iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57752

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Security Issues and Defensive Approaches in Deep Learning Frameworks

Shashank R R[1], Indudhara S[2], Abhijith Hegde[3], Adarsha N Hirematha[4]
[1]*B.E in Computer Science & Engineering, Dept of CSE, Jawaharlal Nehru New College of Engineering*
[2]*B.E in Computer Science & Engineering, Dept of CSE, Jawaharlal Nehru New College of Engineering*
[3]*B.E in Information Science & Engineering, Dept of ISE, NMAM Institute of Technology*
[4]*B.E in Information Science & Engineering, Dept of ISE, Jawaharlal Nehru New College of Engineering*

Abstract: ***Deep learning frameworks are instrumental in advancing artificial intelligence, showcasing vast potential across diverse applications. Despite their transformative impact, security concerns pose significant risks, impeding widespread adoption. Malicious internal or external attacks on these frameworks can have far-reaching consequences for society.***
Our research delves into the intricacies of deep learning algorithms, conducting a thorough analysis of vulnerabilities and potential attacks. To address these challenges, we propose a comprehensive classification system for security issues and corresponding defensive approaches within deep learning frameworks. By establishing clear connections between specific attacks and their defenses, we aim to enhance the robustness of these frameworks.
In addition to theoretical considerations, our study extends to real-world applications, exploring a case involving the practical implications of deep learning security issues. Looking ahead, we discuss future directions and open issues within the realm of deep learning frameworks, aspiring to inspire further developments. We anticipate that our research will not only contribute valuable insights but also attract attention from both academic and industrial domains, fostering a collective commitment to fortifying the security of deep learning frameworks.
Keywords: adversarial examples; deep learning frameworks; defensive approaches; security issues

I. INTRODUCTION

The widespread success of deep learning across various domains [1–3] has fueled increased interest in Artificial Intelligence (AI). The advent of Graphics Processing Units (GPUs) has empowered deep learning algorithms and large-scale datasets to tackle diverse challenges. This technological progress has facilitated practical applications in fields ranging from Information Technology (IT) to automotive industries, with companies like Google, Tesla, Baidu, Mercedes, and Uber actively testing driverless cars that rely on deep learning techniques.

Furthermore, the integration of deep learning algorithms has become pivotal in various systems, exemplified by major phone manufacturers incorporating facial authentication features for unlocking devices. Additionally, a multitude of behavior-based malware and anomaly detection solutions now leverage deep learning [4, 5].

While deep learning offers significant advantages and convenience, it is not without its vulnerabilities. Recent research has revealed that well-designed adversarial samples can exploit vulnerabilities in deep learning models, effectively deceiving even well-behaved models. This susceptibility highlights the need for ongoing efforts to enhance the robustness of deep learning systems in the face of potential adversarial threats.

In their groundbreaking work, Szegedy et al. [6] pioneered the creation of slight modifications in image data, effectively tricking even the most sophisticated Deep Neural Networks (DNNs) with a high success rate. Consequently, instances misclassified by a DNN are termed adversarial samples.

The generation of adversarial samples involves manipulating the model's structure and parameters to disrupt the processes or induce incorrect predictions within the deep learning model. This form of attack, which encompasses techniques such as obfuscated gradient [7] and root mean square gradient [8], is referred to as a white-box attack. In a white-box attack, the attacker possesses knowledge about the internal structure and parameters of the targeted model.

In contrast, a black-box attack is constrained by a lack of information regarding the model's internal structure. Unlike white-box attacks, black-box attacks rely on limited knowledge about the targeted model, making them distinct in terms of their approach and constraints.

Goodfellow et al. [9] emphasized the neural network's susceptibility to slight input disturbances and introduced the Fastest Gradient Sign Method (FGSM) as a means to easily generate adversarial samples. Su et al. [10] extended this by proposing a black-box Deep Neural Network (DNN) attack, achieving effective results with minimal perturbations to a single pixel across various image sizes.

In response to these adversarial threats, various defense measures have been proposed. For instance, Goodfellow et al. [9] introduced the gradient masking method. He et al. [11] argued for the inadequacy of single-defense methods, advocating for a comprehensive defense system comprising multiple measures to better counter adversarial examples.

The impact of adversarial examples extends beyond theoretical concerns, with real-world applications in security-critical environments. Notably, adversaries can create physical adversarial examples to confound autonomous vehicles, such as manipulating a traffic sign recognition system [11].

Existing surveys on deep learning framework security have taken different perspectives. Xu et al. [12] classified issues based on black-box/white-box attacks, poisoning attacks, and escape attacks. Tariq et al. [13] categorized attacks into causative, exploratory, targeted, and indiscriminate types. However, these surveys lacked a comprehensive and systematic view of security and defense approaches in deep learning frameworks.

In contrast, our classification approach is based on attack phase, adversarial knowledge, attack frequency, attack target, and attack scope, providing a more nuanced understanding. Bae et al. [14] discussed deep learning security and privacy issues using mathematical principles, while our approach includes visual illustrations of attack principles and mechanisms. Qiu et al. [15] discussed AI attack methods in training and testing phases but did not establish a clear connection between attacks and corresponding defense technologies. In our research, we establish a one-to-one connection between attacks and defense technologies, offering a more cohesive understanding.

Furthermore, we not only analyze current security issues but also explore future directions and open challenges within deep learning frameworks. This comprehensive and methodical research aims to contribute a thorough analysis of the security landscape in deep learning frameworks and serve as a valuable resource for further exploration and development in this critical domain.

The paper is structured as follows:

In Section 2, we provide an introduction to general deep learning models and processes.

Section 3 delves into deep learning principles, highlighting vulnerabilities, and discussing types of attacks stemming from third-party libraries.

Section 4 offers a classification of attacks from various perspectives.

Section 5 comprehensively details defense measures designed to counteract a range of attacks.

Section 6 focuses on a specific application scenario within deep learning—automatic driving. This includes the identification of traffic signs and an analysis of associated security challenges.

Finally, in Section 7, we conclude the study and outline potential future research directions.

II. OBJECTIVES

The objectives of exploring security issues and defensive approaches in deep learning frameworks are multi-faceted and comprehensive. The primary goals include:

1) *Identification of Security Risks:* Systematically identify and understand potential security risks and vulnerabilities within deep learning frameworks. This involves analyzing the underlying algorithms, model structures, and dependencies on third-party libraries.
2) *Classification and Categorization:* Develop a robust classification system for different types of attacks on deep learning frameworks. This includes categorizing attacks based on various criteria such as adversarial knowledge, attack frequency, target, and scope.
3) *Defensive Strategies and Measures:* Propose and evaluate effective defensive strategies to mitigate the impact of identified security risks. This involves exploring and developing techniques that enhance the robustness of deep learning models against adversarial attacks.
4) *Comprehensive Analysis:* Conduct a thorough and systematic analysis of existing security issues in deep learning frameworks. This includes examining the vulnerabilities introduced by third-party libraries and understanding how they can be exploited in real-world scenarios.
5) *Application to Real-World Scenarios:* Investigate the practical implications of security issues by analyzing a specific application scenario, such as deep learning in automatic driving. This allows for the identification of potential security challenges and the development of context-specific defensive measures.

