adversarial attacks and defenses in communication application classification models based on deep neural networks, lacking a broader survey and thus falling short of effectively guiding the design of general adversarial defense strategies [8].

To address the aforementioned challenges, this paper conducts a systematic analysis of current robustness research from the perspectives of adversarial attacks and defenses, focusing on typical AI systems such as deep learning systems. The study aims to inspire developers to design more effective defense mechanisms, thereby supporting the further application of trustworthy artificial intelligence technologies. The main contributions of this paper can be summarized as follows:

(1) It reviews adversarial attack and defense techniques for artificial intelligence systems;

(2) It builds a bridge between academic researchers and application engineers, taking appropriate measures to enable potential in-depth collaboration in the future.

The remainder of this paper is structured as follows: Section 2 introduces related basic concepts; Sections 3 and 4 review current research on adversarial attacks and defenses in AI system, respectively; finally, Section 5 provides a comprehensive summary of this paper.

2. Basic conception

2.1. Artificial intelligence robustness

AI algorithms and traditional software lack physical entities and realize their intended functions through programs. Therefore, AI algorithms can be regarded as a special category of software. This paper refers to software systems deploying at least one AI algorithm as artificial intelligence systems. However, the methods by which traditional software systems and AI systems acquire their decision-making logic are fundamentally different [3]. In traditional software, decision-making logic is directly determined by the control flow and data flow defined by developers in the program. As a result, its operational outcomes are robust. In contrast, the development of AI systems follows a completely different paradigm: after developers define the architecture (such as a deep learning model), the decision-making logic is autonomously learned by the system from training data. Developers cannot directly influence this decision-making logic, as illustrated in **Figure 1**. Consequently, the decision-making logic of AI systems is not explainable and is highly sensitive to data anomalies, making them susceptible to biased data. This inherent characteristic has spurred related research into robustness. The academic community holds various definitions for the robustness of AI systems. For instance, Mannor et al. define machine learning robustness as the bounded difference in the loss function between any subsets of the training and test sets [9]. Despite varying definitions, the robustness of an AI system can be broadly defined as the stability of its performance when confronted with anomalous inputs, that is, the system's tolerance to variations in data.

