Object detection: review and improvement based on deep learning

Tingyu Wang

Tianjin University, Tianjin, China, 300354

tingyuwangx@gmail.com

Abstract. In recent years, with the quantum leap in deep learning, self-driving vehicle as one of its applications has been gaining tremendously increasing popularity as well as making a multitude of achievements. Object detection, which has made significant contribution to driver-less vehicle, has had been applied to a tremendously wide range of fields. However, reports relevant to automatic vehicle stating that accidents are caused by Automatic driving technology, present problems pointing out that existing target detection algorithms, which are already fairly well reliable, can probably be interfered by adverse conditions such as high temperature, raise dust and transmission loss, and be not capable of providing precise output. This paper recaps on these previous classic algorithms, and their large-scale application domain. Meanwhile, this paper presents improvements focusing on enhancing the robustness of these algorithms to overcome these problems caused by adverse conditions and improve the accuracy. Thus these improvements could augment the security of these driver-less vehicles, and eventually reduce traffic accident mortality relative to self-driving vehicles and safeguard road safety, and may potentially benefit to further research.

Keywords: target detection, deep learning, computer vision.

1. Introduction

Although machine learning has marvelous achievements in image recognition, decision making and Natural Language Process over last decades, technology of self-driving seems out of that tendency. The reason why this progress is slow can probably attribute to approaches that require much hand-engineering, over-reliance on road testing and high fleet deployment costs [1], however the accuracy of the system in identifying objects probably matters yet.

Object detection, which is aiming at detecting the existence of targets and locating the fixed position of them, is an important branch of computer vision. Object detection results have been improved rapidly through last decades. To a large extent, it is powerful baseline systems that promote these [2]. The extraodinarily challenging research field is capable of providing valuable information of semantic understanding for image processing, and intimately relative to a huge number of application. Traditional target detection were mainly manual designed, nowadays however, target detection algorithms based on deep learning take advantage of Convolutional Neural Networks. This feature learning method can automatically figure out the features needed for detecting and classifying targets, transform the original input information into higher dimensional and more abstract features, which enhance its performance under complex conditions to meet most of the needs of industry. Target detection based on deep learning could be greatly helpful for these areas with high requirement of both accuracy and promptness, such as self-driving vehicle, face recognition and probably medical image processing.

Generally, for the field of computer vision, the following cases often occur: 1) The brightness of image sensor is uneven; 2) Noise from the environment and circuit components mutually influence each other; 3) The high temperature of image sensor after working for long time. These all threaten to the noise on images collected. Focusing on these probable factors, based on deep learning and evaluate from Facebook detectron2, this study utilizes Gaussian noise to simulate the real condition and develop a method improving on existing Object detection algorithm to weaken the influence of these adverse condition, especially increases the accuracy of pedestrian recognition hence improve the safety of the transportation system.

2. Deep learning

In the past few years, deep learning and deep neural networks have been very popular in reinforcement learning and the fields of games, robots, natural language processing and so on. We have witnessed some breakthroughs, such as deep Q network (dqn), alphago and deepstack; Each delegates a series of problems and a swarm of applications. Dqn (applicable to single player games and general single agent control. Dqn has influenced most algorithms and applications in the area of reinforcement learning. Dqn ignited the popularity of this round of deep reinforcement learning. Alphago is applicable for two persons perfect information zero sum games. It has made amazing achievements on a very difficult problem and set a milestone in the field of artificial intelligence. The success of alphago directly affects similar games, and alpha zero has achieved great success in chess and Shogi. The basic technologies of alphago and alphago zero, namely deep learning, reinforcement learning, Monte Carlo tree search (MCTS) and self game, will have broader and deeper meanings and applications. As recommended by the author of alphago in his paper, the following applications deserve further research: General Games (especially video games), classical planning, partial observation planning, scheduling, constraint satisfaction, robots, industrial control, online recommendation system, protein folding, reducing energy consumption, and finding revolutionary new materials. Deepstack aims at the zero sum game between two people with incomplete information, which is a series of difficult problems to solve. Similar as alphago, deepstack has also made extraordinary achievements in a difficult problem and set a milestone in the field of artificial intelligence. It will have sufficient meanings and a range of applications, for example, in defending the sound decision-making of strategic resources and medical advice.

