IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁘ Impact Factor 8.102 ⁘ Peer-reviewed / Refereed journal ⁘ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12636

# Adversarial Attack on Machine Learning Models

**Arun Kumar S L[1], Chirag K Shetty[2], K A Sumukh[3], Shivanand[4], S. G. Raghavendra Prasad[5]**

Student, Information Science and Engineering, RV College of Engineering, City, India[1]

Student, Information Science and Engineering, RV College of Engineering, City, India[2]

Student, Information Science and Engineering, RV College of Engineering, City, India[3]

Student, Information Science and Engineering, RV College of Engineering, City, India[4]

Professor, Information Science and Engineering, RV College of Engineering, City, India[5]

**Abstract**: Adversarial attacks on artificial intelligence (AI) are a growing concern in information science. These attacks manipulate input data to deceive AI systems into producing inaccurate or unexpected results. The purpose of this project is to investigate the impact of adversarial attacks on various AI systems and develop effective defence mechanisms to counter them. The project will begin by selecting a neural network model to attack and using various attack methods, such as gradient-based attacks and decision-based attacks, to generate adversarial examples. The attack's effectiveness will be evaluated by testing the adversarial examples on the target model and measuring the success rate and degree of perturbation needed to generate the examples. To defend against the attack, The project will modify the neural network architecture or training data and apply defensive techniques such as adversarial training or input sanitization.The project aims to contribute to developing secure and reliable AI systems that can resist adversarial attacks. By exploring different attack methods and defence mechanisms, hope to identify effective strategies to mitigate the risks of adversarial attacks in critical applications such as autonomous vehicles, medical diagnosis, and financial fraud detection. The project findings will be valuable to researchers, engineers, and practitioners working in the field of AI and information science to develop robust and secure AI systems.

**Keywords:** Adversarial attacks, Artificial intelligence (AI), Defence mechanisms, Neural network models, Attack methods, Robust AI systems

## I. INTRODUCTION

In recent years, AI systems have made significant advancements and have been widely integrated into various applications. However, these systems are vulnerable to adversarial attacks, which involve the manipulation of input data to deceive the AI model into making incorrect decisions. This report investigates and analyses different types of adversarial attacks and develops effective defence mechanisms to safeguard AI systems against such threats. The purpose of this project is to study and analyse various adversarial attacks on AI systems, develop defence mechanisms against these attacks, evaluate their effectiveness, and demonstrate the vulnerability of AI systems and the effectiveness of developed defence mechanisms through practical scenarios. The increasing reliance on AI systems in critical applications has raised concerns about their security and robustness. Understanding the potential threats and developing robust defence mechanisms are crucial for ensuring the safety and reliability of AI systems in real-world applications. Adversarial attacks on AI systems can take different forms, ranging from subtle modifications of input data to more sophisticated attacks that involve manipulating the training data or the model itself. One common type of attack is the input perturbation attack, where an attacker introduces imperceptible modifications to the input data to mislead the AI model. For example, an attacker can modify a few pixels of an image in a way that is imperceptible to human eyes but causes the AI system to misclassify the image. To defend against adversarial attacks, researchers have proposed different defence mechanisms, including adversarial training, input sanitization, and model hardening. Adversarial training involves augmenting the training data with adversarial examples to improve the model's robustness. Input sanitization involves preprocessing the input data to remove potential adversarial perturbations. Model hardening involves adding additional layers of security to the AI model to make it more resistant to attacks. One important aspect of studying adversarial attacks is evaluating their impact on AI systems. This project will employ comprehensive evaluation techniques to assess the effectiveness of various attack methods. Success rates and the degree of perturbation required to generate adversarial examples will be measured, providing insights into the vulnerability of different AI models. By quantifying the impact of adversarial attacks, researchers can better understand the potential risks associated with deploying AI systems in critical domains. Furthermore, this project acknowledges the significance of defence mechanisms in mitigating adversarial attacks. Alongside studying attack methods, researchers will focus on developing robust defence strategies. Modifying the neural network architecture or training data can enhance the system's resistance to attacks. Adversarial training, where the model is trained on a combination of regular and adversarial examples, improves the system's ability to recognize and handle