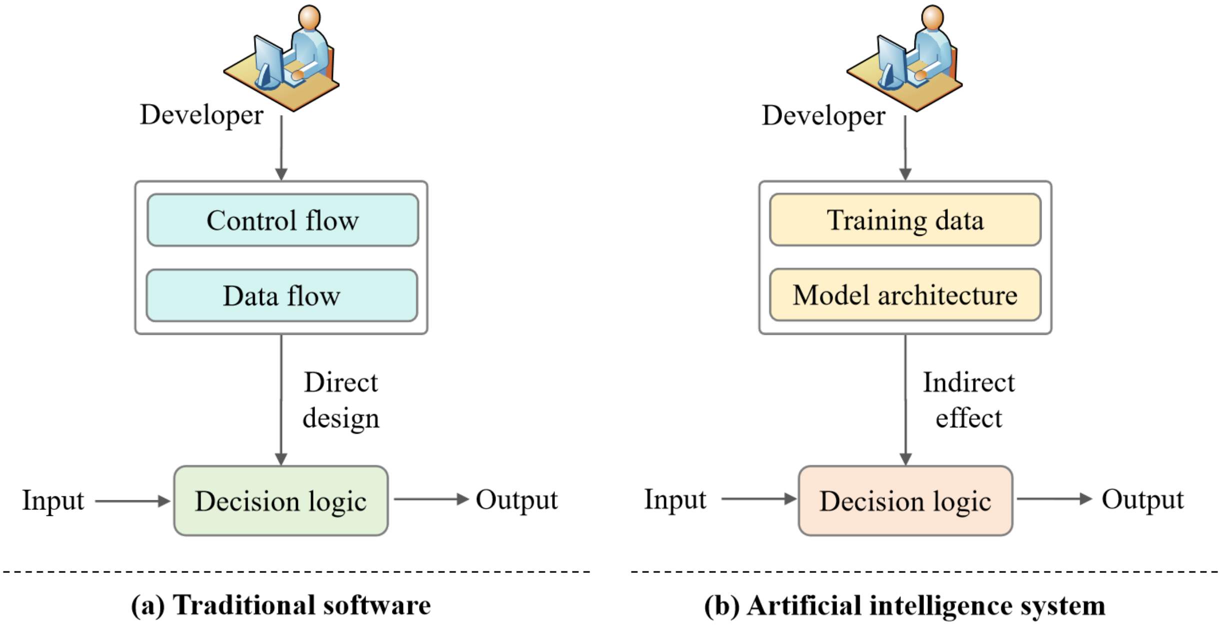

Figure 1. The differences in decision logic acquisition between traditional software and AI algorithms.

2.2. Adversarial attacks and defenses

Szegedy et al. firstly demonstrate that introducing carefully crafted minimal perturbations to images could mislead deep learning models into making completely erroneous decisions [4]. Such deliberately designed input perturbations are generally referred to as adversarial attacks. This technique involves generating adversarial examples by adding subtle perturbations to input data and using these examples to maliciously attack AI systems, such as deep neural networks, causing the system to produce incorrect or biased outputs. The objective of adversarial attacks is to deceive AI systems, preventing them from accurately performing tasks such as classification or regression, thereby compromising the system's robustness. The general representation of an adversarial attack is expressed as $F(x+\eta) \neq F(x)$, where x represents the original sample, η denotes the added perturbation, $x+\eta$ is the resulting adversarial example, and $F(\bullet)$ represents the AI model. It is important to note that the perturbation η is typically imperceptible to humans and does not affect human judgment, yet it can mislead the AI system's outputs. Since Szegedy's seminal work, researchers have discovered that adversarial examples widely exist in almost all deep learning models [10]. Beyond the field of computer vision, AI systems across various domains such as natural language processing, audio and video recognition, recommendation systems, and large language models also face threats from adversarial examples [11–15].

3. Adversarial attacks

Based on the underlying principles of adversarial attack methods, they can be classified into three main categories: gradient-based attacks, optimization-based attacks, and generation-based attacks, as illustrated in **Figure 2**.