Deep learning is an emerging field in machine learning. Its basic motivation is establishing and simulating the neural network of human brain when analytical learning. It simulates the human brain's mechanism of interpreting data such as images, sounds and texts. Deep learning is a sort of unsupervised learning. The concept of deep learning comes from artificial neural network. Specifically, the multi-layer perceptron with multiple hidden layers is a deep learning structure. Deep learning combines low-level features to form more abstract high-level representations, and distributed feature representations of attribute categories or feature discovery data [3].

Metaphorically, the relationship between the information and data processed by the deep learning network is similar to the relationship between water flow and pipeline valves. Information and data are input from the input layer of the deep learning network and output from the output layer after passing through the deep network, just like water flows in from the inlet of a water pipe and then flows out from the outlet of a pipe. The pipeline system is generally composed of complex multi-layer control valves, which control the flow and direction of water flow. And adjust the state and number of valves according to the needs of the task. The water pipe connects the regulating valves of different layers in the water pipe network. It forms a water flow system connecting the front and the back and between layers, that is, a deep learning network [4]. Compared with traditional neural networks, deep learning uses a similar neural structure, but based on the rapid growth of computing power and open source data sets, more layers of deep learning networks have been strongly supported.

3. Mainstream Classic Previous methods in target detection

Most preceding algorithms are mainly focusing on constructing artificial features and for classification, and many outstanding works have emerged, whereas the existing problem is that the characteristics of artificial design may not have strong applicability and generalization ability. One kind of characteristics may be merely heuristic for certain kind of problems, and the influence on other problems is very limited.

3.1. R-CNN and fast R-CNN

R-CNN is based on such a very simple idea. For the input image, through the selective search method, this study first determines, for example, 2000 windows that are most likely to contain objects. For these 2000 windows, we hope it can achieve a very high recall rate for the detected objects. For these 2000 windows, we hope it can achieve a very high recall rate for the detected objects. Then, demonstrated in Figure 1, CNN will be used to extract and classify features in each of the 2000.[4]. Running CNN once for these 2000 regions will limit the speed to a low level. Even if the time cost is merely 0.5 seconds each, it takes 1000 seconds for all 2000 windows. In order to speed up, in 2014, he Kaiming proposed spp net. What is novel is to run CNN once for the whole picture instead of each window alone. However, there is a small barrier that the size of each window may be different , for resolving this problem, spp net designs spatial pyramid pooling. So that different sizes of small windows have the characteristics of the same dimension. Although This method evades calculating convolution for each candidate window, it is not fast yet.

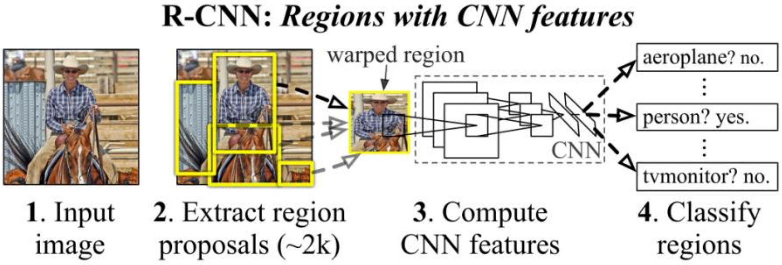

Figure 1. The flow of R-CNN [5].

As for the Fast R-CNN, Figure 2 shows some details: Roipooling is implemented in the network so that the input image block does not need to be cut to a unified size, so as to avoid the loss of input information. Secondly, the whole graph is input into the network to obtain the feature graph, and then the target frame obtained by selective search algorithm on the original graph is mapped to the feature graph to avoid repeated feature extraction. The Fast R-CNN network looks upon the unmodified picture and a set of object coordinates as input. The network first uses several convolution layers and maximum pool layers to process the whole image to generate convolution feature map. After which, for each object, a fixed length feature vector will be extracted from the feature map by the region of interest (ROI) pool layer [6].

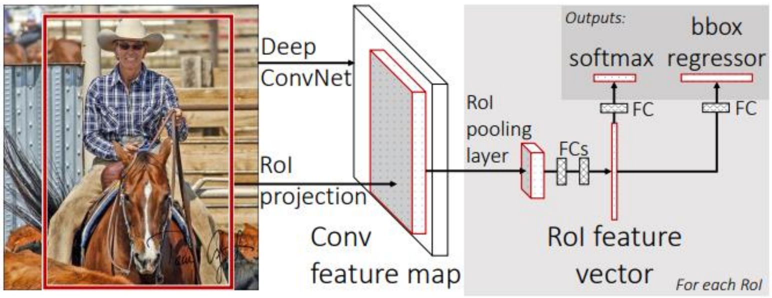

Figure 2. The flow of Fast R-CNN [6].

