Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230556

Comparing one-stage and two-stage learning strategy in object detection

Hanqing Bi[1, 4, †], Vincent Wen[2, †], Zhenyu Xu[3, †]

[1]Tianjin No.21 High School, Tianjin, 300400, China
[2]The Affiliated High School of Fuzhou Institute of Education, Fuzhou, Fujian, 350108 China
[3]International Department, The Affiliated High School of SCNU, Guangzhou, Guangdong, 510000, China

[4]g04@xsy.edu.pl
[†]These authors contributed equally.

Abstract. Object detection plays a vital role in computer social perception and computer vision. It could be applied to computer navigation, video monitoring, industrial detection, and so on. It greatly reduces the human labours by automatically locate and identify objects. Nowadays, the mainstream methods of object detection could be separated into the one- and two-stage method. The one-stage method leverages Convolutional Neural Network (CNN) for obtaining features and directly locate the target objects and their corresponding category probabilities. Different from the two-stage solutions, its accuracy is lower and the recognition speed is higher. The two-stage method is a straight forward solution, which process is mainly completed through a complete CNN, so CNN features will be leveraged to extract the feature description of the target among candidate regions through a CNN. The accuracy of the two-step method has been greatly improved, but the running speed is much slower than the one-step method. While the one-step method is less accurate, it is much faster. In this work, representative works for object detection are conducted and compared. The results could further demonstrate their respective advantages.

Keywords: Object Detection, Deep Learning, One-stage Learning, Two-stage Learning.

1. Introduction

Object detection, also known as object extraction, could simultaneously locate and classify target objects. It combines the recognition and localization of the object into one. It is accurate and almost real-time which distinguishes the CNN solutions. Especially in complex scenarios, where multiple objects with different shape and size require processing in real-time.

With the development of computer vision, artificial intelligence (AI) based real-time tracking achieves more attention, including the intelligent monitoring system, the detection of the military target, and medical navigation surgery surgical instruments positioning has extensive application value.

Traditional identification methods laborious and expensive and not necessarily with good accuracy. In recent years, deep learning-based target recognition has become a popular research method. This method can lower the consumption of manual recognition and improve accuracy. This topic is very hot in academic circles. Detection has become one of the important branches. This task also has many

DOI: 10.54254/2755-2721/5/20230556

application scenarios in life, such as unmanned driving, monitoring, recognition, and so on. The most famous intelligent object detection code YOLO [1] is widely used to identify vehicles, people, and objects. Now it is widely used in unmanned driving systems. In addition, there are many fields where deep learning can be used for object detection. Security field includes fingerprint recognition, face recognition, etc.; military field includes terrain investigation, flying object identification, etc.; traffic field includes license plate identification, driverless driving, traffic sign identification, etc.; medical field includes electrocardiogram, B-ultrasound, health management, nutrition, etc.; life field includes smart home, shopping, intelligent skin measurement, etc. This paper aims at studying and analyzing the difference between one-stage and two-stage target recognition schemes.

In one-stage object detection, the key strategy is to extract features directly by the convolutional neural network. These features could further be used classify and locate objects. While for the two-stage one, the model region proposals are extracted first, and then the classification and localization of the target are predicted by a convolutional neural network. The two strategies have their respective advantages and disadvantages. The one-stage method is very fast and not easily distracted by the background and learned most of the basic features. However, its accuracy is generally difficult to have a good detection effect for most small objects. As for the two-stage one, it has high accuracy to identify most objects and is not easy to make mistakes. However, it requires a large dataset and takes a long time for identifying. In the following paragraphs, these two strategies are separately introduced and quantitatively compared.

2. One stage method

2.1. Introduction

One-stage target detection has the same series of calculation methods, and the detection accuracy continues to improve, even exceeding the accuracy of the two-stage target one. Usually, a deeper backbone network could lead to high detection accuracy, but the speed is lowered. In one-stage framework, small targets are difficult to locate, which can be alleviated through the fusion of different layers of characteristics to achieve multi-scale and high-precision prediction results. It is of great importance in practical, since the detection speed could satisfy the need of real-time application, which can realize rapid deployment on the end side.