6) *Guidance for Future Research:* Conclude the study by summarizing key findings and proposing future research directions in the realm of deep learning framework security. This serves as a guide for researchers, practitioners, and policymakers to address emerging challenges and advance the field.
7) *Promotion of Awareness:* Raise awareness in academic and industrial domains about the critical importance of security in deep learning frameworks. By highlighting potential risks and providing effective defensive approaches, the research aims to foster a proactive approach to security considerations.
8) *Integration of Visual and Practical Insights:* Incorporate visual illustrations and practical insights to enhance the understanding of security issues and defensive mechanisms. This approach facilitates clear communication of complex concepts and promotes a more intuitive grasp of the subject matter.

III. LIMITATIONS

While the exploration of security issues and defensive approaches in deep learning frameworks provides valuable insights, it is essential to acknowledge certain limitations. The dynamic nature of the cyber security landscape may result in emerging threats not fully covered by existing research. Additionally, the effectiveness of defensive measures could be contingent on the specific deep learning applications and evolving attack strategies. The scalability of proposed defenses across different frameworks and their adaptability to rapidly evolving technologies may pose challenges. Furthermore, the integration of defensive measures could potentially introduce performance trade-offs, impacting the efficiency of deep learning models. Addressing these limitations requires ongoing research, collaboration between academia and industry, and a proactive approach to staying abreast of evolving security threats in the deep learning ecosystem.

IV. LITERATURE SURVEY

The study by W. W. Jiang and L. Zhang, titled "Geospatial data to images: A deep-learning framework for [1] traffic forecasting," published in Tsinghua Science and Technology in 2019, contributes to the burgeoning field of traffic forecasting by leveraging a deep-learning framework. The authors focus on the transformation of geospatial data into images, proposing a novel approach to address traffic-related challenges. By utilizing deep learning, the framework aims to enhance the accuracy of traffic forecasting models. This work underscores the interdisciplinary nature of deep learning applications, bridging geospatial data and image processing techniques for more effective traffic predictions. The findings from this study contribute to the broader understanding of deep learning's role in optimizing transportation systems and provide valuable insights for researchers and practitioners engaged in the intersection of geospatial data and traffic forecasting.

The research [2] conducted by L. Zhang, C. B. Xu, Y. H. Gao, Y. Han, X. J. Du, and Z. H. Tian, titled "Improved Dota2 lineup recommendation model based on a bidirectional LSTM," and published in Tsinghua Science and Technology in 2020, addresses the domain of online gaming strategy by introducing an enhanced recommendation model for Dota2 lineups. The authors leverage bidirectional Long Short-Term Memory (LSTM) networks to improve the predictive capabilities of the model. By incorporating bidirectional information flow, the proposed model aims to capture complex dependencies within Dota2 team compositions more effectively. This work contributes to the evolving field of recommendation systems for online gaming, offering insights into the application of advanced neural network architectures to enhance lineup recommendations in multiplayer gaming environments like Dota2. The study provides valuable implications for both the gaming industry and the broader realm of recommendation systems leveraging deep learning techniques.

In [3] the realm of soft robotics, the study conducted by H. M. Huang, J. H. Lin, L. Y. Wu, B. Fang, Z. K. Wen, and F. C. Sun, titled "Machine learning-based multi-modal information perception for soft robotic hands," and published in Tsinghua Science and Technology in 2020, presents a significant contribution to the field. By leveraging machine learning techniques, the authors focus on enhancing the multi-modal information perception capabilities of soft robotic hands. The incorporation of machine learning methods in this context aims to facilitate more adaptive and sophisticated interactions with the environment. This work provides valuable insights into the intersection of soft robotics and machine learning, offering a foundation for the development of intelligent and versatile soft robotic systems capable of perceiving and responding to diverse sensory inputs. The study's implications extend to advancements in the broader field of robotics, where the integration of machine learning enhances the capabilities of soft robotic systems in complex and dynamic environments.

[4] The research paper by X. Y. Yuan, P. He, Q. L. Zhu, and X. L. Li, titled "Adversarial examples: Attacks and defenses for deep learning," published in the IEEE Transactions on Neural Networks and Learning Systems in 2019, constitutes a comprehensive literature survey in the domain of adversarial attacks and defenses in deep learning. This study extensively reviews and analyzes

various techniques employed in generating adversarial examples that exploit vulnerabilities in deep neural networks. The authors explore a range of attack strategies and subsequently delve into defensive mechanisms proposed to mitigate the impact of adversarial attacks on deep learning models. By synthesizing and evaluating the state-of-the-art methods in both adversarial attacks and defenses, this survey contributes a nuanced understanding of the evolving landscape of adversarial examples in deep learning, providing valuable insights for researchers and practitioners working on securing deep neural networks against adversarial threats.

[5] The study by J. C. Hu, J. F. Chen, L. Zhang, Y. S. Liu, Q. H. Bao, H. Ackah-Arthur, and C. Zhang, titled "A memory-related vulnerability detection approach based on vulnerability features," and published in Tsinghua Science and Technology in 2020, presents a literature survey in the domain of memory-related vulnerability detection. Focusing on the identification of vulnerabilities in memory systems, the authors propose an approach grounded in vulnerability features. This research contributes to the ongoing efforts in enhancing cybersecurity by addressing vulnerabilities that are specifically related to memory operations. By systematically reviewing existing literature and providing insights into the characteristics of memory-related vulnerabilities, the study aims to advance the field's understanding of potential security threats in memory systems and foster the development of effective detection approaches. The findings of this survey have implications for cybersecurity researchers and practitioners engaged in fortifying systems against memory-related vulnerabilities.

The seminal paper [6] by C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, titled "Intriguing properties of neural networks," and published as an arXiv preprint in 2013, serves as a cornerstone in the literature survey on the vulnerabilities and behaviors of neural networks. This pioneering work introduces the concept of adversarial examples, demonstrating that small, carefully crafted perturbations in input data can lead to misclassifications by sophisticated neural networks. By revealing the susceptibility of neural networks to adversarial attacks, the paper sparks subsequent research in understanding and mitigating these vulnerabilities. The findings have profound implications for the robustness and security of deep learning models, shaping the trajectory of subsequent studies focused on adversarial machine learning. This work remains foundational for researchers and practitioners seeking to comprehend and address the intriguing and potentially exploitable properties of neural networks.

