A Review of Research on Deep Learning-based Target Detection Technology for Automated Vehicle Driving Systems

Yuze Li

Xi'an Technological University, China

Abstract. In recent years, with the rapid development of artificial intelligence technology, major technology companies around the world have strategically shifted their attention to the field of autonomous driving, momentarily pushing the research on autonomous driving to a climax. Target detection is one of the core technologies in the field of autonomous driving. For this reason, this paper provides a research review on driverless technology, deep learning target detection algorithms, and briefly summarizes the difficulties faced by autonomous driving target detection, and then introduces five common current target detection algorithms.

Keywords: Deep learning; autonomous driving; target detection technology; research.

1. Introduction

With the rapid improvement of the economic level, the car ownership of Chinese residents also grows rapidly accordingly. According to relevant surveys and statistics, in 2021, the number of motor vehicles owned by residents in China has exceeded 400 million. In recent years, the number of motor vehicle drivers in the country has shown a continuous and substantial increase, and the popularity of motor vehicles has brought great convenience to people's life and travel, but the ensuing traffic accidents and the increasingly serious urban traffic congestion have become an important factor limiting the further development of urban economy and people's living standards[1, 2].

According to the statistics of the Ministry of Transportation of China, 211,074 traffic accidents caused by motor vehicles occurred nationwide in 2020, which brought huge casualties and economic losses. In addition to the safety factors of motor vehicles themselves, the majority of accidents are caused by human illegal driving[3].

People's active demand for traffic safety and their desire for intelligent life has driven the rapid development of autonomous driving. In recent years, artificial intelligence technology has developed rapidly, and the traditional automotive industry has made great progress in autonomous driving by closely combining with the information industry[4]. Many large companies at home and abroad have gone forward and backward to join the battlefield of autonomous driving. Technology companies such as Google, Baidu, and Tesla have all made great research progress in the field of autonomous driving respectively. Looking at China, BYD and Baidu have reached an agreement to work together to further explore and promote the development of self-driving cars. It is foreseeable that shortly, with the gradual maturity of the technology, autonomous driving will enter the practical and popular stage, which will greatly reduce the chances of traffic accidents caused by human factors such as unlicensed driving, fatigue driving, and drunk driving.

Target detection is one of the core technologies of autonomous driving, which includes the detection of traffic signals, pedestrians, vehicles, and other targets[5]. Self-driving vehicles rely on computer vision technologies such as target detection to identify targets and perceive the traffic environment to provide judgment conditions for a series of operations such as starting, stopping, steering, and accelerating, so self-driving target detection requires "fast" and "accurate" algorithms. The scenario of autonomous driving is very complex, including different climatic conditions, target occlusion, small target detection, and other difficulties. Traditional target detection algorithms based on image feature extraction rely too much on the feature extraction effect, which is difficult to achieve high detection accuracy and poor generalization performance and cannot adapt to the needs of this scene. In recent years, deep learning[6] has developed rapidly, and Convolutional Neural Networks (CNN)[7] are used to train on a large number of data sets of autonomous driving scenes, and the

resulting model has high detection accuracy and generalization ability and can meet the requirements of real-time detection of autonomous driving scenes when equipped with GPU for acceleration.

2. Status of domestic and international research

2.1 Driverless technology

Driverless cars are divided into different levels from L0 non-driverless to the highest level of L5 fully autonomous driving. Currently, the L4 level of highly autonomous driving is achievable in recent years, and the main scenarios are large parking lots and other structured roads, which are the current joint efforts of major smart car companies[8].

Driverless technology began to be studied abroad in the early 19th century, and the first driverless vehicle was researched in 1925, nearly a century ago. Until the early 21st century, driverless technology stopped. In recent years, computer technology has driven the development of this artificial intelligence, making driverless technology increasingly change. Currently, all major foreign car manufacturers are beginning to research driverless technology.

Waymo, the leader in driverless technology, is the first company to launch a driverless technology rental service, and Daimler, the world's largest commercial vehicle manufacturer, plans to launch driverless technology at the L3 level in 2021.

Tesla is an indispensable representative of driverless technology, and both Model S and Model X models have semi-autonomous driving systems that allow drivers to change lanes automatically by operating the steering wheel differently[9].

Domestic exploration of driverless technology started late, beginning at the end of the 20th century, and has made good progress with the rapid development of domestic technology and the cooperation of car companies, universities, and Internet companies. Baidu, the domestic leader with its driverless platform Apollo (Apollo), developed the self-driving car "Apollo" in 2018 and put it into production. FAW Hongqi brand is equipped with a self-driving parking system in its 2020 production model[10]. The emergence of driverless technology has made tremendous progress up to today, and various current systems to assist drivers are helping drivers to travel more safely and securely. With the research of scholars from all walks of life, the travel method will move step by step towards a true driverless segment.