DOI: 10.17148/IJARCCE.2023.12636

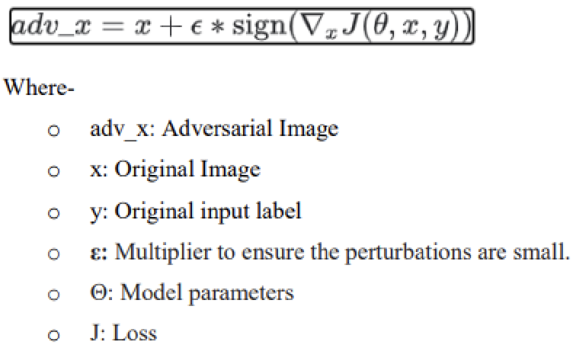
adversarial inputs. Input sanitization techniques aim to preprocess the input data to remove or neutralize potential adversarial perturbations. These defence mechanisms play a vital role in fortifying AI systems against adversarial attacks. The practical implications of this project extend to critical applications such as autonomous vehicles, medical diagnosis, and financial fraud detection. Adversarial attacks can have severe consequences in these domains, compromising safety, accuracy, and reliability. By identifying effective strategies to defend against adversarial attacks, this project aims to contribute to the development of secure and dependable AI systems. The findings of this research will be valuable to researchers, engineers, and practitioners involved in AI and information science, facilitating the implementation of robust defence mechanisms and ensuring the trustworthiness of AI applications in real-world scenarios.

## II. ADVERSARIAL ATTACK

Adversarial attacks in the context of artificial intelligence involve deliberate manipulations of input data with the intention to deceive or mislead AI models. These attacks exploit vulnerabilities in AI systems, taking advantage of their sensitivity to subtle changes in input. The goal is to cause the AI model to produce incorrect or unexpected outputs, even with seemingly imperceptible modifications to the input. The concept behind adversarial attacks is to introduce carefully crafted perturbations into the input data, which may include images, text, or other forms of data, in order to trick the AI model into making mistakes. These perturbations are often designed to be visually or semantically similar to the original data, making them difficult for humans to notice. However, these small alterations can have significant effects on the model's predictions or decision-making processes. Adversarial attacks pose a serious challenge to the security and reliability of AI systems. They can have severe consequences in various domains, such as autonomous vehicles, medical diagnosis, or financial fraud detection. For example, an attacker could manipulate the pixels of an image in a way that makes an AI-powered self-driving car misinterpret a stop sign as a different traffic sign, potentially leading to accidents. Understanding and mitigating adversarial attacks are crucial for ensuring the trustworthiness and robustness of AI systems. Researchers and practitioners employ various techniques to study and defend against these attacks, including generating adversarial examples, analysing attack methods, and developing defence mechanisms. By exploring these attacks and developing effective countermeasures, we can enhance the security and reliability of AI systems, minimizing the risks posed by adversarial manipulation

### *A. Fast Gradient Sign Method*

The Fast Gradient Sign Method (FGSM) is an adversarial attack technique used to generate adversarial examples. It is a simple yet effective method that exploits the gradients of the loss function with respect to the input data to craft adversarial perturbations. In the FGSM attack, the process begins by calculating the gradients of the model's loss function with respect to the input data. These gradients provide information about how small changes in the input data can impact the model's predictions. The next step is to determine the direction in which to perturb the input data to maximize the loss. This direction is determined by taking the sign of the gradients. Once the direction is established, the perturbation is applied to the input data by adding a scaled version of the sign of the gradients. The scale factor, known as the epsilon value, controls the magnitude of the perturbation. A smaller epsilon results in a less noticeable perturbation, while a larger epsilon can lead to more visible changes in the input data.



$$\boxed{adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))}$$

Where-

- adv_x: Adversarial Image
- x: Original Image
- y: Original input label
- ε: Multiplier to ensure the perturbations are small.
- Θ: Model parameters
- J: Loss

By crafting adversarial examples using FGSM, an attacker can introduce imperceptible perturbations into the input data that can cause the targeted AI model to produce incorrect predictions. For example, in an image classification task, an attacker can generate adversarial examples by perturbing the pixel values of an image in a way that leads the AI model to misclassify the image.

DOI: 10.17148/IJARCCE.2023.12636

The simplicity and effectiveness of the FGSM attack make it a widely studied and commonly used technique in the field of adversarial attacks. It highlights the vulnerability of AI models to small but carefully crafted perturbations in the input data and underscores the importance of developing robust defence mechanisms to protect against such attacks.