The research paper [7] by A. Athalye, N. Carlini, and D. Wagner, titled "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," and published as an arXiv preprint in 2018, contributes significantly to the literature survey on adversarial attacks and defenses in deep learning. This work critically assesses the efficacy of defenses based on obfuscated gradients, revealing their limitations and potential vulnerabilities. By systematically evaluating existing defense mechanisms, the authors highlight the false sense of security that may arise from relying solely on obfuscated gradients. The study underscores the need for a nuanced understanding of the shortcomings of various defense strategies and encourages researchers and practitioners to adopt more robust measures to counter adversarial examples. The findings in this paper have had a lasting impact on the ongoing discourse surrounding adversarial attacks and defenses in deep learning, guiding subsequent research endeavors aimed at developing more resilient models.

[8] The study by Y. T. Xiao, C. M. Pun, and J. Z. Zhou, titled "Generating adversarial perturbation with root mean square gradient," published as an arXiv preprint in 2019, contributes to the literature survey on adversarial attacks in deep learning. Focused on the root mean square gradient (RMSG) method, the authors investigate its efficacy in generating adversarial perturbations. By proposing and evaluating this novel approach, the paper enhances our understanding of the diverse techniques available for crafting adversarial examples. The work provides valuable insights into the landscape of adversarial attacks, informing researchers and practitioners about the potential vulnerabilities in deep learning models and the importance of developing robust defenses. This paper, within the broader context of adversarial attacks, plays a role in shaping the ongoing discourse on the security of deep learning systems.

The [9] paper by I. J. Goodfellow, J. Shlens, and C. Szegedy, titled "Explaining and harnessing adversarial examples," and published as an arXiv preprint in 2014, serves as a cornerstone in the literature survey on adversarial attacks and defenses in deep learning. This seminal work introduces the concept of adversarial examples, demonstrating that imperceptible perturbations in input data can lead to misclassifications by neural networks. The authors provide a foundational explanation of the phenomenon, shedding light on the vulnerabilities of deep learning models. Additionally, the paper proposes the Fast Gradient Sign Method (FGSM), a powerful technique for generating adversarial examples efficiently. The insights from this study have significantly influenced subsequent research, inspiring the development of numerous adversarial attack and defense strategies. The work remains pivotal in shaping our understanding of the security landscape in deep learning and continues to guide investigations into adversarial machine learning.

Table 1 Differences between neural network models.

Model	Advantage	Limitation
DNN	Simple architecture	Too many layers will lead to overfitting
CNN	Extract local features, such as image recognition	Cannot process time series data
RNN	Deal with time series features	Gradient disappearance
GAN	Generate new training data	Experience difficulties reaching Nash equilibrium

The research paper [10] by J. W. Su, D. V. Vargas, and K. Sakurai, titled "One pixel attack for fooling deep neural networks," published in the IEEE Transactions on Evolutionary Computation in 2019, makes a significant contribution to the literature survey on adversarial attacks in deep learning. The authors introduce the innovative concept of a one-pixel attack, demonstrating that a minimal perturbation to just one pixel in an image can effectively deceive deep neural networks. This study highlights the surprising vulnerability of sophisticated models to subtle manipulations, challenging conventional assumptions about adversarial attacks. The proposed one-pixel attack has implications for understanding the robustness of deep learning models and underscores the need for more resilient defenses. This paper has been influential in sparking discussions around the intricacies of adversarial attacks, inspiring further research into unconventional strategies for crafting adversarial examples.

V. DEEP LEARNING FRAMEWORK ARCHITECTURE

Deep Neural Network (DNN) processing unfolds through two distinct phases: training and prediction. In the training phase, existing data is harnessed to grasp the network's parameters, as depicted in Fig. 1. This involves minimizing the cost function by adjusting parameters using known samples. The cost function gauges the disparity between the model's predicted value and the actual value of a sample. Completing the training phase necessitates forward and backward propagations. During the feedforward phase, input traverses through the layers to compute the output. Subsequently, a gradient descent algorithm is employed to minimize the error between the output and the actual label. The results obtained in the training phase are then employed in the inference phase, where the model solely propagates the input forward, treating the output as a prediction.

Convolutional Neural Networks (CNNs), exemplified in Fig. 2, play a pivotal role in image recognition and classification. CNNs encompass operations like convolution, nonlinear transformation, pooling, and classification through fully connected layers. In contrast, Recurrent Neural Networks (RNNs), illustrated in Fig. 3, deviate from traditional forward feedback neural networks by introducing directional loops adept at handling contextual correlations among inputs. RNNs are specifically designed for processing time sequence data. Additionally, the Generating Adversarial Network (GAN) framework, featuring a discriminator (D) and a generator (G), is actively explored in the realms of image/speech synthesis and domain adaptation, as demonstrated in Fig. 4. In this framework, the generator creates synthetic data, and the discriminator determines its authenticity, reflecting the ongoing research and application of GANs in diverse domains.

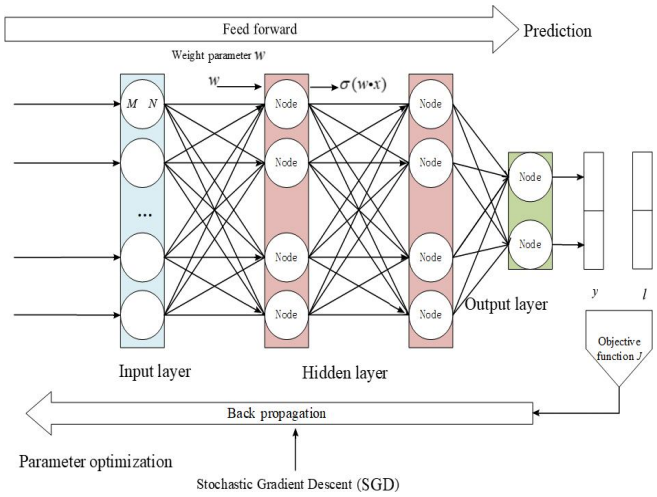


Fig. 1. General DNN training process

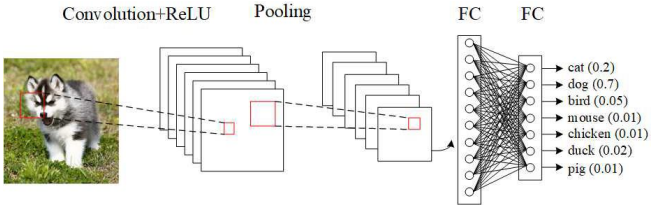


Fig. 2. CNN structure. Here FC stands for fully connected.

