ESTIMATING ROBUSTNESS OF DEEP LEARNING MODELS BY THREE ADVERSARIAL ATTACKS

Truong Phi Ho, Le Thi Ngoc Anh, Phan Xuan Khiem, Pham Duy Trung*
Vietnam Academy of Cryptography Techniques

ARTICLE INFO

Received: **28/4/2023**
Revised: **24/5/2023**
Published: **24/5/2023**

KEYWORDS

Adversarial attack
Targeted attack
Non-targeted attack
Robustness
Benchmark

ABSTRACT

Deep learning is currently an area of interest in research and development by scientists around the world. Deep learning models are deployed and applied in practice for work and social life. However, deep learning has many potential risks related to security in applications, especially recently adversarial attacks using adversarial examples are a big challenge for deep learning in particular and machine learning in general. To test the robustness of the machine learning model, we propose to use three adversarial attacks to calculate the benchmark, the experimental attack methods on the MS-COCO dataset are being used to train the machine learning model, training and testing for the YOLO model. The article summarizes the results of the successful attack rate using the proposed indicators according to the research through the experimental process conducted by the authors to verify the robustness of the deep learning model in general. The comprehensive experiments in the study were performed on the YOLOv7 model to test and evaluate the robustness of the YOLOv7 model, which is also a popularly used deep learning model and is considered to be advanced today.

KIỂM TRA ĐỘ MẠNH MẼ CỦA MÔ HÌNH HỌC SÂU BẰNG BA CUỘC TẤN CÔNG ĐỐI KHÁNG

Trương Phi Hồ, Lê Thị Ngọc Ánh, Phan Xuân Khiêm, Phạm Duy Trung*
Học viện Kỹ thuật Mật mã

THÔNG TIN BÀI BÁO

Ngày nhận bài: **28/4/2023**
Ngày hoàn thiện: **24/5/2023**
Ngày đăng: **24/5/2023**

TỪ KHÓA

Tấn công đối kháng
Tấn công có mục tiêu
Tấn công không mục tiêu
Độ mạnh mẽ
Điểm chuẩn

TÓM TẮT

Học sâu hiện đang là lĩnh vực được quan tâm nghiên cứu và phát triển bởi các nhà khoa học trên thế giới. Các mô hình học sâu được triển khai và ứng dụng nhiều trong thực tiễn phục vụ công việc và đời sống xã hội. Tuy nhiên học sâu lại tiềm tàng nhiều rủi ro có liên quan đến an toàn trong các ứng dụng, đặc biệt gần đây các cuộc tấn công sử dụng mẫu đối kháng đang là thách thức lớn đối với học sâu nói riêng và học máy nói chung. Để kiểm tra được độ mạnh mẽ của mô hình học máy, chúng tôi đề xuất sử dụng ba cuộc tấn công đối kháng để tính toán điểm chuẩn, các phương pháp tấn công thực nghiệm trên bộ dữ liệu MS-COCO đang được dùng để huấn luyện và kiểm tra đối với mô hình YOLO. Bài báo thống kê kết quả tỉ lệ tấn công thành công bằng các chỉ số đề xuất theo nghiên cứu thông qua quá trình thực nghiệm do nhóm tác giả thực hiện để kiểm chứng độ mạnh mẽ của mô hình học sâu nói chung. Các thực nghiệm toàn diện trong nghiên cứu được thực nghiệm trên mô hình YOLOv7 để kiểm tra và đánh giá độ mạnh mẽ của mô hình YOLOv7, đây cũng là mô hình học sâu đang được sử dụng phổ biến và được đánh giá là tiên tiến hiện nay.