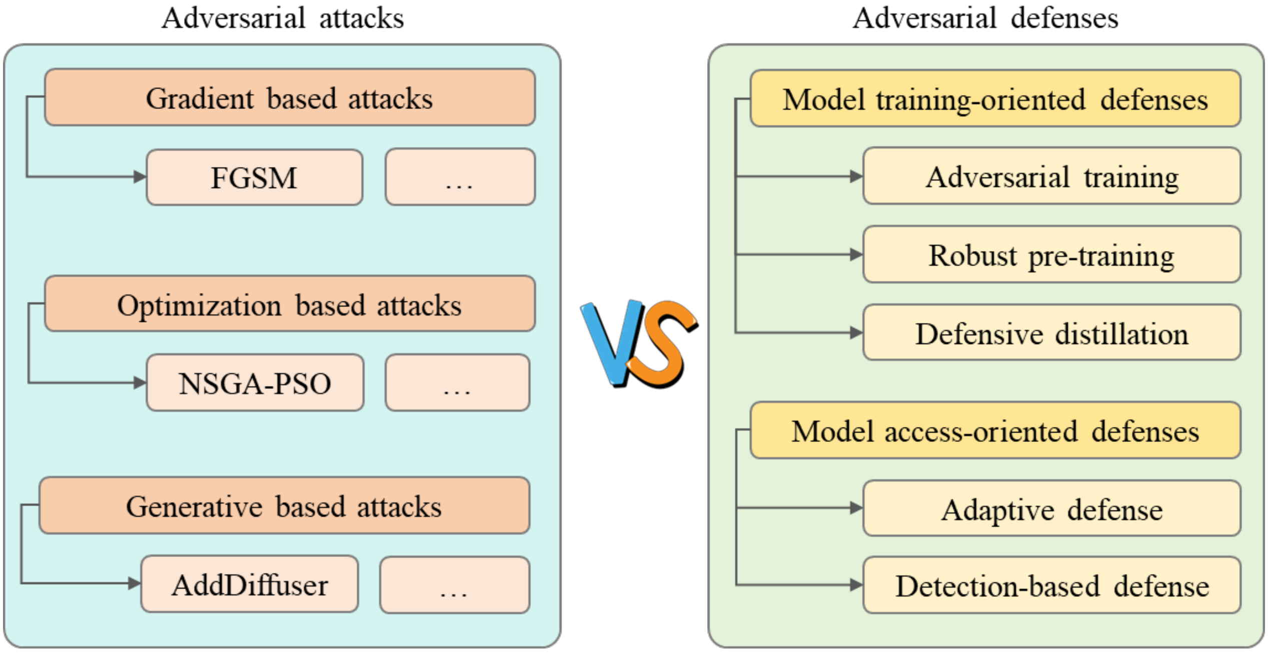

Figure 2. The adversarial attacks and defense methods investigated in this study.

3.1. Gradient-based attacks

Artificial intelligence models, such as deep neural networks, are typically trained by minimizing a loss function. Gradient-based attack methods exploit this very principle by maximizing the loss function to generate adversarial examples. Specifically, these methods compute the gradient of the loss function with respect to the input data, then apply minimal perturbations to the data along the direction (or the opposite direction) of this gradient. This causes the perturbed input data to induce significant errors in the system's predictions.

Classic gradient-based attack methods primarily include the fast gradient sign method (FGSM), its variant R + FGSM, the basic iterative method (BIM), the projected gradient descent (PGD) method, and the Momentum Iterative Method (MIM) [16–20]. The FGSM normalizes the gradient by applying the sign function, *sign* (), causing perturbations to be added in the direction of the gradient of the loss function, and maximizes the loss and thereby deceives the system [16]. To address the limitation of FGSM being easily defended, Kurakin et al. proposed the variant R + FGSM, which incorporates random noise during the generation of adversarial examples to evade gradient masking-based defense strategies [17]. The BIM starts from the original input sample and iteratively applies small perturbations to the current adversarial example to produce the next sample [18]. Through multiple iterations, it gradually increases the magnitude and number of perturbations to more effectively deceive the model. The PGD method is similar to BIM, but it includes an additional random initialization step, and the random perturbation is generated by uniformly sampling random noise within the perturbation threshold [19]. The MIM introduces momentum during the iterative process to accumulate gradient directions, preventing the adversarial perturbation from falling into poor local optima and resulting in more stable updates to the adversarial perturbation [20].

Among the methods mentioned above, the FGSM is characterized by its simple principle and rapid execution, though it achieves a relatively lower attack success rate. Methods such as R + FGSM, the BIM, and PGD are all improvements upon FGSM, while they enhance the attack success rate, they also tend to cause overfitting and exhibit poor generalization ability. The MIM mitigates the overfitting issue to some extent by incorporating momentum into the gradient iteration process. In generall, most gradient-based attack methods focus on enhancing attack potency and transferability. When models are adversarially trained using examples generated by such methods, they tend to develop good robustness against similar gradient-based attacks. However, their generalized robustness against other types of adversarial attacks remains limited.

3.2. Optimization-based attacks

Different from gradient-based methods, the optimization-based attack approaches frame the generation of adversarial examples as a constrained optimization process. These methods employ optimization algorithms to identify the optimal adversarial perturbation, one that minimizes the magnitude of perturbation while maximizing the loss, to create adversarial examples. In contrast to gradient-based techniques, these methods do not explicitly compute gradients; instead, the computation and backpropagation of gradients are embedded within the optimizer. This design allows them to achieve both high attack accuracy and low perturbations.