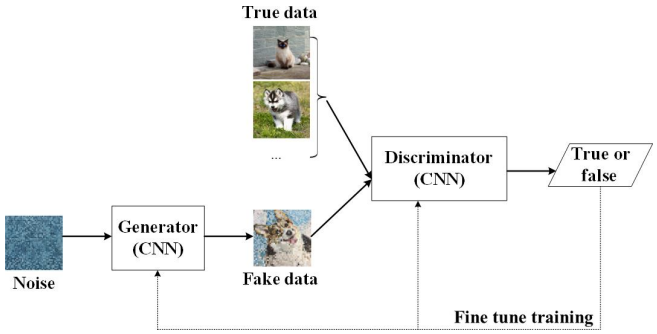


Fig. 3.GAN structure

VI. SECURITY ISSUE IN DEEP LEARNING FRAMEWORKS

A. *Adversarial Example Generation*

Goodfellow et al. [9] introduced the Fast Gradient Sign Method (FGSM) algorithm, focusing on the analysis of the reasons behind the existence of adversarial samples and proposing a method for their generation based on these analyses. The approach entails adding a small imperceptible disturbance to an image, strategically chosen to exert maximum influence on the classifier through the activation function, as illustrated in Figure 5. In this depiction, the input sample is denoted as x, and the resulting adversarial sample is represented as Qx, with · and ! being the parameters of the deep learning algorithm model. In a linear model, where the feature of the input sample is limited, the classifier struggles to differentiate between the original sample x and the adversarial sample Qx if the added perturbation value · for each element in the sample is within the accuracy of the input features. For problems with well-separated classes, the classifier tends to assign the same class to x and Qx as long as the perturbation is sufficiently small. The linear model's output is affected by the product of the weight vector ! and the adversarial sample Qx, where the adversarial perturbation increases the output of the neuron associated with the weight vector !T. Even in a high-dimensional space, small disturbances can significantly impact the final neural network output, demonstrating that linear models can also produce adversarial samples.

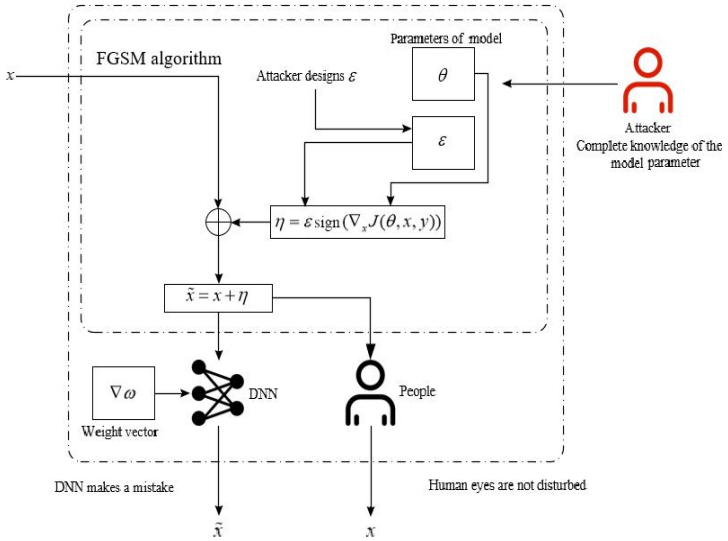


Fig.4.FGSM algorithm process diagram

In the context of a nonlinear model, the introduction of a linear perturbation involves a process governed by a nonlinear differential equation. Assuming the model parameters are denoted by n, with x representing the input, y as the associated target (classification result), and J(n, x, y) as the loss function, the goal is to linearize the loss function near the parameter value n This linearization leads to the derivation of the best max-norm constrained perturbation, given by η = " * sign(rxJ(η, x, y)). Subsequently, directly adding this linear perturbation to the original sample results in the creation of the adversarial sample Qx = x + n leading to a high misclassification rate for the neural network. This method encapsulates the adversarial example generation process of the white-box attack FGSM algorithm. Figure 6 illustrates the calculation of in the FGSM algorithm, where green points correspond to the original sample and their respective loss function values, while red points represent the adversarial sample and corresponding loss function values.

The FGSM and DeepFool[16] are methods employed for generating adversarial samples, both falling under the category of white-box attacks. In a neural network, backpropagation is typically employed for minimizing the loss function. However, the FGSM attack takes a counterintuitive approach by adding a disturbance along the gradient direction, aiming to generate an adversarial sample that maximizes the loss function and thereby deceives the neural network model.

Figure 5 showcases the outcomes of different selections of " in the FGSM algorithm. Notably, while the FGSM can determine the direction of the disturbance addition, it cannot specify the size of the disturbance, which is usually artificially determined. The disturbance direction, as illustrated in Figure 6, opposes the x-axis direction, with "1 and "2 representing two disturbance sizes. Adversarial sample x1, generated by disturbance "1, can lead to misclassification by the classification function. However, achieving the intended misclassification goal is challenging with adversarial sample x2, generated by disturbance "2.

To address this limitation, the DeepFool method enhances the FGSM by not only determining the direction of the disturbance addition but also estimating its distance. Figure 4 demonstrates the application of the DeepFool algorithm in generating an adversarial sample within a linear binary classification context. In this illustration, the classifier is represented by f(x) = !Tx + b, with ! denoting the gradient direction for the decision function, and the scalar corresponds to the optimal perturbation coefficient ". The DeepFool algorithm aims to find the optimal solution by iteratively adjusting the perturbation, overcoming the challenges posed by the FGSM in terms of both direction and magnitude of the disturbance.

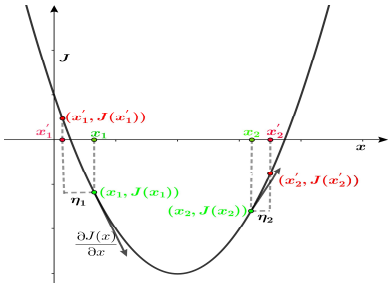


Fig.5. Calculation of η in the FGSM algorithm.

Figure 6 provides insights into the outcomes of different selections for the perturbation coefficient " in the Fast Gradient Sign Method (FGSM) algorithm. While the FGSM excels in determining the direction of the disturbance addition, it lacks the ability to specify the size of the disturbance, a parameter typically set artificially. The depicted disturbance direction opposes that of the x-axis, and "1 and "2 represent two distinct disturbance sizes. Notably, adversarial sample x1, generated with disturbance "1, can lead to misclassification by the classification function. However, the intended misclassification goal is not achieved by adversarial sample x2, generated with disturbance "2.

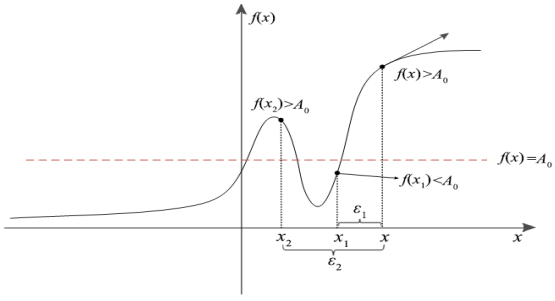


Fig.6.Results of different ε selections in the FGSM algorithm.

To address the limitation of the FGSM, the DeepFool method comes into play, offering enhancements by determining both the direction and the distance of the disturbance. Figure 6 illustrates the application of the DeepFool algorithm in generating an adversarial sample within a linear binary classification framework. In this context, the classifier is represented by f(x) = !Tx + b, where ! denotes the direction of the gradient for the decision function, and the scalar corresponds to the optimal perturbation coefficient ". The DeepFool algorithm iteratively adjusts the perturbation, overcoming the FGSM's limitations and providing more control over both the direction and magnitude of the disturbance.