2.2 Deep learning target detection algorithm

Traditional target detection extracts a limited number of features, which is difficult to use in some modern applications with low accuracy, in addition to training classifiers, which requires training multiple classifiers to complete the detection, which is time-consuming and not very accurate. With the development of technology, the continuous maturity of computer technology, and various fields, convolutional neural networks have become an integral part of the emerging era, driving the development of artificial intelligence. In particular, in 2012, the AlexNet convolutional neural network, which obtained unexpected results, made scholars from all walks of life focus on studying various applications of convolutional neural networks[11].

The emergence of deep learning has brought big changes to target detection, making it no longer necessary to artificially fetch and extract features or train multiple classifiers, but simply to build a neural network that learns to extract features and perform classification on its own. The detection effect is beyond the reach of traditional target detection. Target detection algorithms are divided into two approaches based on candidate regions and regression.

The candidate region-based target detection algorithm is a two-stage detection algorithm that first extracts the region of interest and then performs regression to make the rectangular box closer to the true value box. the R-CNN algorithm uses convolutional neural networks to extract features, which makes the detection significantly better compared to previous detection algorithms due to the diversity of the extracted features. However, there are also significant drawbacks, more than 2000 candidate boxes generated in the algorithm need to go through the convolutional neural network to

extract features, and there are a large number of overlapping regions that are over-computed, resulting in an excessive computational effort[12]. In addition, the narrowing of the candidate region leads to easy deformation of the target, loss of information, and reduced training performance. The Faster R-CNN algorithm in target detection, first extracts the candidate region, and then classifies and locates the candidate region, which has a great advantage in detection accuracy, but the network is complex and has many parameters, leading to slow detection and poor real-time performance. The main regression-based target detections are YOLO[13] and SSD[14-16], and in 2016, Remon et al. first proposed the YOLO algorithm, which divides the input image into a uniformly sized grid, and the target object is at the center point of which grid, then that grid is responsible for detecting that target and regressing the rectangular box. The detection speed is greatly improved due to the reduction of the number of parameters, which speeds up the computation. At the same time, the simplicity of the model leads to weak expressiveness, so the detection accuracy is still not as high as that of Faster R-CNN. Liu proposed the SSD algorithm in the same year, which uses VGG-16 as the skeleton network and also adopts the idea of regression, setting a priori boxes of different sizes at each pixel point of the feature map and regression, so that SSD improves both detection speed and accuracy. Of course, SSD also has disadvantages, and to compensate for the disadvantages, Liu et al. proposed the DSSD algorithm[17], which replaced the network framework from the VGG-16 network to ResNet101 and added a deconvolution structure behind the convolutional layer with a more adequate semantic information representation. In 2017, Shen et al. proposed the DSOD model[18], which replaces the base network with DenseNet, which can achieve good training results with fewer parameter data, which is very different from other algorithmic networks and can be fine-tuned in the trained network. In 2017, Redmon et al. proposed a version of YOLOv2 using Darknet-19 as the base network[19] and also added a transfer layer that can fuse fine-grained features of images. There is an improvement in both accuracy and speed compared to the v1 version.

In 2018, BharatSingh et al. designed the SNIP network to train the network by the scale of the target object, and the transformation of the shape, to compare with each other and learn the impact of detection due to scale changes. The summary of the test results found that the convolutional network has poor detection performance when the scale changes, so the sample test of pyramids was proposed. Under the influence of factors such as target size, which makes network training difficult, image pyramid scale normalization is studied. To solve the small target problem, Redmon et al. proposed YOLOv3[20, 21] in 2018, which has improved the detection performance for small target objects. Deep learning-based target detection is still being improved all the time and has made great progress relative to traditional ones, so there is still unlimited room for future development, and remains a challenging project to replace humans for more precise and accurate tasks.

In 2020, Dong Zhang et al. proposed FPT, an efficient feature interaction method for feature pyramids, containing three well-designed feature enhancement operations that can be freely embedded in various detection algorithms containing feature pyramids with high adaptability. In 2021, the BoTNeT network proposed by Aravind Srinivas et al. incorporates self-attention into a variety of computer vision tasks, including image classification, target detection, and instance segmentation. The method significantly improves the baseline in instance segmentation and target detection, while also reducing parameters to minimize latency. In 2021, LiYuan et al. designed the T2T-ViT network proposing a novel Tokens-to-Token mechanism for simultaneously modeling local structural information and global relevance of images, while also drawing on CNN architectural design ideas to guide the backbone design of ViT. At the same time, the proposed approach achieves better performance than the well-designed MobileNet family of solutions by simply reducing the depth and implicit layer dimensions in terms of lightweighting.