One stage method has many advantages such as its fast speed and capability of avoiding background errors by generating false positives. However, it has low accuracy and is difficult to accurately detect small objects.

2.2. Representative works

2.2.1. YOLO family. There are many improvements based on the YOLO models. The YOLOv1 [1] removes candidate area operations and detects targets through grid regression. It is a pioneer work using regressive analysis to achieve object detection tasks. YOLOv2 [2] introduces the batch normalization operation, where the training becomes more stable. Moreover, the introduced feature fusion strategy greatly boosts performance. YOLOv3 [3] introduces residual operation, multi-scale prediction, and cross-scale feature fusion strategies for learning. By enlarging the receptive field, the result takes more context information into consideration and further increases the performance. YOLOv4 [4] leverage the spatial pyramid and a series of tunning skills. YOLOX [5] reduces the calculation amount of the model, alleviates the imbalance problem, accelerate the convergence time of the model, and obtains the optimal sample matching scheme.

2.2.2. RetinaNet. [6] The algorithm combines ResNet and FPN for feature extraction to achieve image multi-scale featurs, and then realize the classification and return tasks through two FCN networks. At the same time, the focal loss could make the model sensitive to small objects. By adjusting the weights

DOI: 10.54254/2755-2721/5/20230556

among difficult samples, the model focuses more on the classification of sparse and difficult samples during training.

2.2.3. CornerNet. [7] For the RCNN series, YOLO series and SSD series, etc. calculation method, we proposed a problem of the law, and the law, Put forward the CornerNet algorithm. CornerNet adopts the prediction corner points versus boundary frame forward line localization, alternative traditional anchor, and area suggestion detection method, avoiding the above-mentioned Anchor leading problem. CornerNet Model Structure with Hourglass as the main trunk network, respectively for prediction of the angle points, simultaneous pair of shift progress prediction, fine adjustment angle point position generation, and a more tight boundary frame. Use predictive heatmaps, embeddings and deflection, and the final boundary frame. At the same time, CornerNet also adopted corner pooling, and the model is more accurate in positioning.

2.2.4. Center Net. [8] For Corner Net's detection strategy based on double-angle point reaching, the calculation is complex and the detection speed is reduced. Therefore, Duan and others proposed the CenterNet algorithm. When Center Net builds the model, it regards the target as a point, that is, the center point of the boundary box. Compared with CornerNet, there is no need to group the corner points, and there is no post-processing such as NMS, which improves the detection speed. CenterNet obtains key points through local peaks on the graph, then predicts the center point through the key points, returns the target-related attributes, and finally realizes the target detection. Compared with the traditional detection algorithm, multi-eigenvalue graph anchor operation, CenterNet outputs through the use of high-resolution feature graphs to omit the above operations and improve the efficiency of the model.

2.2.5. EfficientDet. [9] It is a regression-based detection algorithm, Tan et al. proposed the EfficientDet algorithm. EfficientDet contains 7 structures, namely EfficientDet D1~D7. The speed gradually slows down, but the accuracy gradually improves. Inspired by the PANet algorithm. The pyramid network BiFPN is a feature network. In the same backbone network EfficientNet, BiFPN is 4 percentage points higher than FPN and has fewer parameters. EfficientNet considers the three elements of network width, depth, and resolution when designing the model. The algorithm leverages EfficientNet as backbone, while EfficientNet B0~B6 can control the size of the backbone; the channel numbers and repeat layers of BiFPN Numbers can also be controlled; input the resolution of the graph, which constitutes the structure of EfficientDet.

3. Two stage method

Two-stage method is a straightforward solution towards object detectiom. It uses CNN features to complete the target detection process through a complete convolutional neural network. Typical representatives are: R-CNN [10] and Faster R-CNN [11]. Figure 1 shows the basic data flow of two stage method.