B. Vulnerabilities of Deep Learning Frameworks

Adversarial sample attacks pose a significant challenge to deep learning, and the associated frameworks, including TensorFlow, Caffe, and Torch, exhibit various security issues. The use of these frameworks provides a streamlined approach for application developers, allowing them to focus on application development without delving into the intricacies of underlying implementation details, thereby enhancing the efficiency of AI applications. However, the efficiency gains are counteracted by the inherent complexity of these deep learning frameworks, and as the system complexity increases, so does the likelihood of encountering security risks. Notably, TensorFlow, Caffe, and Torch rely heavily on numerous third-party open-source libraries. An in-depth analysis of these libraries revealed multiple network security vulnerabilities susceptible to denial of service, escape, and system damage attacks. These vulnerabilities, including those related to memory access cross-border issues, present opportunities for hackers to execute various network attacks, manipulate data streams, and deceive AI applications. The complexity and reliance on third-party libraries contribute to the vulnerability landscape, emphasizing the need for robust security measures in deep learning frameworks.

VII. ATTACK CLASSIFICATION IN DEEP LEARNING FRAMEWORKS

Table 2 CVE in deep learning frameworks.

Deep learning framework	CVE-ID	Type
Caffe	CVE-2017-9782	Heap overflow
Caffe/Torch	CVE-2017-12600	Denial of service
Caffe/Torch	CVE-2017-12604	Software crash
TensorFlow	CVE-2017-12852	Out of bounds
TensorFlow	CVE-2018-7577	Memcpy param overlap
TensorFlow	CVE-2018-10055	Heap buffer overflow
TensorFlow	CVE-2019-9635	Denial of service
TensorFlow	CVE-2020-5215	Denial of service

According to the attack phase, adversarial attacks in deep learning can be categorized into poisoning and evasion attacks. Poisoning attacks involve the addition of adversarial data to the training sample, influencing the training process of the classifier and leading to the acquisition of an incorrect classifier. On the other hand, evasion attacks use adversarial examples during the inference stage to prompt the classifier to produce erroneous outputs. In terms of adversarial knowledge, attacks are classified into white-box attacks, black-box attacks, and semi-white-box attacks. A white-box attack occurs when the attacker possesses complete knowledge of the deep learning system, including the dataset, algorithm, network structure, and more. A semi-white-box attack involves partial knowledge, while a black-box attack is executed without any prior knowledge of the system.

White-box attacks, although comprehensive, are often impractical in real-life scenarios. Black-box attacks can further be categorized into transfer-based, score-based, and decision-based attacks. Transfer-based attacks involve training a local model and using the generated adversarial samples to attack the target model. Score-based attacks aim to gain information within a model by assessing the classification confidence of the target model. Decision-based attacks, while practical, are the most challenging, as they only obtain the classification result of a model on the input.

In terms of attack frequency, attacks can be classified as one-time and iterative attacks. One-time attacks require a single instance to generate adversarial samples, while iterative attacks necessitate multiple iterations to update the adversarial samples. Compared with iterative attacks, one-time attacks are characterized by their efficiency and simplicity in generating adversarial samples.

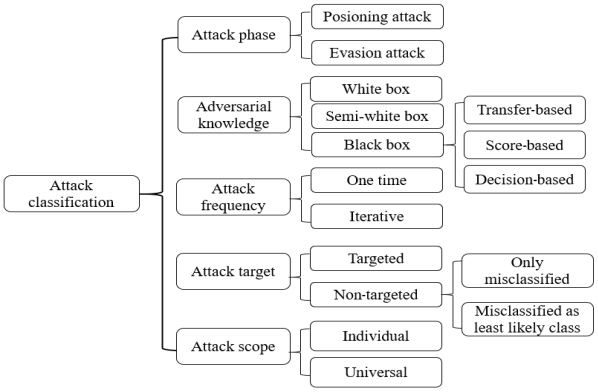


Fig.7. Attack classification in deep learning frameworks.

One-time attacks require less time but result in relatively larger added disturbances. In contrast, iterative attacks, while producing better results, demand extensive computing resources. Attack frequency is a critical consideration, with one-time attacks being efficient and straightforward in generating adversarial samples. However, iterative attacks, although more resource-intensive, can yield superior outcomes. Targeted and non-targeted attacks are differentiated based on the opponent's goal. Targeted attacks aim to change the classifier output to a specific target label, while non-targeted attacks focus on causing the classifier to select any incorrect label. Generally, non-targeted attacks exhibit higher success rates. Non-targeted attacks further subdivide into only misclassification attacks, requiring a model to classify adversarial samples differently from the original class, and least likely attacks, necessitating a model to classify adversarial samples with the least confidence. In terms of attack scope, individual attacks modify a few features, making their adversarial disturbances more imperceptible than those generated by universal attacks, which modify every feature. The timelines of white- and black-box attacks in deep learning, listing algorithms and research schedules. In a white-box attack, attackers can disrupt the learning process by injecting designed samples and adjusting them using gradient methods. References for deep learning security models include proposed attack approaches and security models for wireless sensor networks and cloud computing. White-box attacks are comparatively easier to realize due to attackers' extensive knowledge. In contrast, implementing a black-box attack is challenging due to limitations in model knowledge.

VIII. DEFENSIVE APPROACH IN DEEP LEARNING FRAMEWORKS

Numerous defensive measures have been implemented to address security issues in deep learning, aiming to enhance its application. The intricate relationship between attacks and corresponding defensive approaches is illustrated in Fig. 10. Specifically, measures are strategically aligned with common attack methods. To counter poisoning attacks, which generate adversarial samples, a defense approach involves eliminating outliers with large samples. Evasion attacks can be thwarted by bolstering the robustness of classifiers. Simultaneously, encryption algorithms are employed to safeguard against privacy leakage. Protecting against attacks targeting software vulnerabilities entails writing high-quality code and selecting highly secure third-party libraries. These defensive strategies collectively contribute to fortifying the security of deep learning systems.

Table 3 Historical timeline of white-box attacks in deep learning frameworks.

Timeline	Year	Algorithm	Main contribution
Szegedy et al.[6]	2013	L-BFGS	First proposed the concept of adversarial sample and designed an optimized-based method to generate adversarial samples deliberately.
Goodfellow et al.[9]	2014	FGSM	Designed a method using the gradient of loss function, which can generate adversarial samples quickly.
Papernot et al.[18]	2016	JSMA	Designed a novel method that only needs to modify a few pixels.
Kurakin et al.[19]	2016	iFGSM	Designed the iterative FGSM, which can generate smaller disturbances than the FGSM, and showed that machine learning systems are vulnerable to adversarial examples in physical-world scenarios.
Huang et al.[20]	2017	Attacks on RL	Showed that adversarial attacks are also effective when targeting neural network policies in RL.
Athalye et al.[7]	2018	BPDA	Described the characteristic behaviors of defenses exhibiting effects, discovered three types of obfuscated gradients, and developed attack techniques to overcome them.
Xiao et al.[8]	2019	RMSG	Proposed an adversarial method generating perturbations based on root mean square gradient, which formulates the adversarial perturbation size in the root mean square level and updates gradient direction.
Zhang et al.[21]	2019	Boundary projection	Studied manifold optimization for the classification boundary of an adversarial attack and proposed the boundary projection method to generate adversarial examples that reduce the number of iterations for iterative attacks.