DOI: https://doi.org/10.34238/tnu-jst.7842

* Corresponding author. *Email: trungpd@actvn.edu.vn*

1. Introduction

Deep Learning is a robust and highly accurate machine learning method. However, it also has drawbacks when it can be easily fooled by adversarial images, also known as adversarial samples, which are samples or images contain minor noise used to trick neural networks (NN) into misclassification. Deep learning models using deep neural network (DNN) were developed and proved efficiency in reality; DNNs when applied in practical applications can be attacked by other structural modifications in compare with classified objects [1], [2]. The DL models attack problems is not only limited to the DNNs used in the field of computer vision but also the DNNs in natural language processing such as speech, text,…

Security in machine learning has always been considered as an important subject for developing and improving robustness of models [3], [4]. The first attacked machine learning model is SVM (Support Vector Machine) [5]. In many subsequent studies, authors found that DNNs also tend to be attacked in a fairly simple way by using algorithms based on Fast Gradient Sign Method (FGSM), which is implemented by computing gradients of input image and then fool the model [6].

Currently, Adversarial Examples (AE) attacks are performed to fool DNNs. There are many ways to classify antagonistic attack methods that can be divided into: white box attack, black box attack; this is called pattern-knowledge classification or attack by the type of outcome desired by the adversary called targeted attack and non-targeted attack [7]. In the introduction of the article, the authors approach and present in the direction of classifying targeted attacks and untargeted attacks.

Non-targeted Attack: In this case, aim of AE is to cause the classifier to predict inaccurately for any object no matter what that object is. Some of typical attacks presented below are those with a certain understanding of the model, included in untargeted attack, such as:

FGSM attack: Studied by Goodfellow et al since 2015 [8], using fast gradients optimization method to create adversarial attacks. This is a quite popular method and has been improved my many group authors. For an input image, this method computes gradients of a loss function with respect to the input image and then uses the sign to create a new image that maximize the loss function. This image is called an adversarial image. This will also be attack tested in this article to give results about robustness of the machine learning model.

Square attack [9]: Square attack technique is quite similar to FGSM attack, however, it uses an optimization method to find larger variations in input data insteade of using only gradients of the model to create small variations. Specifically, this technique searches for an image that nearly identical to the original image but differs in some details to make misclassification.

The above attacks, if we approach the classification according to the understanding of DL model is called white-box attack, which requires attackers to have a certain knowledge about the model. This method can cause serious damage to the security system and lead to loss of important information or data leak. Therefore, software developers and cybersecurity professionals need to regularly check and update the security of machine learning systems and models to deal with these types of attack. In addition, another attack method according to the understanding about the model is black-box attack, where the adversaries can only access and modify the input and output of the model, thus, this is the most commonly used method in reality. According to Bhambri et al. [3], black-box attack is divided into four groups based on the method they use: gradients estimation, transformation likelihood, local and combinatorial search.

Targeted Attack: In this case, adversaries aim to modify predictions of classifier to typical kinds of objects or classes. Some of the attacks belonging to targeted attacks can be mentioned as ordinal optimization methods based on attacks to estimate gradients directly (Zoo in [10]). In this article, targeted attack proposed fool DNN by changing pixels [11], this is classified as a black

box attack, which has a great influence on the model because with some minor changes, namely a single pixel change in the input image, the adversaries are not necessary to have an understanding or architecture of the model. This can cause the model to misclassify the image according to the attackers' intent. This method was chosen in study as a tool to measure robustness of machine learning models because through the attack it is possible to check the vulnerabilities of the machine learning model by experimental results presented in the following sections. Because the attack is aimed of misidentification of the model as predefined classifier, the attacker can make use of it to manipulate the behavior of the system.

Another kind of attack is proposed by the authors by adding Gaussian noise to the image, as an adversarial attack used as a tool to test the robustness of machine learning models. This is an attack the authors classify as a non-targeted attack. Adversarial samples are taken into classification model only for the purpose of fooling the model without knowing exactly what classifier the model will misidentify. Unlike the Pixel Attack, this method also shows the vulnerabilities of the machine learning model through experimentation. However, this attack is more general than the Pixel Attack because it can cause more complex and various perturbations to the input data.