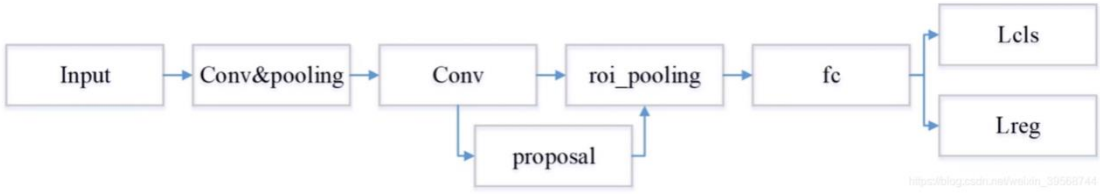

Figure 1. The basic flow of two-stage method.

There are many advantages. The two-stage method is more accurate. Because training the process of the whole network needs two steps. It train the RPN network, and then train the most critical target area localization model. It does not require additional training in classifiers and the process of feature representation. Hence, the two-stage method can be more precise. The two-stage method reduces the

DOI: 10.54254/2755-2721/5/20230556

computational complexity. Fast R-CNN, one of the typical representatives. Besides, it simplifies the SPP layer to ROI Pooling, which is added to the convolution network. ROI Pooling divides each area into smaller pieces, and each piece will get the maximum area. In addition, Fast R-CNN imports the SoftMax function, which replaces SVM processing classification. Therefore, the complexity of the calculation will be decreased.

However, there still exist some disadvantages. Firstly, ROI Pooling is not perfect. The size of the image will be downscaled according to integer multiples if the network is solely downsampled using the pooling layer, but using convolution may result in damage. To put it another way, if the convolution technique without padding is utilized, the original feature map will also be decreased by several pixels, which typically relies on the size of the convolution kernel. Currently, there might be some variations when ROI Pooling is used for matting. When applied to small feature maps, these deviations are relatively small, but when they are returned to their former positions, they will be very far away. As a result, a better ROI Pooling approach is required to increase the network's accuracy. Secondly, features obtained using the two-step method are not efficiently utilized, such as the cascade architecture of Cascade-CNN and the feature pyramid of Libra R-CNN. The model's complexity always rises, because of the current improvement techniques. It is challenging to achieve the real-time goal because, despite ongoing improvements in detection accuracy, training and detection speed are becoming slower. As a result, the target detection method with effective one-stage methods starts to compete with the present development trend of object detection.

In general, One Stage has developed rapidly in recent years, with good accuracy and fast time. If you're after recognition time, one stage is a good choice.

4. Result

4.1. Evaluation matrix

In order to make a multi-directional comparison of different target detection algorithms. A series of recognized evaluation indicators, such as intersection ratio, detection rate, precision rate, recall rate, average precision rate, mean precision rate, etc. Intersection-over -union, refers to the overlap rate of the prediction and ground truth, that is, the ratio between the intersection and union. This index is used to measure the localization accuracy.

Precision means the proportion of positive cases correctly identified as positive cases and represents the proportion of the target detected by the mode as the real target object. Callback rate (recall) refers to the number of positive examples correctly identified as positive examples among all positive examples tested, and represents the proportion of all real targets detected by the model. Precision and recall are aimed at a certain category of indicators in a single picture. At the same time, there are contradictions between precision and some cases. For example, if only one result is detected and accurate, then the precision is 100%, but recall is very low. Therefore, the average precision (AP) is proposed, which is expressed as the integral of the precision rate on the precision rate-call rate curve to the recall rate. AP is for a certain category in the data set and is used to evaluate the detection goodness of the detection algorithm on a category. Mean Average Precision (mAP) refers to the mean of AP values in all categories. MAP is an evaluation of the entire data set and is used to evaluate the detection effect of the detection algorithm on all categories.