Table 4 Historical timeline of black-box attacks in deep learning frameworks.

Timeline	Year	Algorithm	Main contribution
Nelson et al.[22]	2012	Evading convex-inducing classifiers	First proposed existing black-box attacks that do not use a local model for convex-inducing two-class classifiers.
Ateniese et al.[23]	2013	Hacking smart machines	(1) Proposed that releasing trained classifiers is unsafe; (2) defined a model for a metaclassifier; (3) described several attacks against existing ML classifiers.
Narodytska et al.[24]	2016	Greedy local search	Proposed the Greedy Local Search algorithm to generate adversarial samples by perturbing randomly selected pixels with considerable influence on output probabilities.
Chen et al.[25]	2017	ZOO	(1) Showed that a zero-order oracle (without gradient information) can attack black-box DNNs; (2) proposed several techniques, including attack-space dimension reduction, hierarchical attacks, and importance sampling.
Ye et al.[26]	2018	Hessian-aware zeroth-order optimization	(1) Integrated Hessian information into gradient estimation while keeping the algorithmic form similar to the zeroth-order-based gradient descent method; (2) proposed several novel structured Hessian approximation methods; (3) proposed a descent-checking trick for black-box adversarial attacks.
Li et al.[27]	2019	Attack on cloud-based detectors	Designed four types of methods by incorporating semantic segmentation to achieve a high bypass rate with a very limited number of queries to fool cloud-based detectors.
Saxena[28]	2020	TextDecepter	Proposed a novel approach for formulating natural adversarial examples against Natural Language Processing (NLP) classifiers in the hard-label black-box setting.

IX. CASE STUDY—DEEP LEARNING SECURITY SCENARIO RESEARCH

We conducted an analysis on a deep learning software designed for traffic sign identification to delineate potential attacks and threats faced by deep learning frameworks in practical applications. Through simulations of real-world scenarios, we scrutinized potential challenges in algorithm implementation.

Illustrated in Fig. 11 is a case of a deep learning attack. We specifically chose road signs as our research subject due to their relative simplicity, posing a challenge for concealing disturbances. Additionally, road signs are situated in noisy and dynamic environments, influenced by factors such as observation camera distance and angle, as well as lighting conditions. This case holds significant research value as the accurate recognition of traffic signs is crucial for vehicle safety, demanding resilience against adversarial physical disturbances.

Various forms of attacks may be directed at a deep learning algorithm striving for the accurate identification of road traffic signs. Fig. 11 exemplifies an adversarial example employing algorithms to construct robust perturbations against the deep learning implementation.

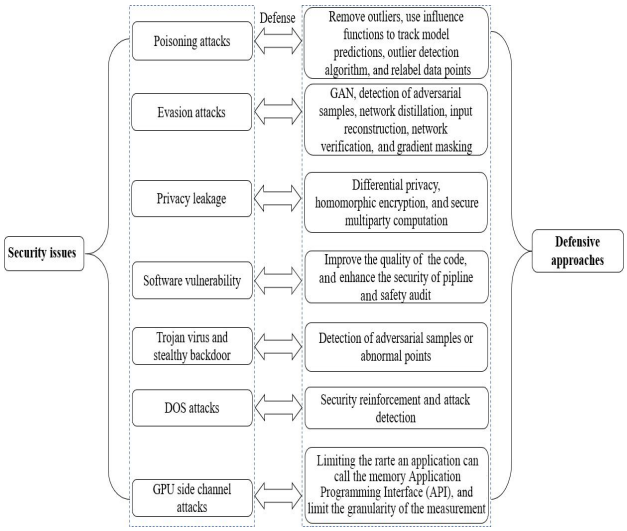


Fig.8. Relationship between attacks and defensive approaches in deep learning

In the study focusing on robust physical-world disturbances [38], researchers employed standard physical science techniques and devised a two-stage experimental design to assess the resilience of the physical-world attack algorithm. The initial stage involved a laboratory test where the viewing camera was adjusted to various distance/angle configurations. The subsequent stage comprised a field test, simulating an autonomous vehicle by driving a car toward an intersection in uncontrolled conditions. Two datasets, the Laboratory for Intelligent & Safe Automobiles (LISA) and the German Traffic Sign Recognition Benchmark (GTSRB), were utilized.

Two classifiers, LISA-CNN and GTSRB-CNN, were trained on the respective datasets, demonstrating high recognition accuracy. Employing object-constrained poster and sticker attacks, the researchers demonstrated the effectiveness of their method in generating robust perturbations for real road signs. The poster attacks achieved a 100% success rate in both stationary and drive-by tests against the LISA-CNN, while the sticker attacks were successful in 80% of stationary testing conditions and 87.5% of extracted video frames against the GTSRB-CNN.

Table 5 showcases several examples of adversarial AI competitions. These competitions play a pivotal role in the realm of machine learning, addressing real-world AI model security and contributing significantly to the advancement of attack and defense methodologies. These competitions not only propel theoretical research but also offer practical insights into AI applications.

Table 5 Adversarial AI competitions.

Competition title	Sponsor	Competition content	Dataset	Champion team
NIPS 2017 Adversarial Attacks and Defenses	Kaggle and NIPS	Targeted attacks, untargeted attacks, and defense against attacks	A new dataset compatible with ImageNet	TSAIL team won all three competitions
ASVspoof 2019	EURECOM, NEC, and so on	Automatic speaker verification spoofing and countermeasures	ASVspoof 2019 dataset contributed by an institution or school, such as Google, USTC, and so on	Tsinghua University
NIPS Adversarial Vision Challenge 2018	NIPS and AWS	To facilitate measurable progress toward robust machine vision models and generally applicable adversarial attacks	NIPS Adversarial Vision Challenge 2018 dataset	Robust Model Track and Targeted Attack Track: Petuum-CMU; Untargeted Attack Track: LIVIA
IJCAI-19	Alibaba Security	To explore the security of AI models; participants can either generate adversarial samples or construct a robust model	Product pictures from the Alibaba e-commerce platform	University of Science and Technology of China (USTC) and so on

X. CONCLUSION AND FUTURE RESEARCH DIRECTION

A. *Conclusion*

Commencing with the fundamental composition and principles of deep learning, this study comprehensively delineates the security challenges inherent in the practical application of deep learning. It encapsulates a compendium of classic attack algorithms relevant to deep learning technologies and developmental processes, substantiating the ubiquity of adversarial samples in the realm of deep learning. The exploration of confrontational algorithms serves a dual purpose—enhancing our understanding of deep learning principles and unraveling the intricacies of its training and prediction processes.