3.2. SSD

SSD (The Single Shot Detector) is that firstly input an image, then let the image extract features through convolutional neural network (CNN) for generating a feature map. Extract the feature map of six layers, then generate a default box on each point of the feature map (different layers, but each point has), collect all default boxes have been generated, and throw them into NMS (maximum suppression), got a output and filtered default box. One input image and ground truth boxes is all SSD needs [7].

The specific implementation details are shown in Figure 3.

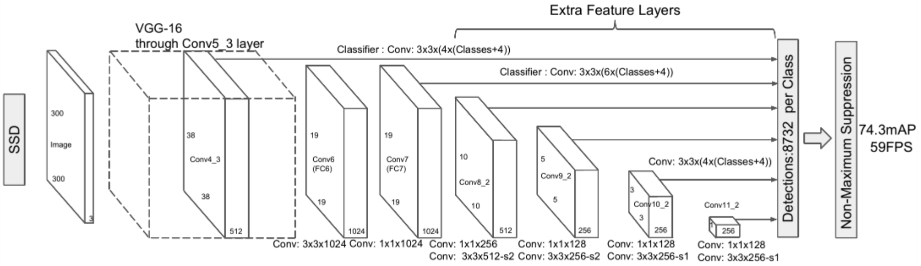

Figure 3. The flow of SSD [7].

3.3. YOLO

There is a tremendously pioneering method, which is named YOLO or You Only Look it Once being presented in 2015. This is a somehow strange method. For a given input image, it will go through a process which is demonstrated in Figure 4: Yolo always divides the grid into 7x7, that is, 49 windows, and then predicts two rectangular boxes in each window. This prediction is completed through the full connection layer. Yolo will predict the four parameters of each rectangular box, the reliability of its contained objects, and the probability that it belongs to each object category. Yolo is very fast and can successfully reach 45fps on GPU [8].

The processing steps of Yolo are: zoom the input picture to 448 × 448 size; run convolution network; the model confidence card threshold is used to obtain the target location and category. For VOC data sets, the opinion of Yolo is to uniformly adjust the pictures to a scale in 448 × 448, divide each figure into 49 average grids, for each grid, there are 2 rectangular boxes and their confidence, with 20 different kinds of probabilities [9]. Abandoning the region proposal stage and the speed is accelerated, but the accuracy of position is relatively lower. Whilst, the question is that the classification accuracy is also relatively low. The performance on various data sets is averagely and approximately 54.5% map.

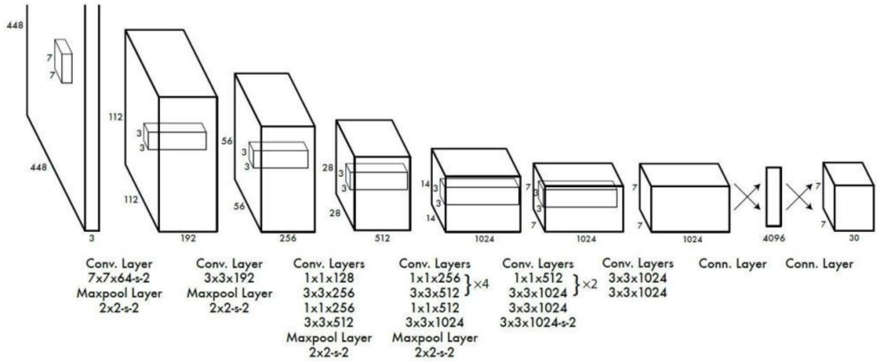

Figure 4. The structure of YOLO [9].

The improvements of target detection algorithms in self-driving vehicles.

The target detection is one of these most significant phases in self-driving technology, since it is the stage that the vehicles detect the objects on the road. Following recognition and policy decision are based on this step.

Specifically, using different basic models and configurations to evaluate detectron2 can achieve faster r-cnn. At the same time,according to the track "global road damage detection challenge 2020" in IEEE big data 2020 big data Cup Challenge dataset, which has test these methods credibly.It is demonstrated that the x101-fpn basic model of faster r-cnn with the default configuration of detectron2 is effective and universal sufficient to transfer to distinctive countries in this challenge. This method showed a result of F1 scores of 51.0% and 51.4% in challenge test 1 and test 2, respectively. Though the visualization performs well in predicting, the F1 score is low [10].

