Comparison of Adversarial Robustness of Convolutional Neural Networks for Handwritten Digit Recognition

Zhen Ren

College of Information, North China University of Technology, Beijing, 100144, China
21152090102@mail.ncut.edu.cn

Abstract. Machine learning has found widespread application in contemporary society, yet it remains vulnerable to the corrosive effects of adversarial samples. These refer to input data that has been deliberately modified in a certain way to mislead machine learning models. While these modifications may be undetectable to human observers, they are sufficient to trigger erroneous outputs from machine learning models, thereby compromising their robustness, and exposing their weaknesses. The purpose of this paper is to examine the vulnerability of machine learning models to adversarial samples. The Fast Gradient Sign Method (FGSM) is used to create adversarial samples from the Modified National Institute of Standards and Technology (MNIST) dataset, which are then used to attack the LeNet and a basic convolutional neural network (CNN) model. The findings reveal that the LeNet model exhibits a higher degree of sensitivity compared to the simple CNN model. As time progresses and models continue to innovate, they are becoming less prone to interference from adversarial samples. This study could offer valuable insights for future endeavors aimed at designing more secure and resilient machine learning models.

Keywords: Convolutional Neural Network, Fast Gradient Sign Method, Adversarial Samples.

1 Introduction

The initial idea of "machine learning" is able to be traced back to Allan et al. and their research paper "Computing Machinery and Intelligence". They raised a foundational question for machine learning: "Can machines think?" [1]. As time has progressed, machine learning has continuously evolved. In the current era, A type of artificial intelligence based on data and algorithms is called machine learning. By continually analyzing data, training samples, and learning models, machine learning discovers patterns within data and models, enabling machines to perform tasks such as prediction, classification, clustering, and discrimination autonomously. However, machine learning is not flawless as envisioned. During the training of samples, minor variations in samples can significantly reduce the accuracy of machine learning. As a result of this phenomenon, adversarial examples emerged [2-4].

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_54

Data inputs that have been subtly modified to deceive machine learning models are called adversarial examples. The presence of adversarial examples in machine learning tasks has the following impacts:

1. Revealing model vulnerability: Adversarial attacks reveal the vulnerability and unreliability of machine learning models. When a model produces incorrect classification results on adversarial examples, it indicates that the model cannot effectively generalize to new data with minor perturbations, thus exposing the flaws and vulnerable points of the model.

2. Challenging model robustness: Adversarial examples challenge the robustness of machine learning models, i.e., the model's tolerance to perturbations and noise in input data. The existence of adversarial examples prompts researchers to develop more robust models and defense mechanisms to enhance the model's ability to withstand attacks.

3. Impacting model interpretability: Adversarial attacks also pose a challenge to the interpretability of models. The model's internal mechanisms and decision-making processes can be involved in the generation process of adversarial examples, researchers need to analyze the effects of adversarial attacks on models to improve the interpretability and understandability of models.

Based on these reasons, people have begun to generate adversarial examples through datasets and various methods to train models. By continuously training and improving models, the robustness of machine learning is ultimately enhanced.

The focus of research in recent years has been on methods to generate adversarial examples, evaluate model robustness, and study how adversarial examples can influence models. Szegedy and colleagues first discovered that deep neural networks are sensitive to adversarial examples, which involve intentionally adding imperceptible disturbances to the input samples. Subsequently, the Fast Gradient Sign Method (FGSM) was introduced by Goodfellow and colleagues, which is an efficient technique for generating adversarial examples [5]. Later, researchers proposed more sophisticated attack methods, such as DeepFool and Projected Gradient Descent (PGD) [6,7]. Moreover, various defense mechanisms are also proposed, including adversarial training, input transformation, and model regularization, etc. Nowadays, adversarial attack methods have evolved from the initial FGSM to more advanced methods such as PGD, Carlini and Wagner (C&W) attacks [8]. Researchers continue to propose new adversarial attack techniques, which increase model prediction errors while keeping input perturbations imperceptible to humans.

Despite numerous studies, there is relatively little research on the comparison of the sensitivity of different machine-learning models to adversarial examples. This paper intends to study the sensitivity of convolutional neural networks to adversarial examples by exploring the sensitivity of convolutional neural network model structures to adversarial examples and summarizing and analyzing possible approaches to improve model robustness.

2 Method

First, import relevant libraries and define the selected and constructed models. Then, load the dataset, instruct the model, test the model, define the attack method, use this attack method to create examples that are adversarial, and finally visualize the results.

This study selects two models: LeNet and simple convolutional neural networks (CNN) [9,10]. These two models are relatively mature and widely used in convolutional neural networks, representing typical choices. Two convolutional layers and three fully connected layers are present in LeNet, one of the early convolutional neural networks. Handwritten digit recognition was one of the first uses of convolutional neural networks. The simplicity of its structure and fast training speed makes it an ideal choice for evaluating basic performance and robustness. Compared to LeNet, simple CNN has a deeper network structure, allowing it to have better feature extraction capabilities and classification performance. Additionally, simpleCNN is utilized in a variety of applications, including image classification and segmentation, object detection, and facial recognition. Therefore, selecting these two models can help this study efficiently and accurately analyze the sensitivity of different models to adversarial examples.

The Fast Gradient Sign Method (FGSM) is employed in this study to generate adversarial examples. FGSM is an efficient method for generating adversarial examples by adding small perturbations to input data to deceive neural network models:

$$x_{adv} = x + \varepsilon * sign(\nabla x\, J(x, y)) \tag{1}$$

In this method, the direction of perturbation aligns with the direction of the gradient, with the magnitude controlled by ε. FGSM possesses efficiency and intuitiveness, enabling the generation of adversarial examples with a single gradient computation, devoid of complex optimization procedures. It facilitates the rapid generation of adversarial perturbations for large datasets and intuitively introduces perturbations along the most sensitive direction of the model, thereby inducing erroneous prediction outcomes.

To accommodate both models and ensure the dataset's broad applicability and randomness, the Modified National Institute of Standards and Technology (MNIST) dataset was selected. MNIST is a classical dataset of handwritten digit images extensively utilized in machine learning and computer vision research. Comprising 70,000 grayscale images of size 28x28 pixels, the dataset encompasses handwritten digit images ranging from 0 to 9. The performance of the two models in this study can be evaluated using it due to its widespread adoption and standardization. Being relatively compact, user-friendly, and comprehensible, MNIST possesses sufficient complexity to assess the performance of image classification algorithms in this study.

3 Result and Discussion

According to the investigation, it was found that irrespective of the degree of perturbation, the robustness of the simple CNN model surpasses that of the LeNet model, as depicted in Fig.1 and Fig.2. As ε increases, the accuracy of both models

begins to decline, indicating the influence of adversarial examples on the predictive capabilities of the models. Specific results are demonstrated in Table 1. Particularly noteworthy is the significant decrease in the hit rate of the LeNet model when ε=0.1, whereas the performance of the simple CNN model exhibits normal fluctuations. When ε=0.3, it was observed that the hit rates of both models dropped to 0.06 and 0.1, respectively. Compared to ε=0, these hit rates decreased by 94% and 90%, respectively. These data corroborate the following assertions: The sensitivity of the LeNet model is higher than that of the simple CNN model, and the models are highly sensitive to samples with larger perturbations, making it challenging to correctly classify adversarial examples.

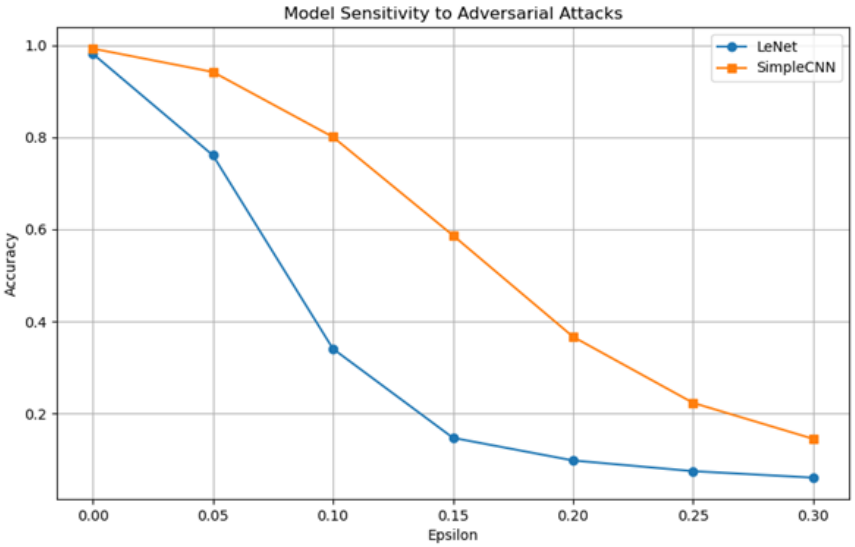

Fig. 1. Model susceptibility to adversarial attacks.

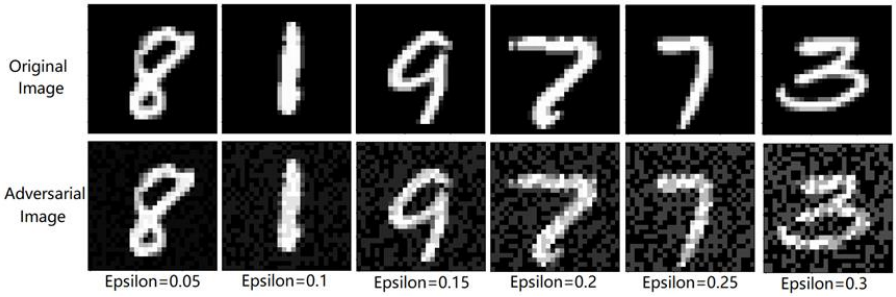

Fig. 2. Representative adversarial examples under different perturbations.

Table 1. Model sensitivity to adversarial attacks

Epsilon	LeNet Accuracy	SimpleCNN Accuracy
0	0.9811	0.9924

0.05	0.7608	0.9417
0.10	0.3408	0.8006
0.15	0.1475	0.5869
0.20	0.098	0.3669
0.25	0.075	0.2234
0.30	0.061	0.1454

The reasons underlying the experimental results can be attributed to the following factors: Deeper Network Structure and Complex Feature Extraction Capability of SimpleCNN: The SimpleCNN model possesses a deeper network structure and more sophisticated feature extraction capability. This characteristic potentially enables SimpleCNN to better extract useful information from perturbations, thereby maintaining higher accuracy when confronted with adversarial examples. In contrast, the relatively simpler architecture of LeNet may render it more susceptible to the influence of adversarial perturbations; Utilization of Rectified Linear Unit (ReLU) Activation Function and Max Pooling: The SimpleCNN model employs the ReLU activation function and max pooling, contributing to enhanced model nonlinearity and learning capability. The ReLU activation function facilitates sparse activation and rapid convergence, while max pooling aids in extracting salient features and reducing sensitivity to position. These characteristics likely contribute to the superior performance of the SimpleCNN model when confronted with samples bearing subtle perturbations; Larger Receptive Field: SimpleCNN maintains a larger receptive field by employing larger convolutional kernels and padding. This implies that the model can observe more extensive contextual information within the input data. A larger receptive field assists the model in better understanding the overall structure and relationships within the image, thereby improving its ability to recognize adversarial perturbations; Influence of Data Preprocessing and Optimizer: The structural differences between models may lead to varying sensitivities to data preprocessing methods and adaptability to optimizers. SimpleCNN may be better suited to handle the MNIST dataset, preprocessed using the provided methods, and may effectively leverage the Adam optimizer for efficient learning.

These factors collectively contribute to the observed differences in model performance when subjected to adversarial examples.

4 Conclusion

In summary, the idea of analyzing the sensitivity of different convolutional neural network models to adversarial examples by varying perturbations is feasible. This paper aimed to assess and compare how two distinct models respond to adversarial examples and visualize the understanding of model sensitivity when facing potential adversarial attacks. To perturb the LeNet and SimpleCNN models, adversarial examples were generated using the FGSM method and MNIST dataset. These adversarial examples introduced varying degrees of perturbation to both models, with analysis revealing the heightened sensitivity of the LeNet model. This phenomenon can be attributed to the superior structural characteristics of the SimpleCNN model. It also implies that with the advancement of technology, more superior models are being designed. These new

models possess better image recognition and various other capabilities. However, this study still has limitations as it only employed two models and one adversarial example generation method, thus failing to represent the full complexity of machine learning models and adversarial attacks. Additionally, the MNIST dataset has its limitations and cannot fully encompass real-world characteristics. In future work, employing a wider range of models and adversarial example generation methods can optimize this study.

References

1. Turing, A.M.,: Computing machinery and intelligence. Springer Netherlands (1950).
2. Akhtar, N., and Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, **6**, 14410-14430 (2018).
3. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., and Kohane, I. S.: Adversarial attacks on medical machine learning. Science, **363**(6433), 1287-1289 (2019).
4. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D.: A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology, **6**(1), 25-45 (2021).
5. Goodfellow, I. J., Shlens, J., and Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
6. Moosavi-Dezfooli, S. M., Fawzi, A., and Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574-2582. IEEE, Las Vegas (2016).
7. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017).
8. Carlini, N., and Wagner, D.: Towards evaluating the robustness of neural networks. In 2017 IEEE symposium on security and privacy. pp. 39-57. IEEE, San Jose. (2017).
9. Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, **33**(12), 6999-7019 (2021).
10. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et, al.: Recent advances in convolutional neural networks. Pattern recognition, **77**, 354-377 (2018).