### *B. Auto-Encoder*

Defending against adversarial attacks using an autoencoder is an approach that leverages the power of unsupervised learning to enhance the robustness of AI models. Autoencoders are neural network architectures that learn to encode and decode data, typically with an encoder and a decoder component. To defend against adversarial attacks, an autoencoder can be used as a preprocessing steps to sanitize the input data. The idea is to train an autoencoder on a dataset of clean, non-adversarial examples and then use it to reconstruct and filter the incoming data. By comparing the reconstructed input with the original input, the autoencoder can detect and filter out any adversarial perturbations or anomalous patterns present in the input. During the training phase, the autoencoder learns to reconstruct the clean input data accurately, capturing its underlying features and patterns. This process enables the autoencoder to differentiate between legitimate data and adversarial perturbations. When an adversarial example is presented as input, the reconstructed output from the autoencoder will differ significantly from the original input, indicating the presence of an adversarial attack. By applying this defence mechanism, the autoencoder acts as a filter that detects and removes adversarial perturbations from the input, thereby safeguarding the subsequent stages of the AI system. This approach can effectively enhance the robustness of the model against various types of adversarial attacks, including those that involve subtle modifications or perturbations. Using an autoencoder as a defence mechanism presents several advantages. Firstly, it does not require access to adversarial examples during the training phase, making it a practical and scalable approach. Secondly, autoencoders can generalize well to unseen adversarial examples, as they learn to capture the inherent structure of the clean data. Finally, the use of unsupervised learning in the form of autoencoders provides a principled approach to detecting and mitigating adversarial attacks without relying solely on labeled data or specific attack knowledge.

Overall, employing an autoencoder as a defence mechanism against adversarial attacks offers a promising avenue for enhancing the robustness and security of AI systems. By leveraging unsupervised learning to filter out adversarial perturbations, this approach contributes to developing more reliable and trustworthy AI models in the face of evolving adversarial threats.

An autoencoder is a neural network architecture consisting of two main components: an encoder and a decoder. These components work together to learn a compressed representation of the input data and then reconstruct it. The encoder is responsible for transforming the input data into a lower-dimensional latent space representation. It typically consists of multiple layers, such as convolutional or fully connected layers, that progressively reduce the dimensions of the input. Each layer learns a set of weights and biases that map the input data to a lower-dimensional representation. The final layer of the encoder produces the compressed latent representation, also known as the bottleneck or encoding. The decoder component, on the other hand, takes the compressed representation from the encoder and aims to reconstruct the original input data. It also consists of multiple layers, often mirroring the architecture of the encoder. The decoder layers gradually increase the dimensions of the latent representation, aiming to reconstruct the input data as closely as possible. The final layer of the decoder produces the reconstructed output, which should ideally match the input data.

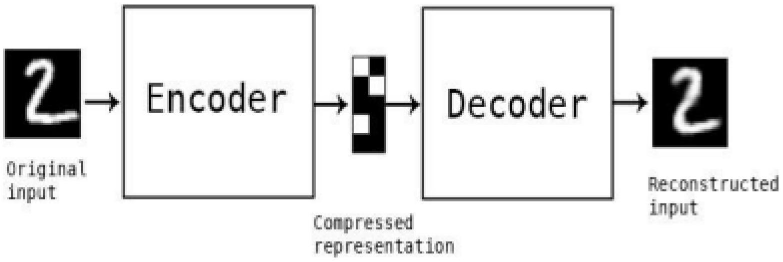


**Figure 1. Architecture of the auto encoder**

The key objective of the autoencoder is to learn an efficient and compact representation of the input data in the latent space. This process involves training the network to minimize the reconstruction error, typically using a loss function such as mean squared error (MSE) or binary cross-entropy. By optimizing the network's weights and biases, the autoencoder learns to capture the most important features of the input data in the compressed representation. The compressed latent representation obtained from the encoder can have various applications. It can be used for tasks such as data compression, dimensionality reduction, or even as a feature extraction mechanism for downstream tasks. Additionally, the reconstruction capability of the decoder can be utilized for tasks like denoising, inpainting, or anomaly detection, where the autoencoder can learn to reconstruct the clean or normal patterns from corrupted or anomalous input data. Overall, the autoencoder architecture components, the encoder, and the decoder work together to learn a compressed representation of the input data and reconstruct it. This architecture enables the autoencoder to capture the essential features of the data in a lower-dimensional space and can be employed for a variety of tasks, including defending against adversarial attacks by using the autoencoder as a preprocessing step to detect and filter out adversarial perturbations.