The study meticulously summarizes and analyzes recent algorithmic instances of deep learning attacks while presenting an inventory of defense techniques against these adversarial methodologies. Additionally, it sheds light on specific software vulnerabilities in certain implementations. The susceptibility of deep learning predictions to minor disturbances underscores substantial flaws in the deep learning structure, impeding its further evolution. Despite achieving remarkable accuracy in fixed scenarios like image classification, the intricate dynamics of real-time environments with complex interactions pose challenges, leading to errors and misjudgments. This predicament constitutes a bottleneck in AI technology. Hence, an in-depth exploration of the security issues ingrained in deep learning architecture algorithms assumes profound significance.

B. *Future Research Directions*

1) The development trajectory of attacks and defensive strategies in deep learning spans a protracted journey, evolving from the revelation of deep learning's susceptibility to the emergence and continual enhancement of diverse defensive methodologies. This ongoing process signifies a persistent commitment to refining both offensive and defensive elements.

2) The pervasive presence of adversarial samples serves as a valuable catalyst for enhancing the resilience of deep learning algorithms. The substantial deviations in deep learning prediction results under minor disturbances underscore the need for prolonged refinement. The current state of development is dynamic and characterized by an ongoing, yet incomplete, maturation.

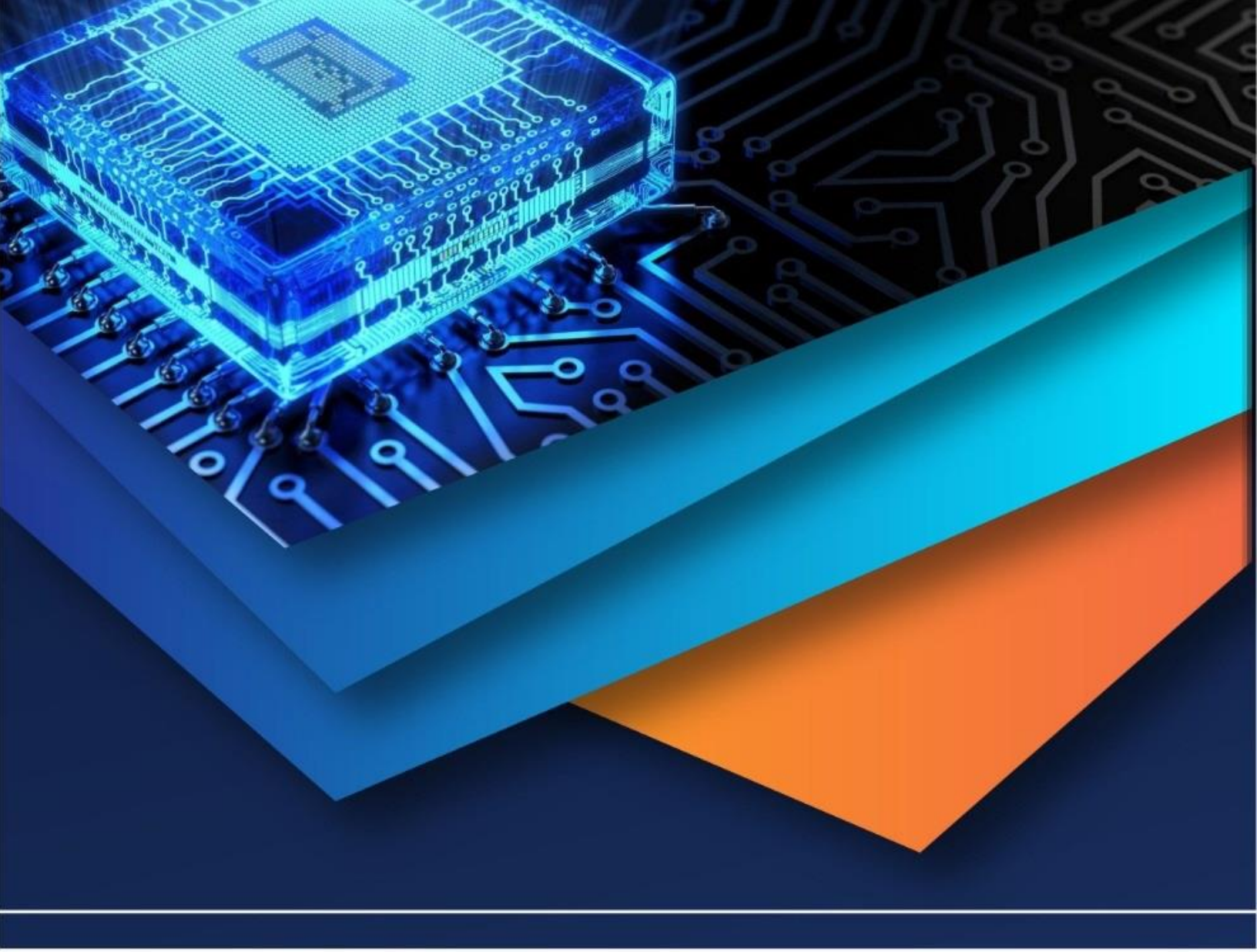
3) Deep learning technologies necessitate significant parallel computing power due to the voluminous training data, prompting a transition from CPUs to GPUs equipped with multiple nodes. However, challenges in underlying software architecture support, such as concerns related to data privacy and security, persist within the industry.

4) Despite commendable performance in experimental settings, neural networks within deep learning systems often fall short during practical application. The real-world environment is intricate, marked by dynamic complexities and concealed unknown variables. A comprehensive consideration of diverse influencing factors and their integration into the training process is imperative for the construction of robust deep learning systems.