4.2. Comparison result

In this section, the performances of representative one stag and two stage methods are compared and displayed in Figure 2. All these methods are implemented on the ImageNet dataset.

DOI: 10.54254/2755-2721/5/20230556

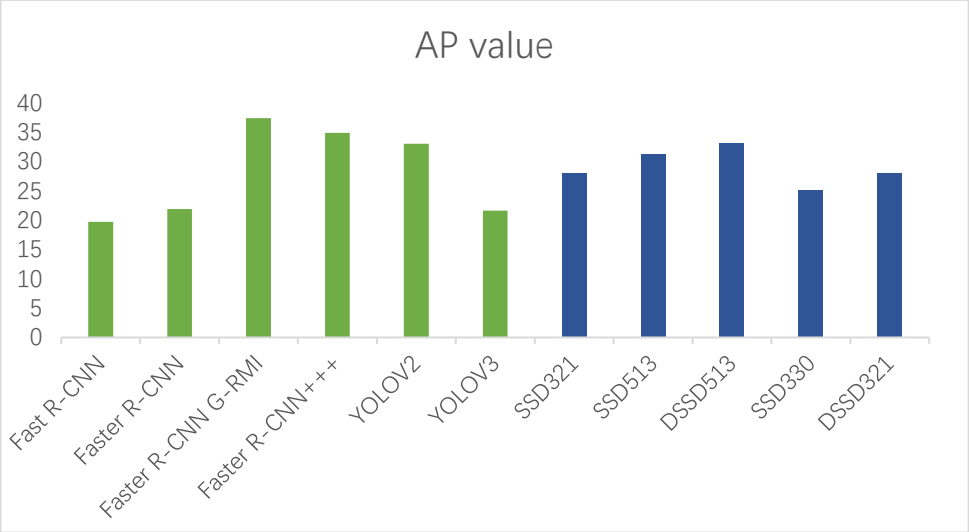

Figure 2. Comparison results of two-stage (green) and one-stage (blue).

For the Two Stage approach, it has many advantages. For example, the two-stage method is done in two steps. The features of CNN convolutional neural network are extracted mainly through convolutional neural network. It first train the RPN network, and then the target detection network. Therefore, the advantage of two stages is very accurate, especially the comparison with one stage. But the disadvantage is also obvious, the speed is relatively slow.

Two-stage method reduces the computational complexity. Fast R-CNN, one of the typical representatives. Besides, it simplifies the SPP layer to ROI Pooling, which is added to the convolution network. ROI Pooling divides each area into smaller pieces, and each piece will get the maximum of the area. In addition, Fast R-CNN imports SoftMax function, and replaces SVM processing classification. Therefore, the complexity of calculation will be decreased. All in all, the advantages of two-stages are obvious. It is difficult to recognize multiple objects in complex scene. Or when the recognition accuracy is very high, the Two Stage method is used. However, because of the need to use RPN network learning and network training. Therefore, it takes a long time. It is not recommended to use two-stage if you have time requirements

In recent years, the one-stage recognition efficiency has been significantly boosted. The recognition efficiency of some one-stage methods is even not inferior to that of some two-stage methods. The typical representative of one stage is the familiar YOLO 1-5 series. Compared with two stages, the advantage is that only one step is needed, that is, the length, width, height and thickness of the object can be identified by learning the CNN and bringing it into the regression. It saves a lot of time. However, small objects are difficult to detect, and this problem can be alleviated by integrating different levels of features, such as dividing more detailed grids to identify small objects.

5. Discussion

The one-stage algorithm combines the candidate region generation, integrates feature extraction, target classification and position regression into one stage. Only need to feed into the network once can predict all the bounding box, so the speed is fast, very suitable for mobile terminal.

The two-stage target algorithm has two stages: selecting the candidate region of the image, and carrying out regression localization and classification of the possible targets in the candidate region. Although the accuracy of the two-stage target detection algorithm is relatively high, the deepening of the algorithm model brings higher computational complexity, which leads to the decrease of the detection rate of the algorithm, and there are still difficulties for real-time applications.