This paper uses adversarial attacks selected and presented in the introduction by the authors, which will be shown in detail in the following sections; through experiments to calculate benchmarks, then draw conclusions about the robustness of machine learning models in general, and deep learning in particular, which are tested on YOLOv7 model. This will help us understand more clearly the weaknesses of the DL model and suggest solutions to improve the model before implementing it in practice.

The rest of the paper is organized as follow: Section 2 presents the research methodology. Section 3 is experiments and results. Conclusions will be discussed in section 4.

2. Methods

In this part, the authors present theory of three adversarial attack methods measured and used in model testing, including: Fast Gradient Sign Method (FGSM), Pixel change attack method, attack method by adding Gauss noise.

2.1. FGSM method

FGSM is one of the most simple and effective adversarial attack techniques for machine learning models using DNNs. This method was introduced by Ian Goodfellow et al in 2015 [8].

The aim of FGSM Attack is to cause minor changes in the input data, small enough to make the machine learning model predict wrongly, but not large enough to change the meaning of the data. Since this method only uses the gradients of machine learning models, it is very fast and simple to implement. However, FGSM Attack also have certain limitations.

Firstly, this method only works well for gradient – based machine learning models, and is not as suitable for non – gradient models as decision tree-base model. Secondly, this method can be easily detected and countered if the machine learning model is trained with adversarial techniques. Expression of FGSM is presented generally and cited in many studies according to formula (1):

$$X^{adv} = X - \varepsilon sign(\nabla_X J(X_N^{adv}, y_t)) \quad (1)$$

Where X^{adv} is the adversarial image, X is the original input image, y is the original input node, ε is the multiplier to ensure low noise and J is loss function. It is not possible to guarantee the adversarial samples created by this method are the same as the real object in its original image. Figure 1 illustrates the difference between the original image and the image when implementing FGSM method.

(a) *(b)* *(c)*

Figure 1. *Illustrating images: (a) original image; (b): noised image with low ε 0.05; (c): medium ε 0.25*

The use of FGSM Attack to test robustness machine learning models is necessary because it allows us to evaluate the model's ability to resist external attacks. If a machine learning model cannot cope with FGSM attacks, it may be vulnerable to other attack techniques. Therefore, the use of FGSM is necessary and many groups of authors study in practice.

2.2. Attack by using Gaussian noise

Gaussian Noise is a common type of noise used in image and audio processing. This method adds an amount of Gaussian noise with a mean of 0 and standard deviation of a given value ($\varepsilon \in Z^{+}$) to the input data of DL model. This amount of noise is very small, but enough to cause misprediction of the model. In the case of digital images, natural noise can be generated due to various reasons: damaged image sensor or affected by external factors; lack of light or overheating of device at the time of taking picture; channel noise,…

In this study, the authors used the Gaussian noise added by computer programming shown in Figure 2, with the noise form according to formula (4) (see in [12]):

$$w(x,y) = s(x,y) + n(x,y) \quad (2)$$

Where x, y are the coordinates of the pixel noise applied: $s(x,y)$ is the intensity of the original image; $n(x,y)$ is the noise added to original image; $w(x,y)$ is the altered image obtained by applying noise.

(a) *(b)* *(c)*

Figure 2. *Illustrating images: (a) original image; (b): noised image with low ε 50; (c): medium ε 250*

Attack by using Gaussian noise can also help detect vulnerabilities and weaknesses of DL models. By adding noise to the input data, we can create input cases where the model misclassifies. Analyzing these cases can help us understand more about the way model processes data and find the errors that the model is dealing with.