2.3 Difficulties of self-driving target detection

Due to the complexity of targets in the autonomous driving environment, there are many different types of obscured targets or small targets, which makes it necessary to have a high accuracy rate and real-time target detection to ensure the safety of road drivers and pedestrians and the efficiency of

travel. For example, in driverless technology, the algorithm is embedded in the mobile terminal to detect the target accurately and in real-time to ensure the safety of driverless technology. The future research direction is also the research difficulty is how to improve the detection accuracy while reducing the model parameters and locating the target more accurately.

3. Common target detection algorithms

3.1 R-CNN algorithm

The skeleton network of the R-CNN algorithm is the AlexNet network. After receiving an image, the similarity of neighboring image blocks is calculated by the Selective Search algorithm, and then about 2000+ candidate regions are generated, and all these candidate regions are reduced to 227×227, and the target object features are extracted by the convolutional neural network, and finally, the classification is completed by the classifier SVM.

R-CNN algorithm is a more successful target detection algorithm in the early stage of deep learning, using a convolutional neural network to extract features. However, there is also a big drawback that all the 2000+ candidate frames generated in the algorithm need to go through the convolutional neural network to extract features, and there are a large number of overlapping regions that are over-computed, resulting in too much computation. In addition, the narrowing of the candidate region leads to the target is easy to produce deformation, loss of information, and training performance degradation.

3.2 SPP-Net algorithm

To solve the above characteristics of image shrinkage and distortion, the SPP-Net algorithm incorporates the SPP structure in the fully connected layer of the network, regardless of the size of the input image, without cropping, scaling and other operations, the feature vector of fixed size can be sent to the fully connected layer, reducing the amount of computation.

Since the training process is multiple stages, the data during the process needs to be saved and the time becomes more. And when it comes to fine-tuning, only the fully connected layer behind is updated, and the convolutional layer cannot continue the training and cannot be back-propagated.

3.3 Faster R-CNN algorithm

The Regin of Interest Pooling Layer (ROI layer) structure is designed in the Faster R-CNN, which samples different sizes of feature maps into features of the same size. In the algorithm flow, all the features of the whole image are taken first, then feature mapping is performed, and the feature maps are converted to the same size in the ROIlayer. In addition, the Fast R-CNN algorithm also proposes a multi-task loss function, which performs both classification and regression tasks on the fully connected layer, and constructs a multi-task model to solve the problem of repeated computation of features, which greatly reduces in training and detection time.

The Fast R-CNN algorithm still gets the candidate regions by the sliding window method, the computation volume is still relatively large and the computation time is longer. The Faster R-CNN algorithm adds the RPN network to the backbone network, and each pixel point will generate an anchor box of different sizes, and then the feature vector is fed into the classifier to determine the probability of whether the box contains the target object, while the border regression network gets the coordinate position of the anchor box and fine-tunes the anchor box to get a more accurate candidate box. The algorithm completes the roles of candidate region extraction step, classification, and regression in the convolutional network, and achieves end-to-end training. And these steps can share the convolutional network, which makes the training model parameters reduced and the speed has been greatly improved.

3.4 YOLO algorithm

3.4.1 YOLOv1 algorithm

Using the idea of regression to solve the target detection speed fps 45, the YOLOv1 algorithm divides the input regiment image into a 7 × 7 grid, the center of the target object in the grid, then the grid is responsible for detecting the object. yOLOv1 algorithm sets each grid to detect B target borders, and each border generates five prediction values: x, y, w, h, and confidence. Confidence is the information about the magnitude of the probability of the target contained in the border and the accuracy of the prediction. The confidence is defined as Pr(object) × IOU. a Pr(object) of 0 means that there is only a background; a Pr(object) of 1 means that there is a target.

The YOLOv1 algorithm is about ten times faster than the Faster R-CNN algorithm, and the detection speed is substantially improved. It meets the real-time requirements for most purposes, and the network structure is simple, but the drawback is also obvious that an object intelligence is predicted by a grid.

3.4.2 YOLOv2 algorithm

Since YOLOv1 target localization is prone to errors, the YOLOv2 algorithm, adds a BN layer after the convolution layer to normalize the data and the input image size is 448 × 448.