Carlini and Wagner (C&W) is a classic optimization-based attack method. It transforms the constrained optimization problem into an unconstrained one, uses optimized parameters to represent adversarial examples, and applies optimization techniques to precisely attack the target model. Feng et al. proposed an adversarial attack method based on a non-dominated sorting genetic algorithm and particle swarm optimization (NSGA-PSO), which generates adversarial digital watermarks in black-box attack scenarios. This approach demonstrates strong generalization capability and exhibits great robustness against image transformation defense methods [21,22].

Optimization-based attack methods generally exhibit good transferability across different network architectures and show strong resistance to example transformation defense strategies. Moreover, as they do not rely on access to predictions and labels, they are more suitable for real-world scenarios. However, the performance of these methods may vary significantly across different datasets.

3.3. Generation-based attacks

Generation-based attack methods typically involve generative adversarial networks (GANs) or other generative models. These approaches learn the distribution of real data and subsequently generate adversarial examples that closely resemble original samples yet are capable of deceiving the target AI system. The adversarial transformation network (ATN), the generative adversarial perturbation (GAP) model, and AddDiffuser are three classic generation-based attack methods [23–25].

The ATN was the first method to employ a generative model for producing adversarial examples [23]. It first uses an autoencoder to convert input samples into adversarial perturbations, then superimposes these perturbations onto the original clean images. A loss function is utilized to guide the generator in performing targeted attacks. The GAP represents a further development based on ATN [24]. It enhances the network architecture by adopting a U-Net structure to generate adversarial perturbations. Additionally, this method decouples the perturbation magnitude constraint from the adversarial attack loss, allowing more fine-grained control over perturbation generation while optimizing adversarial effectiveness. AddDiffuser is a novel approach that utilizes a diffusion model to generate adversarial examples [25]. This method guides the latent code into the adversarial example space of a specific classifier through perturbed predicted images and employs adversarial repair based on class activation mapping to preserve salient regions of the image while perturbing less important areas.

Compared to the other two categories of methods, generation-based attacks do not require repeated access to the target system, resulting in higher attack efficiency. This makes them more suitable for generating adversarial examples in large quantities.

4. Adversarial defenses

To counter diverse and potent adversarial attacks, numerous defense methods have been successively proposed to enhance the robustness of AI systems when facing anomalous inputs. Current adversarial defense designs for AI systems are primarily implemented during two stages: model training-oriented and model access-oriented defenses.

4.1. Model training-oriented defenses

The model training-oriented defense methods refer to techniques applied during the model training stage, where adversarial examples are introduced as input. This enables the model to learn characteristics of adversarial perturbations, thereby equipping the final trained model with inherent capability to recognize adversarial samples and consequently improving system robustness. Such methods primarily include defense paradigms such as adversarial training, robust pre-training, and defensive distillation.

4.1.1. Adversarial training

These methods operate in a self-play manner by incorporating adversarial examples into the training dataset, enabling the model to learn how to resist adversarial attacks. Specifically, during each training epoch, adversarial examples are generated against the current model and included in training. This approach ensures that while the model's robustness is progressively enhanced through each iteration, it also continuously faces increasingly potent adversarial examples, thereby systematically improving the overall robustness of the system. Classical adversarial training methods include ensemble adversarial training, cascaded adversarial training, and projected gradient descent adversarial training [17,19,26]. (see **Figure 3**)

