Road Sign Recognition and Backdoor Attacks of Baidu Pilotless Automobile

Shiwen CAO 12113024

Xuanyu LIU 12110408

Tianyou SONG 12112519

Fanzhong WANG 11911824

Jinlong ZHANG 12112528

***Abstract*—Traffic image recognition is an indispensable function in the field of automatic driving. It requires that the computer can accurately and quickly identify the traffic signs in the captured information. YOLOv5 is an algorithm that can effectively detect the far and small targets with high speed and accuracy. Backdoor attack is an operation aimed at malicious processing of data sets, and it is difficult for human eyes to detect the differences in images, However, after receiving the backdoor attack, many characteristic values of the data have changed, which will affect the judgment of the recognition algorithm. This paper will use the method of adding hidden watermark to simulate the backdoor attack to detect whether the backdoor attack has an impact on the accuracy of YOLOv5.**

***Index Terms*—Backdoor Attack, Watermark, Yolov5**

I. Introduction

A. Background

With the development and application of artificial intelligence technology in license plate detection and driver-less driving. Due to the characteristics of its sliding window model, the traditional target detection algorithm has a certain degree of singularity in feature extraction and matching, and its accuracy and detection speed are relatively poor. Therefore, the deep learn-based target detection algorithm has the advantages of higher efficiency and accuracy, rapidly overtaking the traditional target detection algorithm and becoming the most mainstream target detection algorithm. It is mainly divided into two development directions: two-stage model and one-stage model.

B. Backdoor Attack

Data poisoning attack refers to that the attacker adds a small amount of carefully constructed poison data to the training set of the model, so that the model cannot be used normally in the test stage or assists the attacker to invade the model without damaging the accuracy of the model. The former destroys the availability of the model and is a no-target attack. The latter destroys the integrity of the model for targeted attacks. Data poisoning attack was first proposed by Dalvi, who used this attack to evade detection of spam classifiers. Poisoning attacks that destroy integrity have strong concealment: the poisoned model shows normal prediction ability for clean data, and only outputs wrong results for the target data selected by the attacker. Such attacks that cause the AI model to output incorrect results on certain data can cause significant damage and in some critical scenarios, serious security failures.

C. Motivation

Through the computer to realize the human visual function, grasp and understand the image of the scene, identify and locate the target, determine their structure, spatial arrangement and distribution as well as the mutual relationship between the target and so on, according to the perceived image of the actual target and scene in the objective world to make a meaningful judgment.

II. Experiment

A. Backdoor Attack

1) earlier stage: Through the operation of the frequency domain of the picture, the invisible watermark is added to the picture. The Algorithm used is invisible-watermark [1]. The resulting image is obviously different from the original image and fails to achieve the invisible watermark effect we expected. The reason may be that this method is suitable for large size pictures, and it is difficult to hide the watermark in the picture without being discovered for small size pictures.

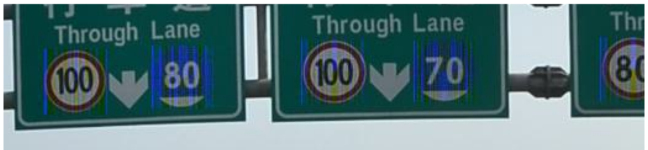


Fig. 1. The invisible watermark images in earlier stage

2) later stage: Eventually the watermark adding methods is FTrojan [2], concrete process is: the parts of the need to add the watermark from the original cutting, part of the invisible watermark image, and will be processed images put back to the original image.

B. Model Training

1) Yolov5: In 2016, Redmon J proposed a new target detection algorithm, YOLO (You Only Look Once). Different from the target detection algorithm based on classification, which

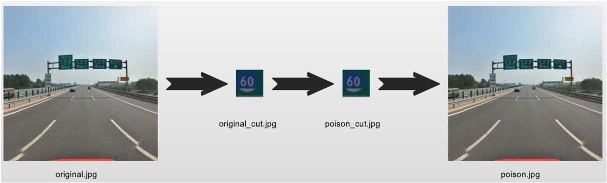
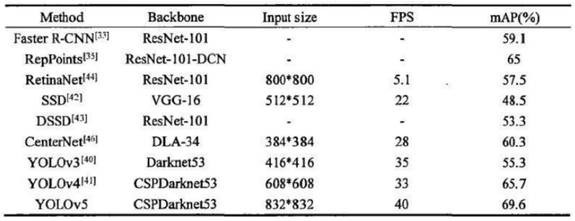


Fig. 2. The invisible watermark images in later stage

uses classifiers to perform detection, the YOLO algorithm treats the target detection framework as a spatial regression problem. A single neural network can obtain the prediction of boundary frame and category probability from the complete image after one operation, which is conducive to the end-to-end optimization of detection performance. Compared to previous networks, YOLOv5 uses Pytorch for the first time, which provides simpler support and easier deployment, and improves image reasoning time and detection average accuracy without any loss of accuracy.