DOI: 10.54254/2755-2721/5/20230556

In the future, although object detection has made remarkable achievements in the past 20 years and has been widely used, there are still many difficult problems. The future research directions of object detection are prospected in the following aspects.

Firstly, weakly supervised target detection problem. At present, the training of mainstream object detection algorithms relies heavily on a large amount of manual annotation data, and the annotation process is time-consuming, inefficient, and expensive, which seriously hinders the development and application of object detection algorithms. The Weakly Supervised Object Detection method (WSOD) is designed to solve this problem by training a detector that contains only image-level annotations but not bounding box annotations. However, weakly supervised target detection faces challenges brought by training under uncertain conditions, including inaccurate training labels, background noise interference, limited diversity of training samples, insufficient training samples, etc. To solve these problems, the ability of weakly supervised learning can be enhanced by embedding useful prior knowledge in the process of weakly supervised learning, or by carrying out reinforcement training in stages.

Secondly, small object detection problem. In large scenario, the small target may be only a few pixels, of the target of this type, how to improve the detection capability is a huge challenge, this includes the potential application of automatic driving distance of pedestrians and nearby of small target detection, medical subtle symptoms, including early tumor detection, using no man-machine for important military targets detection, etc. As more and more complex systems are deployed in the real world, the detection and segmentation of small targets is also a research focus. Small targets can be detected by increasing the resolution of the input image, merging high-resolution features and high-dimensional features in low-resolution images, or oversampling images containing small targets. In recent years, FPN has been used to predict the results of multi-scale feature fusion, which is useful for small object detection. ALFNet adopted a progressive positioning fitting module for pedestrian detection, adjusted the threshold of IoU in a recursive way, and gradually trained multiple positioning modules to improve the positioning accuracy of pedestrians. In addition, low-level fine granularity is used in the perceptive generative adversarial network to supplement the original features with weak feature expression and convert them into high-quality features to improve the small target detection performance.

6. Conclusion

In this paper, the performances of the various one-stage and two-stage object detection methods are compared. The results show that the one-stage model is usually light-weighted and could unify the candidate region generation, feature extraction, and classification into one stage. It is suitable to be applied to some scenarios with limited computational capacity. However, the performance of these models is worse than the two-stage strategies. The two-stage methods sequentially select candidate regions and then implement the classification techniques for identifying the category. These methods perform superior to the one-stage solution, which could be applied to scenarios that require high detection accuracy.

In the future, there remains some problem that requires solving. For example, for current solutions, a large-scale dataset is required to train the model, which is laborious and expensive. Future works could pay more attention to using weakly supervised or semi-supervised strategies for solving the object detection problem. In addition, current methods do not perform satisfactorily on small objects. Some new mechanisms are required to improve the detection accuracy of small objects.

References

[1] Redmon J, Divvala S, Girshick R, et al., 2016, You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp: 779-788.

[2] Redmon J, Farhadi A., 2017, YOLO9000: better, faster, stronger, In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 7263-7271.

DOI: 10.54254/2755-2721/5/20230556

[3] Redmon J, Farhadi A., 2018, Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
[4] Bochkovskiy A, Wang C Y, Liao H Y M., 2020, Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
[5] Ge Z, Liu S, Wang F, et al., 2021, Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430.
[6] Lin T Y, Goyal P, Girshick R, et al., 2017, Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp 2980-2988.
[7] Law H, Deng J., 2018, Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision, pp 734-750.
[8] Duan K, Bai S, Xie L, et al., 2019, Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 6569-6578.
[9] Tan M, Pang R, Le Q V., 2020, Efficientdet: Scalable and efficient object detection, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 10781-10790.
[10] Girshick R, Donahue J, Darrell T, et al., 2014, Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 580-587.
[11] Ren S, He K, Girshick R, et al., 2015, Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, 28.