2020 International Conference on Materials, Control, Automation and Electrical Engineering (MCAEE 2020)
ISBN: 978-1-60595-680-0

Application of Adversarial Sample Attack in Aerial Photo Identification of Transport Vehicle

Mingjiang Zhang, Weihu Zhao*, Hongwei Li and Chengyuan Wang

College of Information and Communication, National University of Defense Technology, Xi'an, Shaanxi, 710106, China

*Corresponding author. Email: zhaoweihuandy@126.com

Abstract. The adversarial samples can cause the convolutional neural network model to output incorrect results. It is proposed to paste the generated adversarial sample patch on the roof of the transport vehicle to prevent aerial identification of the drone and achieve the attack on the target detection system. By producing aerial transport vehicle datasets, a YOLOv2-based target detection model is trained in the Pytorch deep learning framework, and the adversarial patch is trained by the GAN (Generative Adversarial Networks) called adversarial-yolo that can make the target detection failed. After simulation and comparison, the transport vehicle with a small adversarial patch can successfully and stably attack the target detection model, making it unable to detect the target, and the operation is flexible. The research can provide a certain reference value for the defense and camouflage methods of important ground targets against unmanned aerial intelligent detection devices.

Keywords: adversarial sample attack, transport vehicle target detection, YOLOv2.

1. Introduction

In recent years, artificial intelligence is accelerating its penetration into various industries. The continuous breakthrough of deep neural networks in many aspects such as target recognition and natural language processing will significantly increase the level of intelligence in the future society. However, researchers have found that the smartly constructed “adversarial samples” [1] can successfully deceive existing intelligent recognition systems, which has drawn great attention to the safety of artificial intelligence. In modern industry, it has become the norm to make use of drones equipped with intelligent identification system to conduct high-frequency and full-coverage search over the target area, so as to quickly obtain information such as the category, number and location of the targets[2]. In order to combat the drone search, important targets such as vehicles and buildings are usually protected and camouflaged by coating, camouflage nets and false targets. However, these traditional technical means are vulnerable to the impact of the natural environment and have poor reliability, so they cannot adapt to the antagonistic environment under information conditions.

Transport vehicle is an important environmental goals of drone identification, so this article attempts to generate the adversarial patch which can attack the transport vehicle target detection system. Transport vehicles with adversarial patch can be hidden under the drone's camera. This provides a new technical support for the defense camouflage method of important targets in the face of the drone search.

2. Target detection algorithm in drone aerial photo identification

In order to counter the intelligent identification of drone, it is necessary to understand the target detection process of drone. At present, unmanned search has become highly intelligent, and ground targets appear to have nowhere to escape in the drone identification. By creating a patrol mission, using track planning technology and geographic information system, the detection results of ground targets can be transmitted back to the rear quickly and in real time. In the process of drone target detection, both airborne and off-line recognition are inseparable from the target detection algorithm based on deep learning convolutional neural network. At present, target detection algorithms of convolutional neural network are mainly divided into two categories. One is region-based convolutional neural network, such as Faster R-CNN (Faster Region-based Convolutional Neural Network)[3]. The other is regression-based convolutional neural network, which is typically represented by SSD (Single Shot MultiBox Detector)[4], YOLO(You Only Look Once)[5], YOLOv2[6], and YOLOv3[7] series of target detection algorithms. Among them, YOLOv2 improved some shortcomings of the previous YOLO algorithm, adopted darknet-19 network structure as the feature extraction network, combined with image fine-grained features, connected the shallow features with the deep features, which was helpful for the detection of small-size targets. The architecture diagram of YOLOv2 is shown in figure 1.[6]