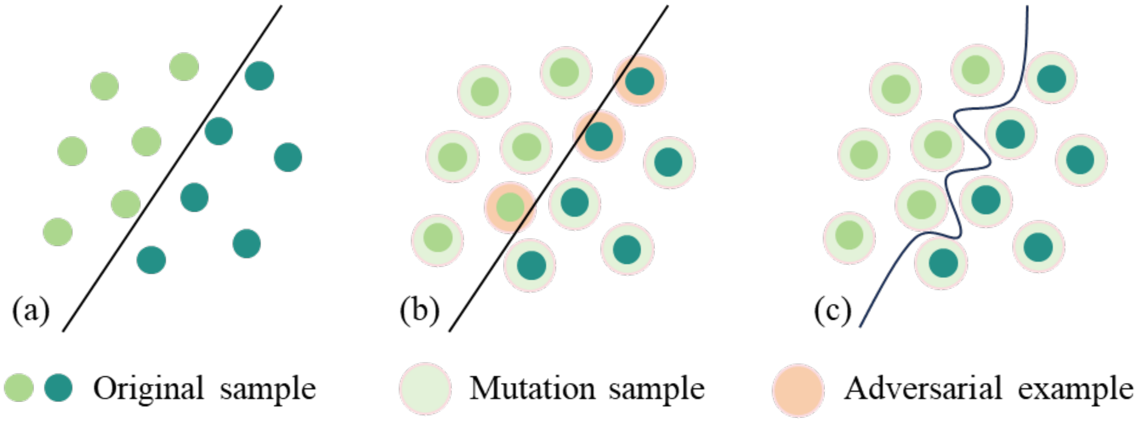

Figure 3. Principle of Adversarial Training. (a) demonstrates the result of the original model on clean samples. (b) shows the result of the original model on perturbed samples, some of which are adversarial examples. (c) illustrates the result of the adversarially trained model on perturbed samples, which can effectively recognize variations and remains robust against adversarial attack.

4.1.2. Robust pre-training

This category of defense methods draws inspiration from the pre-training-fine-tuning paradigm by designing a robust pre-training followed by adversarial fine-tuning strategy. During the pre-training phase, the model is trained on large-scale, diverse datasets to learn more generalized and robust feature representations. Subsequently, the pre-trained model is fine-tuned using adversarial examples to enhance its robustness against such attacks. Compared to traditional adversarial training, the robust pre-training and adversarial fine-tuning paradigm significantly improves both the model's performance on clean samples and its robustness against adversarial examples. Classical robust pre-training methods include dual-path adversarial contrastive pre-training and adversarial contrastive learning [27,28]. The key to robust pre-training lies in designing superior pre-training strategies that enable the model to learn more effective pre-trained features, thereby achieving better performance in the subsequent adversarial fine-tuning stage [28].

4.1.3. Defensive distillation

Defensive distillation draws upon the concept of knowledge distillation, where knowledge is transferred from a pre-trained teacher model to a student model, enabling the student model to maintain recognition capability when confronted with adversarial examples. Classical defensive distillation methods include AGKD-BML and guided adversarial contrastive distillation (GACD) [29,30].

4.2. Model access-oriented defenses

The model access-oriented defense methods refer to techniques applied to AI systems that lack defensive mechanisms incorporated during the training phase. These methods process either the input samples or the model itself during the system's access (deployment) stage, thereby enabling models originally incapable of handling adversarial examples to function normally. This category primarily includes adaptive defense and detection-based defense approaches.

4.2.1. Adaptive defense

Adaptive defense methods refer to techniques that process either the input data or the model architecture to mitigate attacks from adversarial examples. These can be further categorized into data-adaptive and model-adaptive defense methods. Data-adaptive defense methods involve preprocessing input samples using statistical approaches or prior knowledge to eliminate potential adversarial perturbations, such as image compression and image denoising [31,32]. However, since these methods apply uniform processing to all input samples, they inevitably degrade the model's performance on clean data while enhancing robustness against adversarial attacks. Model-adaptive defense methods involve modifying the model structure to defend against adversarial examples, such as the supergrid method [33]. Since these approaches require retraining the model after architectural modifications, they generally entail higher computational costs and lower efficiency.

4.2.2. Detection-based defense

Detection-based defense methods treat adversarial examples as anomalous inputs, detecting and subsequently rejecting them from processing [34,35]. The core of these approaches lies in accurately identifying adversarial samples. Based on the detection techniques employed, they can be further classified into statistical-discrepancy-based detection methods and statistical-modeling-based detection methods. The former involves extracting features from input samples using statistical methods and detecting adversarial examples based on discrepancies in these features. Commonly used statistical measures include density ratio, mutual information, and activation function invariance. The latter refers to modeling the differences between genuine samples and adversarial examples using techniques such as deep learning models such as binary classification detectors, followed by detection. The advantage of detection-based defense methods is that they require no modifications to the input data or the target model, resulting in relatively lower complexity.