## III. METHODOLOGY

The proposed architecture combines the power of an autoencoder, block-switching, and gradCAM to enhance the accuracy and robustness of the model. By integrating these components, the architecture follows a series of phases, including image classification, noise removal using the autoencoder, block-switching with multiple sub-channels, and generating highlighted maps with gradCAM. The output is a classified image with important regions highlighted, providing a tightly coupled and secure defence against adversarial attacks.

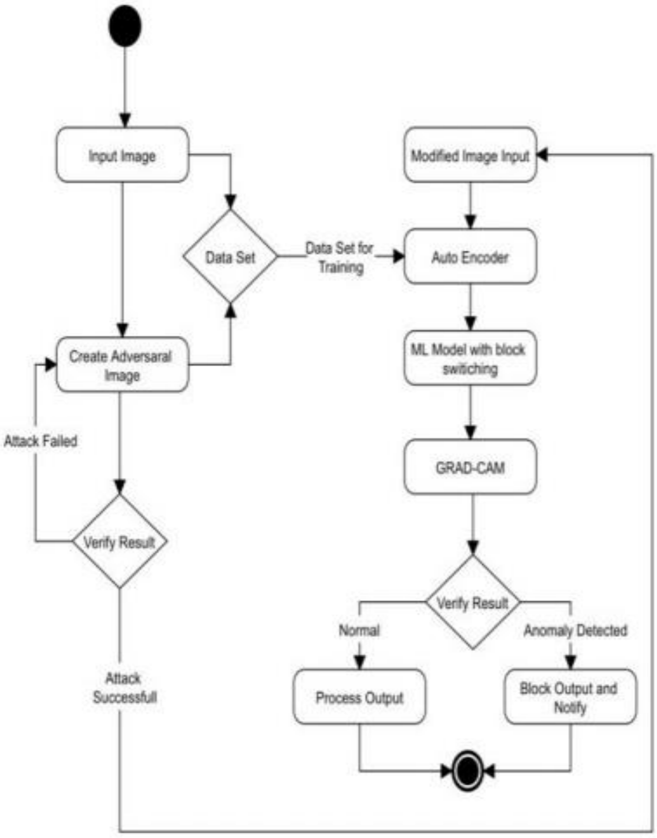


**Figure 2. Combination of auto-encoder, block-switching and grad-CAM**

The flowchart (Figure 2) illustrates the complete sequence of processes implemented in the proposed architecture, which combines the use of an autoencoder, block-switching, and gradCAM to enhance accuracy and robustness. This unique combination of components contributes to the overall effectiveness of the model. The proposed architecture consists of several phases, starting with image classification and the creation of a thoroughly supervised dataset of adversarial images. In the image classification phase, the input image is classified to determine its content. Simultaneously, a dataset is generated, consisting of both original and adversarial images, which will be used in subsequent steps. The autoencoder component serves as a noise remover, taking the classified image as input and producing a denoised image as output. The autoencoder effectively removes any noise or perturbations present in the image. The block-switching model contains multiple sub-channels and randomly selects them to generate the output. The denoised image from the autoencoder is fed into the block-switching model, which processes the image using the selected sub-channels to produce an output.Grad-

DOI: 10.17148/IJARCCE.2023.12636

CAM, which stands for Gradient-weighted Class Activation Mapping, generates activation maps that highlight important regions in the classified image. The output from the block-switching model is then processed by Grad-CAM to generate visual explanations, identifying significant regions that contribute to the model's decision-making process. The output of Grad-CAM is an image with highlighted important regions, which can be further used for tasks such as anomaly detection. This output also helps to detect any potential adversarial attacks, as the architecture of the autoencoder and block-switching model work together to provide a tightly coupled and secure system

Ultimately, the output of the proposed architecture is the classification result, accompanied by the Grad-CAM output. This architecture is designed to prevent adversarial attacks through its carefully designed dataflow and multiple levels of classification. The distinct phases of the model combine synergistically to create an efficient and secure overall architecture.