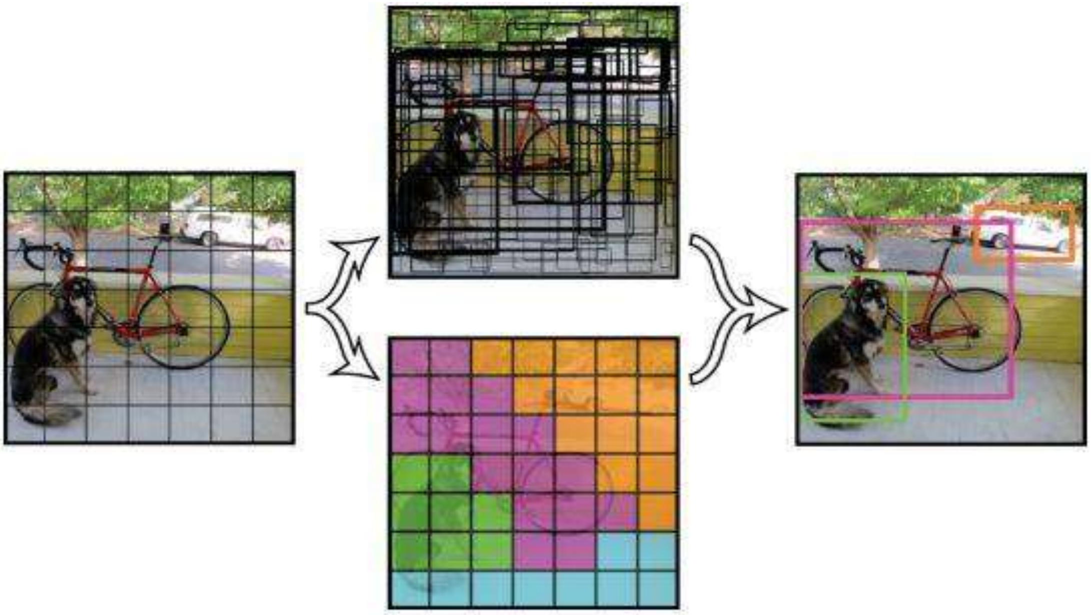


Figure 1. YOLOv2 structure diagram.

Because YOLOv2 has a good detection effect on small targets, while in the pictures collected by drone, the target is usually small, many Chinese drone technology companies have adopted YOLOv2 as the main algorithm of their target detection module. Therefore, this paper will generate adversarial sample targeted at the transport vehicle detection to attack the recognition system of drone based on YOLOv2.

3. An attack method of transport vehicle target detection based on adversarial patch

3.1. Adversarial sample for transport vehicle target detection

Adversarial samples were first discovered and proposed by Szegedy[1], which were generated by the GAN. The generation of adversarial sample is also called the attack based on adversarial sample. Due to the convolutional neural network has highly nonlinear, when add some weak disturbance to network input, can mislead the trained neural network model output error classification results, the disturbance is called “adversarial disturbance”. Data added with this “adversarial disturbance” are called adversarial sample, but the samples with this disturbance are hard for humans to find. According to whether the output error result is specified in advance, the counter attack can be divided into Targeted Attack and Non-targeted Attack[8]. Targeted Attack refers to the adversarial sample that needs to be determined as the specified type, and if it is classified successfully, the attack is successful. However, a Non-targeted Attack is when the adversarial sample is misclassified or missed. The research scenario in this paper is a Non-target Attack. When the target detection system of drone fails to identify the transport vehicle with the adversarial sample, the attack is considered successful.

3.2. Adversarial patch generation method for transport vehicle target detection based YOLOv2

The adversarial patch is equivalent to adversarial sample, so how to generate the adversarial patch? The most direct method is to modify the sample within the given range of disturbance, so that the loss function of the modified sample on the target detection model is optimal, so that the problem of generating adversarial patch can be transformed into the optimization problem of spatial search. At present, there are two main types of attack methods: pixel attack[9] and patch attack[10]. For pixel attack, although its concealment is good, it is difficult to realize in the physical domain. In combination with the application scenario in this paper, we choose the patch attack method.