5. Conclusion

Research on adversarial attacks against AI systems focuses on developing efficient algorithms to generate adversarial examples in various scenarios, thereby revealing vulnerabilities in embedded deep learning models and other AI components. On the other hand, adversarial defense research aims to mitigate the harm caused by such attacks by enhancing model robustness, thereby advancing the development of secure and reliable artificial intelligence. This paper surveys current research on the robustness of AI systems from the perspectives of adversarial attacks and defenses. It reveals that as the interplay between attacks and defenses evolves, their respective technical capabilities continue to deepen, significantly promoting progress in this field. Through a systematic investigation, we identify several promising research directions. For adversarial attacks, future work could focus on:

(1) Multimodal adversarial attacks

Current research predominantly targets image data, with relatively less attention to text, audio, and other data types. Developing effective multimodal attack techniques based on existing foundations remains a challenge.

(2) High-efficiency and low-cost adversarial attacks

Effective attack methods should achieve high success rates while maintaining low computational cost. These objectives often conflict, making their balanced realization a significant challenge.

For adversarial defenses, beyond developing countermeasures against emerging attack methods, greater emphasis should be placed on the generalization capability of defense strategies. Customized defenses may struggle against rapidly evolving attack techniques, underscoring the need for versatile and broadly applicable adversarial defense methods.

The analysis and discussion presented in this paper can provide valuable insights to help researchers develop their own methods and tools for AI system adversarial attacks and defenses, ultimately contributing to the creation of more secure and trustworthy artificial intelligence for end-users.

Disclosure statement

The author declares no conflict of interest.

Reference

[1] Zhao J, Zhao W, Deng B, et al., 2024, Autonomous Driving System a Comprehensive Survey. Expert Systems with Applications, 242: 122836.

[2] Al Kuwaiti A, Nazer K, Al-Reedy A, et al., 2023, A Review of the Role of Artificial Intelligence in Healthcare. Journal of Personalized Medicine, 13(6): 951.

[3] Pei K, Cao Y, Yang J, et al., 2017, DeepXplore: Automated Whitebox Testing of Deep Learning Systems. Proceedings of the 26th Symposium on Operating Systems Principles: 1–18.

[4] Szegedy C, Zaremba W, Sutskever I, et al., 2013, Intriguing Properties of Neural Networks. arXiv Preprint arXiv:1312.6199.

[5] Hamon R, Junklewitz H, Sanchez I, 2020, Robustness and Explainability of Artificial Intelligence. Publications Office of the European Union, 207(40): 1–40.

[6] Javed H, El-Sappagh S, Abuhmed T, 2024, Robustness in Deep Learning Models for Medical Diagnostics: Security

and Adversarial Challenges Towards Robust AI Applications. Artificial Intelligence Review, 58(1): 12.

[7] Tocchetti A, Corti L, Balayn A, et al., 2025, AI Robustness: A Human-Centered Perspective on Technological Challenges and Opportunities. ACM Computing Surveys, 57(6): 1–38.

[8] Wang Y, Sun T, Li S, et al., 2023, Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey. IEEE Communications Surveys & Tutorials, 25(4): 2245–2298.

[9] Xu H, Mannor S, 2012, Robustness and Generalization. Machine Learning, 86(3): 391–423.

[10] Xu H, Ma Y, Liu H, et al., 2020, Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. International Journal of Automation and Computing, 17(2): 151–178.

[11] Goyal S, Doddapaneni S, Khapra M, et al., 2023, A Survey of Adversarial Defenses and Robustness in NLP. ACM Computing Surveys, 55(14s): 1–39.

[12] Zhang W, Sheng Q, Alhazmi A, et al., 2020, Adversarial Attacks on Deep-Learning Models in Natural Language Processing: A Survey. ACM Transactions on Intelligent Systems and Technology, 11(3): 1–41.

[13] Carlini N, Wagner D, 2018, Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. IEEE Security and Privacy Workshops: 1–7.