2.3. Pixel Attack

The research team [11] proposed a simple method of generating AE by shuffling random segments of the image to generate AE. Through research, attack by changing pixels can be summarized as follow:

First, using an image $x \in R^{(3,w,h)}$, the main idea of this algorithm is to sample a random portion of adjacent pixels from it. As P is a list of indexes (i, j) are varied pixel coordinates, defined to lie within a rectangle whose vertices are $\left[(o_x, o_y), (o_x + w_{p,} o_y), (o_x, o_y + h_p), (o_x + w_p, o_y + h_p)\right]$. Assign this list as $P = \left[(o_x + i, o_y + j)\right] \forall\, i \in, \{(0, \dots, w_p\}, j \in (0, \dots, h_p)\right]$ has the value $|P| = w_p \times h_p$ assumes that the patches do not exceed the boundary of the image, otherwise, it is moved in such a way the indexes are all inside the images. The mapping function m is defined according to (3):

$$m\colon (i, j) \in \mathrm{P} \mapsto N_+^2 \in [1, \mathrm{h}] \times [1, w] \tag{3}$$

For each point in the patch, the position where the pixel should be moved in the original image. With this function, the authors identify each pixel adversarial image according to (4):

$$\bar{\mathrm{x}}_{\mathrm{i,j}} = \begin{cases} m_{(i,j)}, & (\mathrm{i}, \mathrm{j}) \in \mathrm{P} \\ x_{(i,j)}, & \text{otherwise} \end{cases} \tag{4}$$

Testing the robustness of machine learning models by Pixel Attack can help developers improve their models by increasing resistance to small variations in the input data. Therefore, the Pixel Attack was chosen by the authors to apply in the experiment of this study.

2.4. Indicators for evaluating deep learning models

In this section, the authors aim to measure the success of adversarial samples in test execution in order to gain a more realistic view of the actual impact of applied attacks. Through research [13], we measure robustness of model and evaluate based on two different indicators:

Adversarial Success Rate (ASR) is calculated according to (5):

$$\mathrm{ASR} = \frac{\#\ \text{perturbed samples}}{\#\ \text{all samples}} \tag{5}$$

The turbulence test image rate is successful and it provides a basis about the adversaries' abilities to fool the unprotected target model. Therefore, ASR indicators provide information like RA (Robust Accuracy) of the attacks.

Attack Success Rate under Defense (ASRD): This is an extension of the ASR indicator by assuming the fact that excessive perturbations can be detected at the time of the attack. To measure the performance of attacks under protection of the model, the authors formula (6) to calculate the rate of successful attacks:

$$ASRD = FNR.ASR \tag{6}$$

FNR is False Negative Rate (the rate of not detected perturbed samples).

3. Experiments and results

3.1. Experimental method

The authors use two datasets MS-COCO [14] and ImageNet [15] to experiment in this paper to measure robustness of deep learning model. Microsoft's MS-COCO dataset is a large dataset in the field of computer vision and natural language processing. This dataset includes more than 330000 images and 4 million labels on objects: people, animals, objects and vehicles. For the ImageNet dataset, which is also commonly used in deep learning and computer vision, it includes: more than 14 million images labelled as keywords, representing more than 20000 different object classes.

The selected deep learning model is the YOLO model version 7 (YOLOv7), the YOLOv7 model [16] was selected in this study to experiment to measure the robustness of the

representative deep learning model. Through research and understanding of the authors, YOLOv7 model is a new updated version with many improvements such as increased learning speed, improved accuracy and better multitasking ability.

FGSM Attack and Gaussian noise: Because these are two attacks using perturbation that need to be done in large numbers and the image used does not require a single object, selected images are added noise at low and medium levels (ε index is different in each experiment type according to the theory of each attack).

Pixel Attack: performed with the original images and transformed images, using the images in the ImageNet dataset selected by the author group due to the requirement of a single-object according to the research [11].