In 2019, scholar Thys[10] tricked the automatic surveillance camera perfectly with a simple printed pattern, and they succeeded: when you put a printed adversarial patch on your body, the monitor could not find that you. Similarly, in order to make the drone unable to find the real transport vehicles on the ground during aerial photo identification, we can generate the adversarial patch that can deceive and attack the transport vehicle detection system. The network flow chart to generate the patch is shown in figure 2, and the algorithm of drone is YOLOv2.

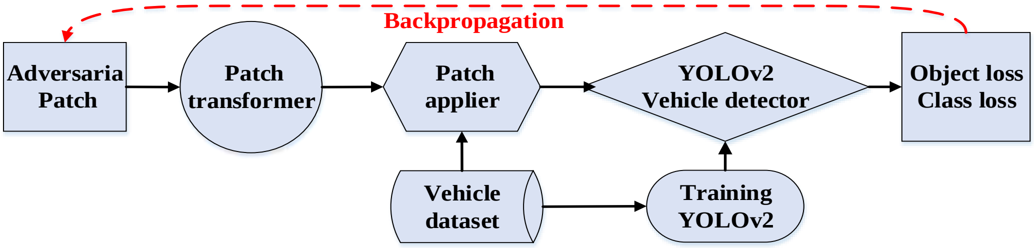


Figure 2. Network flowchart for generating transport vehicle adversarial patch.

As shown in figure 2, firstly, the transport vehicle dataset in the aerial photography environment should be made, and some random grayscale blocks should be placed in specific positions of the image, and then the pre-trained transport vehicle target detection model should be used, which will output the target, category, confidence and other recognition results. Then, after a comprehensive consideration of the recognition result and the printability and color smoothness of the adversarial patch, a loss function L is set, the loss is reverse-transmitted, and the pixel in the image block is updated. Through several hundred iterations, the final adversarial patch can be obtained. The value of function L here can be expressed by equation 1:

$$\begin{cases} L = \alpha L_{nps} + \beta L_{tv} + L_{obj} \\ L_{nps} = \sum_{p_{patch} \in p} \min_{c_{print} \in C} \left| p_{patch} - c_{print} \right| \\ L_{tv} = \sum_{i,j} \sqrt{\left(p_{i,j} - p_{i+1,j}\right)^2 + \left(p_{i,j} - p_{i,j+1}\right)^2} \end{cases} \quad (1)$$

The L_{nps} in the expression stands for the loss value of controlling printing, the L_{tv} stands for the loss value of controlling smoother image block, and the L_{obj} represents the output confidence of the target after the target detection model based YOLOv2. The purpose of training the adversarial patch is to minimize the value of L.

4. Experimental process

4.1. Experimental environment

- Hardware: GPU is NVIDIA GeForce RTX2070; Video memory is 8G; Memory is 32G.
- Software: Deep learning of Pytorch framework, based on Windows10 operating system, CUDA9.2, cudnn7.2.1, Pycharm and Python3.6.

4.2. Make transport vehicle dataset

In order to produce the aerial images of transport vehicle, the rotorcraft drone was used to collect about 200 aerial images of transport vehicles at a height of 200~300 meters from the ground. By means of angular rotation, the images were expanded to 400 pieces. Then, the corresponding XML and Labels files were generated for the photos to conform to the YOLO training format, and the transport vehicle label was abbreviated to “Tran-Veh”.

4.3. Training target detection model

The pretraining weight file “darknet19_448.conv.23” was loaded during the training. Since only one type of target was needed to be detected, the parameters "classes" and "filters" in the CFG configuration file were set to 1 and 30 respectively. In addition, the main parameter settings during the training were shown in table 1.

Table 1. Main parameter settings during training detection model.

Parameter name	Value	Parameter name	Value
Batch	64	subdivisions	32
max_batches	8020	policy	steps
stepsize	4000,6000	learning_rate	0.001
momentum	0.9	decay	0.0005

Through training, the model has a good detection effect on transport vehicles, as shown in figure 3.