As it has been mentioned previously, the adverse conditions could have abominable influence on the accuracy of object detection on these images transmitted. This study adds layers and adjust the hyperparameter in classic detectron2 to process and denoise the image, so the accuracy of target detection is increased. The method is based on an existing framework, namely detectron2, this paper throws a new and slight change on and reach a better result. Our change is mainly focus on the part the the model associated with the input data. There is already a data enhancement system in detecrtron2, namely DatasetMapper. This study changes the RandomBrightness class, augment the brightness of Pixel point with large pixel value to make the edge of image more sharp. And to adapt to the variety, the granularity of the enhanced RGB channel light intensity is adjusted. In the Imagenet training set, principal component analysis (PCA) is used to obtain the eigenvector group and eigenvalue group of the RGB covariance matrix. The eigenvalue group multiplies the normal distribution random value with the mean value of 0 and the given standard deviation element by element, multiplies the eigenvector group and the eigenvalue group vector to obtain the RGB increment, and adds the increment to the image pixel by pixel. So that the target feature is not affected by light intensity and color change.

Because the main purpose is to show the improvement in noisy data, and with limited computing power, this paper chooses to identify single target. One category and 255 instances.

Specifically, this study firstly trains and tests the existing model detectron 2 with dataset COCO2017, reaches a similar result with the result of the developer. And then use the same dataset to train our updated neural network, taking down the results and then use Gaussian noise to contaminate images and test it again. Finally compare the outcome of each situations, as shown in Table 1 and Table 2.

Table 1. The evaluation result for bbox.

	AP	AP50	AP75	APS	APm	APl
classic model	71.6	83.7	83.7	0.00	53.2	87.8
Classic model (noisy simulation)	64.9	79.9	79.9	0.00	44.8	82.0
Our model	71.7	80.2	80.2	0.00	48.0	90.6
Our model (noisy simulation)	67.1	80.9	80.1	16.8	46.5	83.9

Note: original by the author

Table 2. The evaluation result for segm.

	AP	AP50	AP75	APS	APm	APl
classic model	76.0	80.8	80.8	0.00	52.04	96.36
Classic model (noisy simulation)	73.9	79.9	79.9	0.00	47.4	94.8
Our model	75.3	80.2	80.2	0.00	49.7	96.2
Our model (noisy simulation)	74.4	81.0	81.0	23.56	49.0	93.9

Note: original by the author

4. Conclusion

This paper summarizes the research progress of deep learning in recent years and outlines the general situation, reviews the development process of deep learning, and summarizes its basic principles, including the basic principles of deep learning and target detection. This paper analyzes the current mainstream target detection algorithm based on deep learning, and discusses some of its current cutting-edge and mainstream research. However, due to the limitation of Computing power and equipment as well as the restrictions on data sets, this research does not make it to reach an apparently effective outcome. At present, driverless technology has attracted more and more attention, and has made a lot of research results. In the future, further research will be done on the learning efficiency, running speed, generalization performance and other aspects in target detection.

References

[1] Jain A, Del Pero L, Grimmett H, and Ondruska P, Autonomy 2.0: Why is self-driving always 5 years away? 2021. DOI:10.48550/ARXIV.2107.08142.

[2] He K. M, Gkioxari G, Dollár P, and Girshick R, Mask R-CNN. IEEE International Conference on Computer Vision (ICCV), Venice, 22-29 October 2017, pp.386-397.

[3] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin Ma, Ghemawat S, Irving G, Isard M, et al. Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} Symposium on operating systems design and implementation ({OSDI} 16), 2016; pp. 265–283.

[4] KaiFu Lee, Yonggang Wang, Artificial Intelligence [M] Cultural Development Press, 2017.

[5] Girshick R, Donahue J, Darrell T, and Malik J, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587, Doi: 10.1109/CVPR.2014.81.

[6] Girshick R, “Fast R-CNN”. in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp.1440-1448.

[7] Liu W, et al. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M. (eds) Computer Vision -2016 European Conference on Computer Vision (ECCV), Amsterdam, 2016, pp.21-37.

[8] Ge Z, Liu S, Wang F, Li Z, and Sun J, YOLOX: Exceeding YOLO Series in 2021, arXiv preprint arXiv:2017.08430, 2021.

[9] Redmon J, Divvala S, Girshick R. and Farhadi A, You only look once: Unified real-time objectd etection, Proc. CVPR, 2016, pp. 779-788.

[10] Pham V, and Dang T, Road Damage Detection and Classification with Detectron2 and Faster R-CNN, 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 5592-5601.