The YOLOv2 algorithm uses post-Darknet-19 as the skeleton feature extraction network. The YOLOv2 algorithm also borrows from the anchor mechanism in Faster R-CNN to obtain a more accurate regression bounding box by initially setting a priori boxes to predict the location information on the grid with four parameters {x, y, w, h, which denote the border centroid location and width and height, respectively. c_x, c_y denote the target offset relative to the upper left corner of the image, respectively; p_w, p_h denotes the a priori box height and width, respectively; b_x, b_y, b_h, b_w denote the corrected bounding box centroid and width and height, respectively. The x, y are normalized by the Sigmoid function.

The YOLOv2 algorithm also adds a transfer layer (Passthrough Layer) to fuse fine-grained features of the image, combining deep and shallow features, which improves both accuracy and speed compared to YOLOv1.

3.4.3 YOLOv3 algorithm

YOLOv3 algorithm uses an excitation function of Leaky ReLU to avoid gradient disappearance and overfitting phenomenon.

YOLOv3 differs from the previous version by using the biggest improvement of having Darknet-53 as the skeleton feature extraction network, which has a residual nature. In addition, for better detection, a unit convolution kernel is used in the network to varying the feature image pixel size, making the information extraction more adequate. the YOLOv3 algorithm adds a feature pyramid structure to the network by adding rectangular boxes of different sizes to the top for feature maps of different sizes and then makes progress in detection accuracy by predicting targets of different sizes for all rectangular boxes. In addition to this, multi-label classification is also used. When predicting the category, the output of logistic is used for prediction, and the category loss is calculated using binary cross-entropy loss. The classifier only determines whether the target object in the border belongs to the current label or not, and only a simple binary classification task is needed to complete the classification of multi-label objects.

References

[1] Zhao L. Inventory of China's auto market in 2018[J]. Automotive Vertical, 2019, 95(02): 35-37.

[2] Hazelhoff L, Creusen I M, De With P H N. Exploiting street-level panoramic images for large-scale automated surveying of traffic signs[J]. Machine Vision & Applications, 2014, 25(7): 1893-1911.

[3] Analysis of driver factors before major accidents in key vehicles[J]. China Public Safety, 2019, 319(Z1): 72-74.

[4] Bengler K, Dietmayer K, Farber B, et al. Three Decades of Driver Assistance Systems: Review and Future Perspectives[J]. IEEE Intelligent Transportation Systems Magazine, 2014, 6(4): 6-22.

[5] Meng, X.R. Research on target detection and target tracking methods[J]. Information and Computer, 2019(9):33.34.

[6] Lecun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553): 436-444.

[7] Haykin S. Neural networks: a comprehensive foundation[M]. Prentice Hall PTR, 1994.

[8] Liu W. Analysis of intelligent vehicle assisted driving technology[J]. Practical automotive technology, 2021, 46(02):35-37.

[9] Fu Yufei, Zhu Cing. With automatic driving, car companies open the "head goose fight"[J]. School of Business, 2019, 2(06):75-77.

[10] Li Yao-chen. Research on the development strategy of driverless cars of FAW Group[D]. Jilin University, 2019, 8(03):39-42.

[11] DICKEY L, SHAW L . Proceedings of the IEEE International Conference on Computer Vision[J]. Elsevier ence B.v.amsterdam, 2014, 37(90):v-xiv, 1-387.

[12] ZUO Z, YU K, ZHOU Q, et al. Traffic Signs Detection Based on Faster R-CNN[C]// 2017 IEEE

[13] REDMON J, DIVVALA S K, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. arXiv: Computer Vision and Pattern Recognition, 2016: 2(08)779-788.

[14] XIANG W, ZHANG D, ATHITSOS V, et al. Context-Aware Single-Shot Detector[J]. arXiv: Computer Vision and Pattern Recognition, 2017, 32(4):325-329.

[15] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]. arXiv: European Conference On Computer Vision, 2016: 3(3)21-37.

[16] YAN Zhao, L. Wu, Q. Chen. Multi-scale fusion SSD-based algorithm for small target detection[J]. Computer Engineering, 2020, 46(01):247-254.

[17] FU C, LIU W, RANGA A, et al. DSSD: Deconvolutional Single Shot Detector[J]. arXiv: Computer Vision and Pattern Recognition, 2017, 42(2):437-458.

[18] HUANG G, LIU Z, MAATEN L, et al. Densely Connected Convolutional Networks[C]// CVPR. IEEE Computer Society, 2017 6(9):463-467.

[19] ARGEN Z, LIU Yun. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2019.

[20] LIN Hao-Chi. Improved multi-target detection algorithm for unmanned systems with YOLOv3 network[J]. Industrial Control Computer, 2020, 33(05):77-79.

[21] Tan F, Mu P, Ma Z X. A multi-target tracking algorithm based on YOLOv3 detection and feature point matching[J]. Journal of Metrology, 2021, 42(02):157-162.