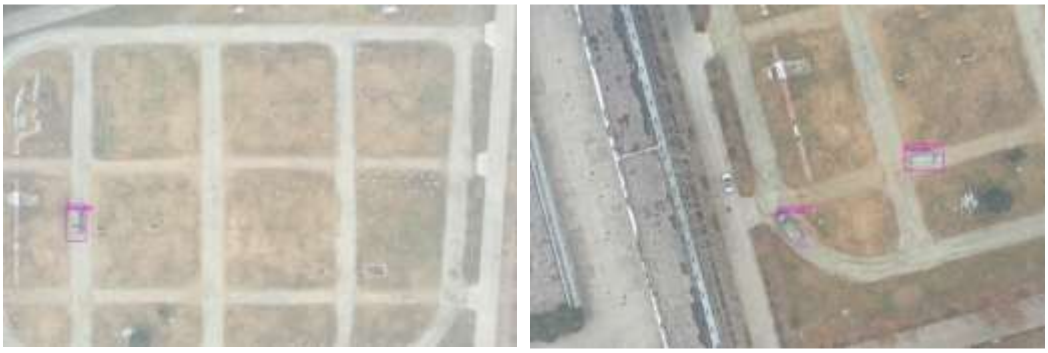


Figure 3. Samples of transport vehicle target detection.

4.4. Training and generating adversarial patch

According to the adversarial patch generation method described in 3.2, the adversarial-yolo[10] framework was used as the GAN in this paper. When training to generate the adversarial patch, the main parameters were set as shown in table 2.

Table 2. The main parameter settings during training adversarial pattern.

Parameter name	Value	Parameter name	Value
n_epochs	500	patch_size	400
batch_size	8	start_learning_rate	0.03
max_lab	14	max_tv	0.165

First, a gray image of 400*400 pixels was taken as the initial status of the adversarial patch. After 500 epochs of training iterations, the patch for the transport vehicle target detection can be generated, as shown in figure 4.

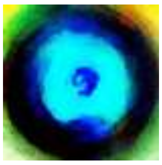
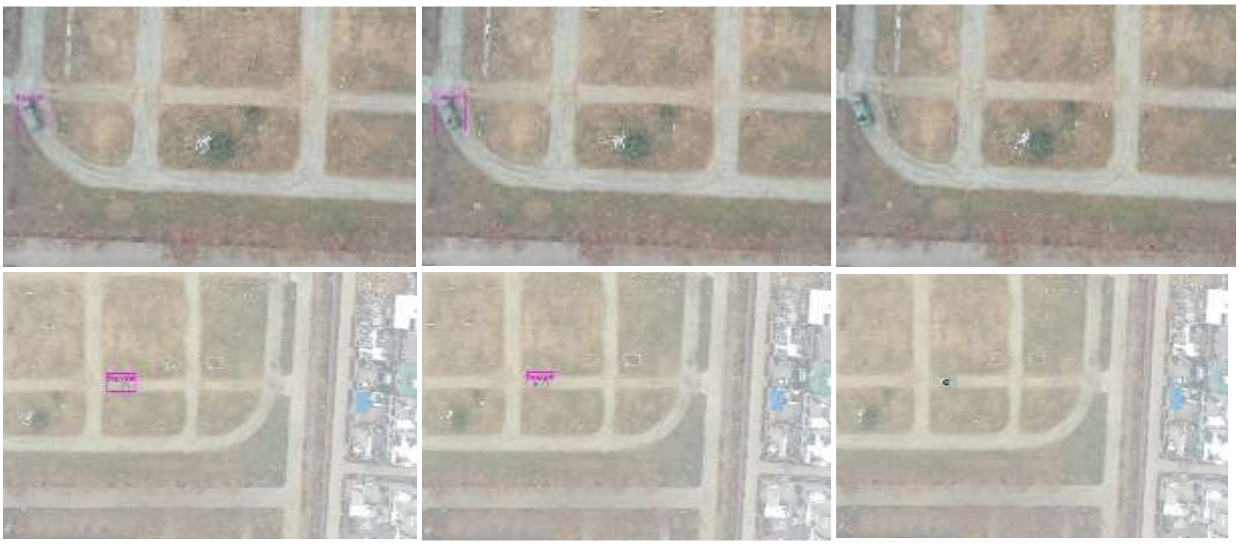


Figure 4. Adversarial patch for transport vehicle detection.