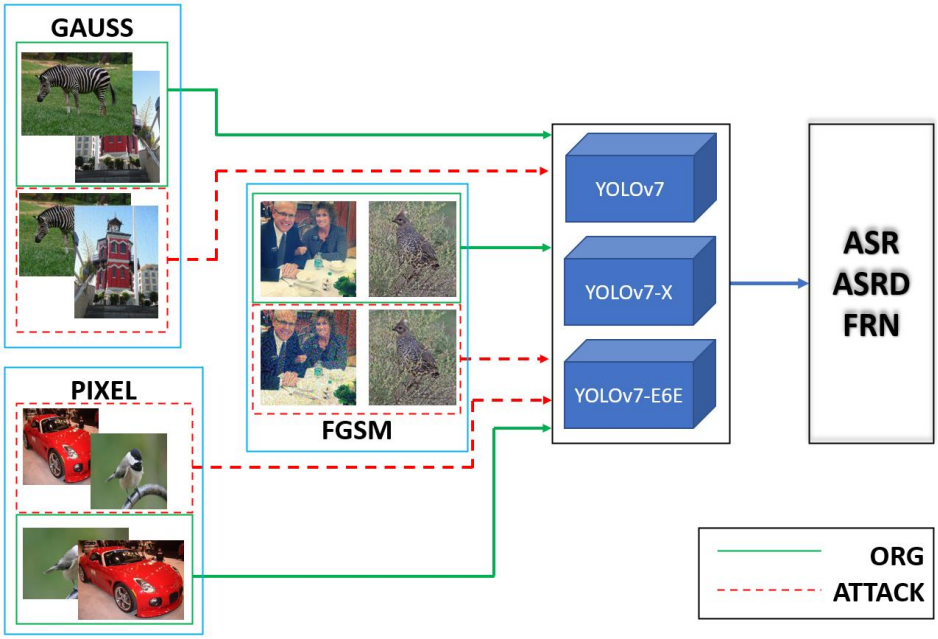

Figure 3. *Experimental method through 3 attacks*

Figure 3 illustrates generally about the process of conducting model evaluation experiments through 3 proposed attacks. We implemented identification by using YOLOv7 model with 3 different pretrained weights to evaluate the model objectively and holistically. After the experiment, we present statistics on the indicators of ASR, ASRD and FNR. The results of experiments are detailed in Section 3.2.

3.2. Experimental statistics

In the experiments, the authors used the Python programming language to experiment and calculate the results. For the FGSM method and the Gaussian noise attack, the authors use 5000 images randomly selected from the MS-COCO dataset.

Results of FGSM attack and Gaussian noise which are used to evaluate YOLOv7 model using 3 pretrained weights selected by the authors team (researched by [16]) with different model parameters are shown in Table 1.

Observing the results of Table 1, the authors found that there is a significant difference in 3 indicators between the two types of FGSM attack and Gaussian noise. Hence, it can be seen that in Gaussian noise attack, since density of noise pixels is randomly generated based on standard distribution, the success rate as well as the efficiency is lower than the FGSM attack, which adds perturbation computationally by modifying the gradients optimization parameters with the aim of increasing the value of objective function.

With the YOLOv7 weights selected similar to the previous 2 attacks, the authors implemented experimental statistics for Pixel Attack on 200 images from ImageNet dataset. Specific results are shown in Table 2.

Table 1. *Statistic of YOLOv7 model evaluation through FGSM attack & Gaussian noise Attack*

Type of attack	Weight	Noise (ε)	Perturbed samples/ All Samples	Indicators (%)		
				ASR	ASRD	FNR
FGSM	YOLOv7	Low	3655/5000	73.10	59.94	0.82
		Medium	4537/5000	90.74	68.96	0.76
	YOLOv7-X	Low	3580/5000	71.60	52.98	0.74
		Medium	4460/5000	89.20	58.87	0.66
	YOLOv7-E6E	Low	3328/5000	66.56	47.92	0.72
		Medium	4088/5000	81.76	53.96	0.66
GAUSS	YOLOv7	Low	2868/5000	57.36	43.59	**0.76**
		Medium	3092/5000	61.84	43.29	0.70
	YOLOv7-X	Low	2869/5000	57.38	43.61	**0.76**
		Medium	3073/5000	61.46	46.71	**0.76**
	YOLOv7-E6E	Low	2484/5000	49.68	37.76	**0.76**
		Medium	2663/5000	53.26	40.48	**0.76**