Method	Backbone	Input size	FPS	mAP(%)
Faster R-CNN[33]	ResNet-101	-	-	59.1
RepPoints[35]	ResNet-101-DCN	-	-	65
RetinaNet[44]	ResNet-101	800*800	5.1	57.5
SSD[42]	VGG-16	512*512	22	48.5
DSSD[43]	ResNet-101	-	-	53.3
CenterNet[46]	DLA-34	384*384	28	60.3
YOLOv3[40]	Darknet53	416*416	35	55.3
YOLOv4[41]	CSPDarknet53	608*608	33	65.7
YOLOv5	CSPDarknet53	832*832	40	69.6

Fig. 3. Different target detection algorithms on COCO datasets

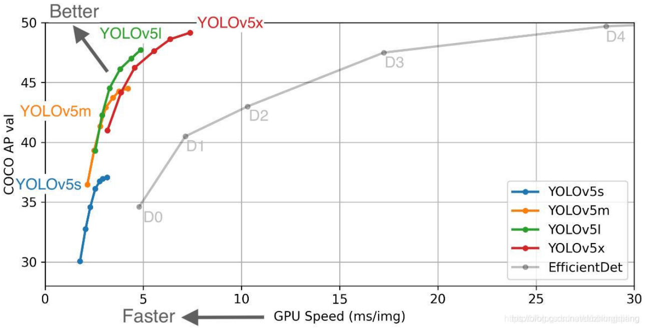


Fig. 4. GPU speed of four Yolov5 models

2) Training Result: AP (Average precision) is the main evaluation standard of target detection model. In the field of target detection, suppose we have a set of pictures containing several targets to be detected, Precision represents how many dozen targets can be detected by our model, and Recall represents how many percentage of all real targets can be detected by our model.

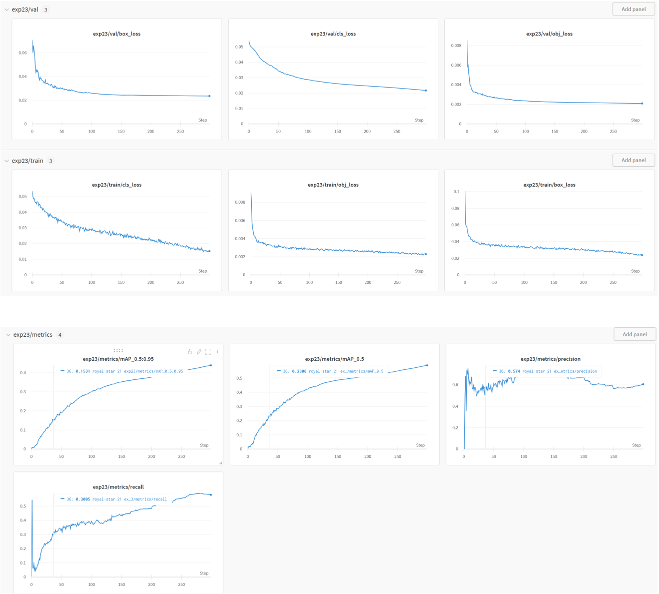


Fig. 5. The test results without poisoning

The possible causes of misidentification are:

- Less similar data, can not train more accurate weight;
- Part type characteristic value of similar images, bring certain difficulty to algorithm identification;
- Datasets with poor quality, image distance is far and less information.

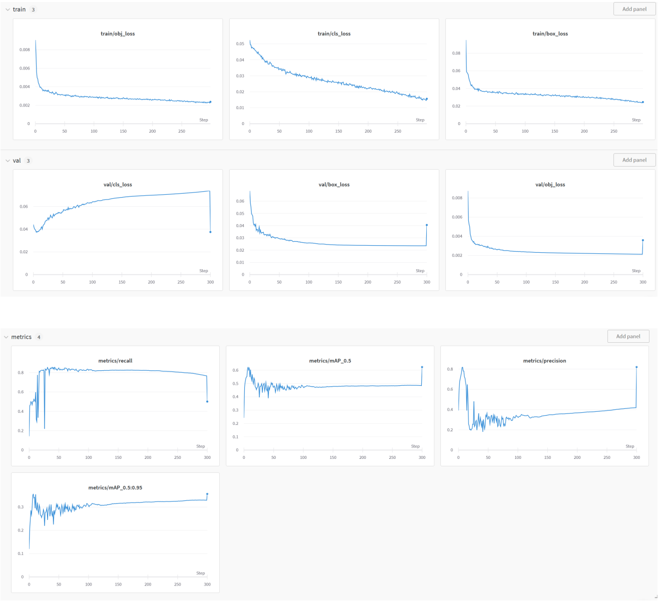


Fig. 6. The test results with poisoning

There are two conjectures as to why the poisoning effect is not obvious. One is that, yolov5 has powerful function, which can effectively filter poisoning information. Second is, Poisoned by the information is not obvious in the image characteristics, has not been identified.

Explain the error in identification after poisoning:

- Less training data, can not train more effective parameters;
- Picture is too small, can identify the characteristics of less;
- Information may be poisoning affects judgment.

REFERENCES

[1] https://github.com/ShieldMnt/invisible-watermark

[2] https://github.com/SoftWiser-group/FTrojan

[3] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. 2022. An Invisible Black-Box Backdoor Attack Through Frequency Domain. In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII. Springer-Verlag, Berlin, Heidelberg, 396–413. https://doi.org/10.1007/978-3-031-19778-9_23

[4] Haoti Zhong, Cong Liao, Anna Cinzia Squicciarini, Sencun Zhu, and David Miller. 2020. Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation. In Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy (CODASPY '20). Association for Computing Machinery, New York, NY, USA, 97–108. https://doi.org/10.1145/3374664.3375751

[5] Y. Li, Y. Li, B. Wu, L. Li, R. He and S. Lyu, ”Invisible Backdoor Attack with Sample-Specific Triggers,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 16443-16452, doi: 10.1109/ICCV48922.2021.01615.

[6] S. Li, M. Xue, B. Z. H. Zhao, H. Zhu and X. Zhang, ”Invisible Backdoor Attacks on Deep Neural Networks Via Steganography and Regularization,” in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 5, pp. 2088-2105, 1 Sept.-Oct. 2021, doi: 10.1109/TDSC.2020.3021407.

[7] Chang Yue, Peizhuo Lv, Ruigang Liang, Kai Chen. Invisible Backdoor Attacks Using Data Poisoning in the Frequency Domain. https://doi.org/10.48550/arXiv.2207.04209