4.5. Adversarial patch based transport vehicle target detection attack display

To test the effect of the generated adversarial patch attacks on transport vehicle target detection, we divided the test dataset into three conditions, the test results are shown in figure 5, namely, the original transport vehicle images(Fig.5-a), the transport vehicles pasted random patch(Fig.5-b) and the transport vehicles with the generated adversarial patch(Fig.5-c) were respectively input into the trained transport vehicle detection network.



a. Original test results b. Test results add random patch c. Test results add adversarial patch

Figure 5. Adversarial patch based attack comparison diagram.

As can be seen from sub-figures a and c in figure 5, after the transport vehicles add the generated adversarial patch, the drone target detection system can no longer detect the transport vehicles. The adversarial patch successfully achieves the stealth from the YOLOv2 algorithm, proving the effectiveness of the attack by adversarial patch. It can be seen from sub-figures b that the vehicles can still be detected after adding a random patch of the same size as the adversarial patch, which proves the uniqueness and pertinence of the generated adversarial patch. In addition, it was found through the test that no matter the adversarial patch was added to the front, middle and tail of the vehicles, or the patch size was slightly adjusted, the attack could all be realized, which indicated that the adversarial patch had certain flexibility in the attack operation.

5. Summary

In this paper, transport vehicles are taken as the detection targets, the drone aerial photography dataset is made, the transport vehicle target detection model based on YOLOv2 is trained by the Pytorch. And based on the GAN adversarial-yolo, the adversarial patch is trained for transport vehicle target detection. Through the test, the transport vehicles with the adversarial patch can realize the attack and stealth of drone target detection system stably. The research has certain reference value for improving the intelligent defense level of ground targets against drone aerial photo identification. Through analysis, the next important research direction is to study the visual concealment of adversarial samples and the migration of adversarial samples to other detection models.

Acknowledgement

This work was supported by the Scientific Research Program Project of National University of Defense Technology, No. “ZK17-03-33”.

References

[1] Szegedy C, Zaremba W,Sutskever I,et al. Intriguing proper-ties of neural networks [EB/OL]. [2013-12-21]. https://arxiv.org/abs/1312.6199v4.

[2] Zhang Mingjiang, Li Hongwei, Zhao Weihu, et al. Application of Deep Learning in the Patrol and Inspection of Military Optical Cable Lines by UAV[J]. Study on Optical Communications, 2018(06): 57-61.

[3] Ren S, He K,Girshick R,et al.Faster R-CNN:Towards Real-Time Object Detection with Region ProposalNetworks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016,39(06):1137-1149.

[4] Liu W, Anguelov D, Erhan D,et al.SSD:Single Shot MultiBox Detector[C]//European Conference on Computer Vision. Springer International Publishing, 2016: 21-37.

[5] Redmon J, Divvala S, Girshick R,et al.You Only Look Once:Unified,Real-Time Object Detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[6] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[J]. arXiv, 2016: 1612.08242.

[7] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv, 2018: 1804. 02767.

[8] Chen Jin-Yin, Shen Shi-Jing, Su Meng-Meng, et al. Black-boxAdversarial Attack on License Plate Recognition System [J/OL]. Acta Automatica Sinica: 1-18[2020-02-06]. https:// oi.org/ 10.16383/ j.aas.c190488.

[9] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019, 23(5):828-841.

[10] Thys S, Van Ranst W, Goedeme T. Fooling automated surveillance cameras: adversarial patches to attack person detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.California, USA: IEEE, 2019.