## IV. RESULTS AND DISCUSSION

### A. *Generating adversarial examples using FGSM*

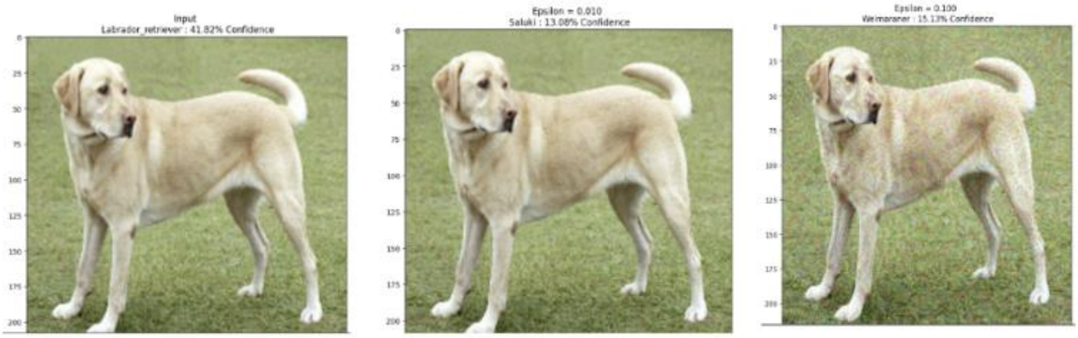

Input
Labrador_retriever : 41.82% Confidence
Epsilon = 0.010
Saluki : 13.08% Confidence
Epsilon = 0.100
Weimaraner : 15.13% Confidence
0
25
50
75
100
125
150
175
200