REFERENCES

[1] W. W. Jiang and L. Zhang, Geospatial data to images: A deep-learning framework for traffic forecasting, Tsinghua Science and Technology, vol. 24, no. 1, pp. 52–64, 2019.
[2] L. Zhang, C. B. Xu, Y. H. Gao, Y. Han, X. J. Du, and Z. H. Tian, Improved Dota2 lineup recommendation model based on a bidirectional LSTM, Tsinghua Science and Technology, vol. 25, no. 6, pp. 712–720, 2020.
[3] H. M. Huang, J. H. Lin, L. Y. Wu, B. Fang, Z. K. Wen, and F. C. Sun, Machine learning-based multi-modal information perception for soft robotic hands, Tsinghua Science and Technology, vol. 25, no. 2, pp. 255–269, 2020.
[4] X. Y. Yuan, P. He, Q. L. Zhu, and X. L. Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 9, pp. 2805– 2824, 2019.
[5] J. C. Hu, J. F. Chen, L. Zhang, Y. S. Liu, Q. H. Bao, H. Ackah-Arthur, and C. Zhang, A memory-related vulnerability detection approach based on vulnerability features, Tsinghua Science and Technology, vol. 25, no. 5, pp. 604–613, 2020.
[6] C. Szegedy, W. Zaremba, I. Sutskever I, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv: 1312.6199, 2013.
[7] A. Athalye, N. Carlini, and D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, arXiv preprint arXiv: 1802.00420, 2018.
[8] Y. T. Xiao, C. M. Pun, and J. Z. Zhou, Generating adversarial perturbation with root mean square gradient, arXiv preprint arXiv: 1901.03706, 2019.
[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv: 1412.6572, 2014.
[10] J. W. Su, D. V. Vargas, and K. Sakurai, One pixel attack for fooling deep neural networks, IEEE Trans. Evol. Comput., vol. 23, no. 5, pp. 828–841, 2019.
[11] W. He, J. Wei, X. Y. Chen, N. Carlini, and D. Song, Adversarial example defense: Ensembles of weak defenses are not strong, in Proc 11th USENIX Workshop on Offensive Technologies, Vancouver, Canada, 2017.
[12] G. W. Xu, H. W. Li, H. Ren, K. Yang, and R. H. Deng, Data security issues in deep learning: Attacks, countermeasures, and opportunities, IEEE Comm. Mag., vol. 57, no. 11, pp. 116–122, 2019.
[13] M. I. Tariq, N. A. Memon, S. Ahmed, S. Tayyaba, M. T. Mushtaq, N. A. Mian, M. Imran, and M. W. Ashraf, A review of deep learning security and privacy defensive techniques, Mobile Inf. Syst., vol. 2020, p. 6535834, 2020.
[14] H. Bae, J. Jang, D. Jung, H. Jang, H. Ha, and S. Yoon, Security and privacy issues in deep learning, arXiv preprint arXiv: 1807.11655, 2018.
[15] S. L. Qiu, Q. H. Liu, S. J. Zhou, and C. J. Wu, Review of artificial intelligence adversarial attack and defense technologies. Appl. Sci., vol. 9, no. 5, p. 909.
[16] S. M. Moosavi-Dezfooli, A. Fawzi and P. Frossard, DeepFool: A simple and accurate method to fool deep neural networks, presented at 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2574–2582.
[17] Q. X. Xiao, K. Li, D. Y. Zhang, and W. L. Xu, Security risks in deep learning implementations, presented at 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 2018, pp. 123–128.
[18] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, The limitations of deep learning in adversarial settings, presented at 2016 IEEE European Symp. Security and Privacy (EuroS&P), Saarbrucken, Germany, 2016, pp. 372–387.
[19] A. Kurakin, I. Goodfellow, and S. Bengio, Adversarial examples in the physical world, arXiv preprint arXiv: 1607.02533, 2016.
[20] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, Adversarial attacks on neural network policies, arXiv preprint arXiv: 1702.02284, 2017.
[21] H. W. Zhang, Y. Avrithis, T. Furon, and L. Amsaleg, Walking on the edge: Fast, low-distortion adversarial examples, arXiv preprint arXiv: 1912.02153, 2019.
[22] B. Nelson, B. I. P. Rubinstein, L. Huang, A. D. Joseph, S. J. Lee, S. Rao, and J. D. Tygar, Query strategies for evading convex-inducing classifiers, J . Mach. Learn. Res., vol. 13, pp. 1293–1332, 2012.
[23] G. Ateniese, G. Felici, L. V. Mancini, A. Spognardi, A. Villani, and D. Vitali, Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers, arXiv preprint arXiv: 1306.4447, 2013.
[24] N. Narodytska and S. P. Kasiviswanathan, Simple black-box adversarial perturbations for deep networks, arXiv preprint arXiv: 1612.06299, 2016.
[25] P. Y. Chen, H. Zhang, Y. Sharma, J. F. Yi, and C. J. Hsieh, ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in Proc. 10th ACM Workshop on Artificial Intelligence and Security, Dallas, TX, USA, 2017, pp. 15–26.
[26] H. S. Ye, Z. C. Huang, C. Fang, C. J. Li, and T. Zhang, Hessian-aware zeroth-order optimization for black-box adversarial attack, arXiv preprint arXiv: 1812.11377, 2018.
[27] X. R. Li, S. L. Ji, M. Han, J. T. Ji, Z. Y. Ren, Y. S. Liu, and C. M. Wu, Adversarial examples versus cloud-based detectors: A black-box empirical study, arXiv preprint arXiv: 1901.01223, 2019.
[28] S. Saxena, TextDecepter: Hard label black box attack on text classifiers, arXiv preprint arXiv: 2008.06860, 2020.
[29] A. Zimba, H. S. Chen, and Z. S. Wang, Bayesian network based weighted APT attack paths modeling in cloud computing, Future Generation Comput. Syst., vol. 96, pp. 525–537, 2019. 448–456.
[30] H. S. Chen, C. X. Meng, Z. G. Shan, Z. C. Fu, and B. K. Bhargava, A novel low-rate denial of service attack detection approach in zigbee wireless sensor network by combining Hilbert-Huang transformation and trust evaluation, IEEE Access, vol. 7, pp. 32 853–32 866, 2019.
[31] J. Steinhardt, P. W. Koh, and P. Liang, Certified defenses for data poisoning attacks, presented at 31st Conf. Neural Information Proc. Systems, Long Beach, CA, USA, 2017, pp. 3517–3529.
[32] P. W. Koh and P. Liang, Understanding black-box predictions via influence functions, arXiv preprint arXiv: 1703.04730, 2017.
[33] A. Paudice, L. Mu˜noz-Gonz´alez, A. Gyorgy, and E. C. Lupu, Detection of adversarial training examples in poisoning attacks through anomaly detection, arXiv preprint arXiv: 1802.03041, 2018.

[34] A. Paudice, L. Mu˜noz-Gonz´alez, and E. C. Lupu, Label sanitization against label flipping poisoning attacks, in Joint European Conf. Machine Learning and Knowledge Discovery in Databases, A. Paudice and L. Mu˜noz Gonz´alez, eds. Cham, Germany: Springer, 2018, pp. 5–15
[35] N. Carlini and D. Wagner, Towards evaluating the robustness of neural networks, presented at 2017 IEEE Symp. Security and Privacy (SP), San Jose, CA, USA, 2017, pp. 39–57.
[36] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy, presented at Proc. 33rd Int. Conf. Machine Learning, New York, NY, USA, 2016, pp. 201–210.
[37] S. Lee, H. Kim, J. Park, J. Jang, C. S. Jeong, and S. Yoon, TensorLightning: A traffic-efficient distributed deep learning on commodity spark clusters, IEEE Access, vol. 6, pp. 27 671–27 680, 2018.
[38] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. W. Xiao, A. Prakash, T. Kohno, and D. Song, Robust physical-world attacks on deep learning visual classification, presented at 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1625–1634.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY