International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Adversarial Attack in Traffic Sign Identification Using Deep Learning Frameworks

**Dr. Yashaswini S[1], Dr. Jayanthi M G[2]**

[1]Assistant Professor, Department of CSE, Cambridge Institute of Technology, Bangalore, India.
[2]Professor, Department of CSE, Cambridge Institute of Technology, Bangalore, India.

**Abstract**
Deep learning frameworks promote the development of artificial intelligence and demonstrate considerable potential in numerous applications. However, the security issues of deep learning frameworks are among the main risks preventing the wide application of it. Attacks on deep learning frameworks by malicious internal or external attackers would exert substantial effects on society and life. We start with a description of the framework of deep learning algorithms and a detailed analysis of attacks and vulnerabilities in them.
We propose a highly comprehensive classification approach for security issues and defensive approaches in deep learning frameworks and connect different attacks to corresponding defensive approaches. Moreover, we analyze a case of the physical-world use of deep learning security issues. In addition, we discuss future directions and open issues in deep learning frameworks. We hope that our research will inspire future developments and draw attention from academic and industrial domains to the security of deep learning frameworks.

**Keywords:** Deep Learning, AI, malicious comprehensive classification, security issues.

**Introduction:**
The successful application of deep learning are in many fields, Artificial Intelligence (AI) has attracted increasing attention. Owing to the development of Graphics Processing Unit (GPU), deep learning algorithms and large-scale datasets can solve problems in various fields. Moreover, many practical applications and systems are driven by deep learning algorithms.
Companies, ranging from Information Technology (IT) firms to automobile makers (e.g., Google, Tesla, Baidu, Mercedes, and Uber), are testing driverless cars, which require deep learning techniques. In addition, major phone manufacturers offer facial authentication features for unlocking phones, and a number of behaviorbased malware and anomaly detection solutions are based on deep learning. Although deep learning can bring certain conveniences, it is prone to numerous vulnerabilities.
Recent research has found that deep learning is vulnerable to well-designed adversarial samples, which can easily fool a well-behaved deep learning model. Szegedy et al. first generated small perturbations in an image classification problem and deceived the most advanced Deep Neural Network (DNN) with high probability. As a result, samples misclassified by a DNN are called adversarial samples.
We classify deep learning attacks by attack type, adversarial knowledge, attack phase, attack frequency, adversarial specificity, and attack method. The attacks can be divided into poisoning and evasion attacks.

Poisoning attacks add adversarial data to the training sample to influence the training of the classifier and obtain the wrong classifier. Evasion attacks use adversarial examples in the inference stage to make the classifier produce an error output.

In terms of adversarial knowledge, attacks can be classified into white-box attacks, black-box attacks, and semi- whitebox attacks. If an attacker fully masters the content of the deep learning system, such as the dataset and algorithm used, the structure of each layer of the network, and so on, then an attack based on this realization is called a white-box attack. An attack that knows only a part of this knowledge is called a semi-white-box attack. A completely ignorant attack is called a blackbox attack.

The generation of adversarial samples is based on understanding model structures and parameters to destroy deep learning model processes or make wrong predictions. This type of attack, including those based on obfuscated gradient and root mean square gradient, is called the white-box attack. Meanwhile, the black-box attack is limited by knowledge on the model structure and parameters. Goodfellow et al. claimed that the neural network is affected easily by small disturbances from inputs.

The proposed the Fastest Gradient Sign Method (FGSM) to generates adversarial samples. Su et al. proposed a black-box DNN attack that makes only differential perturbations to one pixel, which performs well at different image sizes.

Defense measures were proposed to defend against such attacks, like gradient masking method was proposed for a large number of deep learning- security- critical applications.

The use of deep learning frameworks, such as TensorFlow, Caffe, and Torch, allows application developers to not pay attention to underlying implementation details, thereby substantially improving the development efficiency of AI applications. However, the efficiency of these deep learning frameworks is doomed by the complexity of the framework, and the more complex the system, the more likely the security risks.

**Methodology:**

The research paper uses a deep learning software that identifies traffic signs to describe the types of attacks and threats that deep learning frameworks may be exposed to in practical applications. By simulating real physical scenarios, we analyzed potential problems in the implementation of algorithms.

The case of an attack in deep learning is shown in Fig. 1.1. We choose road signs as our research sample, because road signs are relatively simple, thereby making hidden disturbances challenging. In addition, road signs exist in noisy and changeable environments, such as the distance and angle of the observation camera used as well as lighting conditions. Moreover, this case has high research value.

Traffic signs, as important elements affecting vehicle safety, should be accurately recognized by algorithms despite the presence of adversarial physical disturbances. For a deep learning algorithm to realize the correct identification of road traffic signs, various forms of attacks against it may exist. Figure 1.1 shows an adversarial example that applies algorithms to construct robust perturbations against the deep learning implementation.

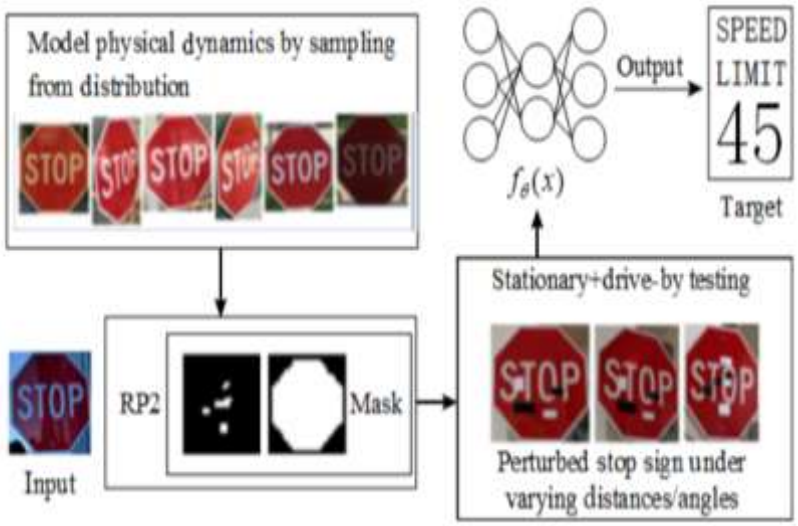


**Fig 1.1 Trafic Sign Adverserial Attack**

The proposed is a two-stage experimental design to verify the physical world attack algorithm. The first stage depends on viewing camera was set up to various distance/angle configurations. The second stage simulates an autonomous vehicle in a field test in which a car was driven toward an intersection in uncontrolled conditions.

**Results & Discussion**

The test, Laboratory for Intelligent & Safe Automobiles (LISA), a US traffic sign dataset containing 47 different road signs, and the German Traffic Sign Recognition Benchmark (GTSRB). The classifiers were built on the LISA-CNN model where, LISA and GTSRB- CNN were trained on the GTSRB.

Using object-constrained poster and sticker attacks, the robust perturbations were created for real road signs. For stationary image the poster attacks were 100% successful and drive-by tests 80% successful. The extracted video frames, demonstrated an accuracy of 87.5%.

**Conclusion:**

Starting from the basic composition structure and principles of deep learning, this study describes security problems behind the application of deep learning and summarizes classic attack algorithms for deep learning technologies and development processes. Moreover, it also confirms that adversarial samples against deep learning are widespread. Studying confrontational algorithms can help us better understand and learn deep learning principles and its training and prediction processes.

**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,andZ.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,andF. C. Sun, Machine learning-based multimodal 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 11t h USENIX Workshop on Offensive Technologies, Vancouver, Canada, 2017.