[14] Deldjoo Y, Noia T, Merra F, 2021, A Survey on Adversarial Recommender Systems: From Attack and Defense Strategies to Generative Adversarial Networks. ACM Computing Surveys, 54(2): 1–38.

[15] Shayegani E, Mamun M, Fu Y, et al., 2023, Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. arXiv Preprint arXiv:2310.10844.

[16] Goodfellow I, Shlens J, Szegedy C, 2014, Explaining and Harnessing Adversarial Examples. arXiv Preprint arXiv:1412.6572.

[17] Tramèr F, Kurakin A, Papernot N, et al., 2017, Ensemble Adversarial Training: Attacks and Defenses. arXiv Preprint arXiv:1705.07204.

[18] Kurakin A, Goodfellow I, Bengio S, 2018, Adversarial Examples in the Physical World. Artificial Intelligence Safety and Security. Chapman and Hall/CRC: 99–112.

[19] Madry A, Makelov A, Schmidt L, et al., 2017, Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv Preprint arXiv:1706.06083.

[20] Dong Y, Liao F, Pang T, et al., 2018, Boosting Adversarial Attacks with Momentum. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 9185–9193.

[21] Carlini N, Wagner D, 2017, Towards Evaluating the Robustness of Neural Networks. IEEE Symposium on Security and Privacy: 39–57.

[22] Feng S, Feng F, Xu X, et al., 2021, Digital Watermark Perturbation for Adversarial Examples to Fool Deep Neural Networks. International Joint Conference on Neural Networks: 1–8.

[23] Baluja S, Fischer I, 2017, Adversarial Transformation Networks: Learning to Generate Adversarial Examples. arXiv Preprint arXiv:1703.09387.

[24] Poursaeed O, Katsman I, Gao B, et al., 2018, Generative Adversarial Perturbations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 4422–4431.

[25] Chen X, Gao X, Zhao J, et al., 2023, AdvDiffuser: Natural Adversarial Example Synthesis with Diffusion Models. IEEE/CVF International Conference on Computer Vision: 4562–4572.

[26] Na T, Ko J, Mukhopadhyay S, 2017, Cascade Adversarial Machine Learning Regularized with a Unified Embedding. arXiv Preprint arXiv:1708.02582.

[27] Hendrycks D, Lee K, Mazeika M, 2019, Using Pre-Training Can Improve Model Robustness and Uncertainty. International Conference on Machine Learning: 2712–2721.

[28] Jiang Z, Chen T, Chen T, et al., 2020, Robust Pre-Training by Adversarial Contrastive Learning. Advances in Neural Information Processing Systems, 33: 16199–16210.

[29] Wang H, Deng Y, Yoo S, et al., 2021, AGKD-BML: Defense Against Adversarial Attack by Attention-Guided Knowledge Distillation and Bi-Directional Metric Learning. IEEE/CVF International Conference on Computer Vision: 7658–7667.

[30] Bai T, Zhao J, Wen B, 2023, Guided Adversarial Contrastive Distillation for Robust Students. IEEE Transactions on Information Forensics and Security, 19: 9643–9655.

[31] Guo C, Rana M, Cisse M, et al., 2017, Countering Adversarial Images Using Input Transformations. arXiv Preprint arXiv:1711.00117.

[32] Liao F, Liang M, Dong Y, et al., 2018, Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. IEEE Conference on Computer Vision and Pattern Recognition: 1778–1787.

[33] Bian H, Chen D, Zhang K, et al., 2021, Adversarial Defense via Self-Orthogonal Randomization Super-Network. Neurocomputing, 452: 147–158.

[34] Alotaibi A, Rassam M, 2023, Adversarial Machine Learning Attacks Against Intrusion Detection Systems: A Survey on Strategies and Defense. Future Internet, 15(2): 62.

[35] Aldahdooh A, Hamidouche W, Fezza S, et al., 2022, Adversarial Example Detection for DNN Models: A Review and Experimental Comparison. Artificial Intelligence Review, 55(6): 4403–4462.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.