Table 2. *Indicators of Pixel Attack on YOLOv7 model*

TT	Weight	Peturbed Samples/ All Samples	Indicators (%)		
			ASR	ASRD	FNR
1	YOLOv7	72/200	36	0	0
2	YOLOv7-X	46/200	23	11.50	0.5
3	YOLOv7-E6E	47/200	23.5	11.75	0.5

According to the results of Table 2, targeted attack by changing pixels have a low success rate (less than 40%). Therefore, attacking in this way may not be very effective against an advanced deep learning model like YOLOv7, or in other words, YOLOv7 is robust to this type of attack. Through ASR indicator in Table 2, it is estimated that in models having better parameters like YOLOv7-E6E, it is more robust, accurate and difficult to deceive than YOLOv7.

4. Conclusion

Several research groups have also studied how to evaluate and measure the robustness of machine learning models [17]. However, it is only limited to datasets with small images such as Cifar-10 or MNIST. Our study is different from previous studies in terms of experimental methods and indicators. In our experiment, the authors found that if the more noise (ε) in the first two attacks, DL model will be biased, and the accuracy loss will be higher. However, the possibility of being detected with data input interference will also be higher. From that, it cannot be concluded that the index ε is proportional to ASR, ASRD indicators.

It can be concluded that, in an attack that wants to deceive or falsify the DL model, if the data input deceives the DL model, it is also easily detected and overcome by defensive methods. In the next studies, the authors want to evaluate the indicators with many others, more advanced deep learning models such as YOLOv8 by using newly published adversarial attacks on more diverse datasets.

REFERENCES

[1] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, PMLR, 2018, pp. 284-293.

[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physicalworld attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625-1634.

[3] S. Bhambri, S. Muku, A. Tulasi, and A. B. Buduru, "A survey of black-box adversarial attacks on computer vision models," *arXiv preprint arXiv:1912.01667*, 2019.

[4] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, 2006, pp. 16-25.

[5] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," *arXiv preprint arXiv:1206.6389*, 2012.

[6] F. Behnia, A. Mirzaeian, M. Sabokrou, S. Manoj, T. Mohsenin, K. N. Khasawneh, L. Zhao, H. Homayoun, and A. Sasan, "Code-Bridged Classifier (CBC): A Low or Negative Overhead Defense for Making a CNN Classifier Robust Against Adversarial Attacks," *arXiv:2001.06099v1*, 2020.

[7] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, 2019, pp. 2805-2824.

[8] S I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[9] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square Attack: a query- efficient black-box adversarial attack via random search," *arXiv preprint arXiv:1912.00049v3*, 2020.

[10] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, doi: 10.1145/3128572.3140448.

[11] P. H. Truong, T. N. Hoang, Q.T. Pham, M. T. Pham, and D. T. Pham, "Adversarial attacks into deep learning models using pixel tranformation," (in Vietnamese), *TNU Journal of Science and Technology*, vol. 228, no. 02: Natural Sciences - Engineering - Technology, pp. 94-102, 2023.

[12] Tristan, Alex, Kostya, I. J. Roth, J. Hallberg, and T. Spiegel, "Gaussian Noise," *Hasty's end-to-end ML platform*, 2019. [Online]. Available: https://hasty.ai/docs/mp-wiki/augmentations/gaussian-noise. [Accessed Dec. 21, 2022].

[13] P. Lorenz, D. StraBel, M. Keuper, and J. Keuper, "Is RobustBench/AutoAttack a suitable Benchmark for Adversarial Robustness?," *arXiv:2112.01601v2*, 2022.

[14] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of 13th European Conference on Computer Vision–ECCV*, Springer International Publishing, 2014, pp. 740-755.

[15] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 248-255.

[16] C. Y. Wang, A. Bochkovskiy, and H. Y. Liao, " YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv: 2207.02696*, 2022.

[17] C. Ma, C. Zhao, H. Shi, L. Chen, J. Yong, and D. Zeng, "Metaadvdet: Towards robust detection of evolving adversarial attacks," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 692-701.