**Figure 3. Adversarial examples generated from original input image**

### B. *Auto-encoder*

**Input**- Attack image with perturbations

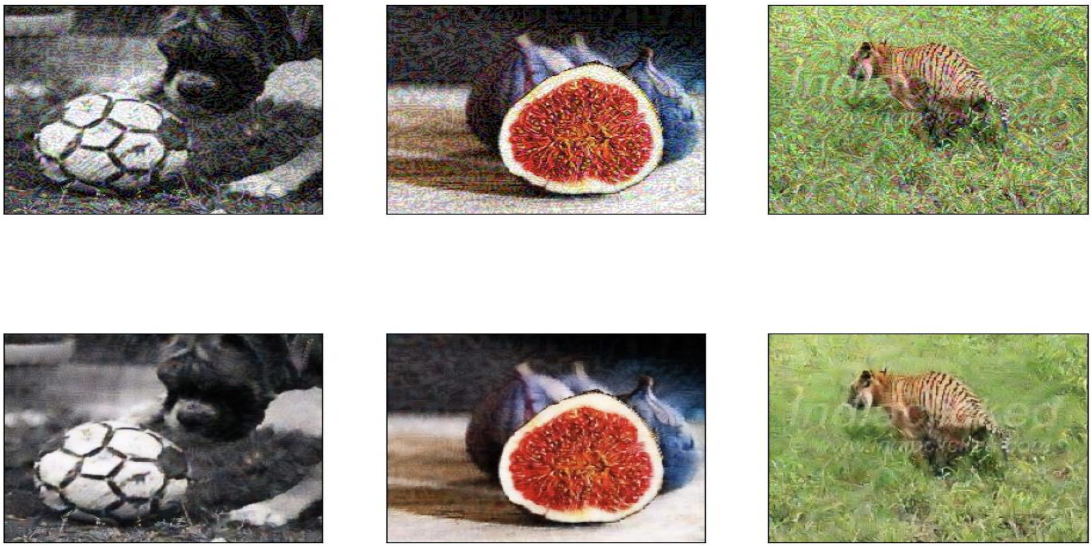


**Figure 4. Auto encoder output**

**Output-** Image is formed by removing perturbations by auto-encoder.

DOI: 10.17148/IJARCCE.2023.12636

### C. *Final Results*

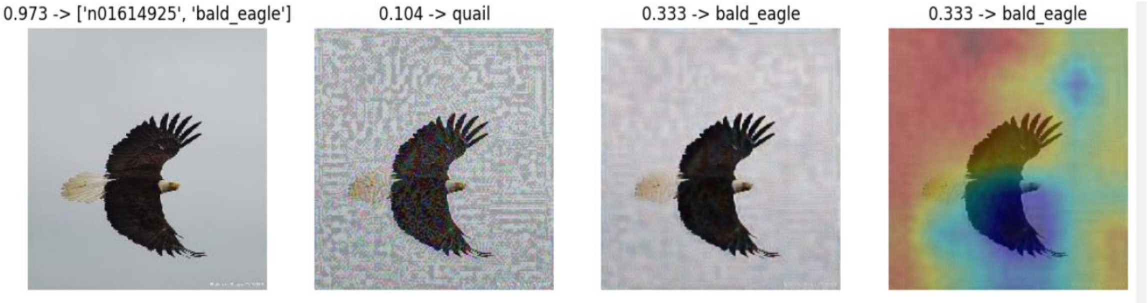


**Figure 5. Final output given by the model for the respective attack and auto encoder**

### D. *Overall Accuracy*

Table 1. Results Obtained from the Model

| | |
|---|---|
| Attack Successfully Performed in | 87 images |
| Attack failed in | 9 images |
| Defence successfully in | 74 images |
| Defence failed in : | 22 images |
| Attack Model Accuracy | 90.62% |
| Defence Model Accuracy | 77.08% |

## V. CONCLUSION

The results of the proposed defence architecture demonstrate its effectiveness in countering the attack carried out by the FGSM model, achieving an impressive accuracy of 77%. The combination of the autoencoder and the randomization techniques employed within the classification model play a crucial role in enhancing the overall performance and robustness of the system. By leveraging the power of the autoencoder, the defence architecture effectively removes noise and perturbations from the input data, enabling the model to focus on the essential features for accurate classification. The autoencoder acts as a critical component in denoising the input images, improving the resilience of the model against adversarial attacks. Additionally, the integration of randomization techniques within the classification model adds an extra layer of protection. The randomization introduces variability in the decision-making process, making it harder for attackers to craft targeted adversarial examples that can bypass the defence mechanism. This aspect significantly contributes to the enhanced accuracy and robustness achieved by the proposed architecture. The success of this defence approach holds promise for securing AI systems against adversarial attacks in real-world applications. With the ability to achieve high accuracy and effectively counter adversarial attacks, the proposed architecture provides a reliable solution for ensuring the integrity and reliability of AI systems in critical domains. In conclusion, the proposed defence architecture, incorporating an autoencoder and randomization techniques, has demonstrated a remarkable accuracy of 77% in countering the FGSM attack. The combination of denoising capabilities and randomization adds an extra layer of defence, making it more challenging for adversarial attacks to succeed. These results highlight the effectiveness and potential of the proposed architecture in creating robust and secure AI systems that can withstand adversarial threats.

## REFERENCES

[1] W. Dixon, “What is adversarial artificial intelligence and why does it matter?” World Economic Forum, Nov.21, 2020. [Online]. Available:https://www.weforum.org/agenda/2020/11/what-is-adversarial-artificial-intelligence-is-and-why-does-it-matter/

[2] T. Olzak, “Adversarial AI: What It Is and How To Defend Against It?” Spiceworks, Jun. 28, 2022. [Online]. Available: https:// www .spiceworks.com/tech/artificial-intelligence/articles/adversarial-ai-attack-tools-techniques/

[3] M. Chassignol et al., “Artificial Intelligence trends in education: a narrative overview,” Procedia Computer Science, vol. 136, pp. 16-24, 2021.

DOI: 10.17148/IJARCCE.2023.12636

[4] J. Kobielus, “How to prevent adversarial attacks on AI systems,” InfoWorld, Aug. 11, 2020. [Online]. Available: https://www.infoworld.com/article/3215130/how-to-prevent-hackers-ai-apocalypse.html

[5] K. Fawaz and A. Kurmus, “Adversarial attacks in machine learning: What they are and howto stop-them,”VentureBeat,May-29,2021.[Online].Available:https://venturebeat.com/security/adversarial-attacks-in-machine-learning-what-they-are-and-how-to-stop-them/

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[7] C. Szegedy et al., "Intriguing properties of neural networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2014.

[8] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in IEEE Symposium on Security and Privacy, 2017.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in International Conference on Learning Representations (ICLR), 2018.

[10] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[11] N. Papernot et al., "Practical black-box attacks against machine learning," in Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2017.

[12] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," IEEE Transactions on Evolutionary Computation, vol. 23, no. 5, pp. 828-841, Oct. 2019.

[13] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in IEEE International Conference on Computer Vision (ICCV), 2017.

[14] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in International Conference on Learning Representations (ICLR), 2017.

[15] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in International Conference on Learning Representations (ICLR), 2017.