Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 4, No. 2, 2023

A Review of YOLO Object Detection Algorithms based on Deep Learning

Xiaohan Cong *, Shixin Li, Fankai Chen, Chen Liu, Yue Meng

College of Electronic Engineering, Tianjin University of Technology and Education, Tianjin 300222, China
*** Corresponding author:** Xiaohan Cong (Email: 1136094608@qq.com)

Abstract: Object detection is a research hotspot in the field of computer vision, and YOLO series shows good performance in object detection, and has been widely used in robot vision, unmanned driving and other fields in recent years. This paper first introduces the YOLO series algorithm, including the principle, innovation points, advantages and disadvantages of various algorithms, then introduces the application field of YOLO series, and finally analyzes its future development trend to provide reference for the topic research.

Keywords: Object Detection; YOLO; Deep Learning.

1. Introduction

Object detection is a kind of image segmentation based on object geometric and statistical features, also known as object extraction. With the development of computer and object detection theory, the object detection field of deep learning should be in intelligent security, security monitoring, intelligent medical treatment, traffic detection, robot vision, unmanned driving and other fields. [1]-[3]

Based on the Convolutional Neural Network (CNN), in 2012 AlexNet [4] won the champion of ImageNet image recognition competition by a significant margin, and since then deep learning has received wide attention. OverFeat was born in 2013 and is regarded as the pioneer of single-stage object detection due to its integration of location and detection technology. The proposal of R-CNN in 2014 has achieved a very big breakthrough in object detection tasks. After that, the object detection algorithm based on deep learning began to emerge in the field of natural image recognition, and the two detection algorithms of single-stage object detection and two-stage object detection also competed with each other in their fields and excelled. The performance of R-CNN on VOC2007 has been significantly improved, and the average detection accuracy (mAP) has been improved by nearly 25% compared with the traditional method. But the regression process of this method is extremely time - and memory-intensive. In the same year, in order to improve the complex training process of R-CNN and improve the detection speed, He proposed the Spatial pyramid pool network (SPPNet) [5], which can effectively reduce the information redundancy caused by repeated calculation. The detection speed is more than 20 times that of R-CNN, and the mAP can reach 66%. The network only fine-tuned the full connection layer, but did not deal with other feature layers in the training process. In 2015, Girshick once again proposed the Fast R-CNN detector [6], which integrated the R-CNN and SPPNet network structure, and was able to train both the target category detector and the bounding box regresser during detection. It was found that the FAST R-CNN detector significantly accelerated the training process and test speed, and further improved the detection accuracy. With mAP reaching 66.9%. Soon after, Ren proposed Faster R-CNN [7], which replaces slow selective search with efficient regional candidate network (RPN), overcomes the speed bottleneck of Fast R-CNN, improves detection accuracy and operational efficiency, and achieves end-to-end target detection. mAP achieved 69.9% on the VOC2007 dataset. One of the fast versions, mAP, achieved 59.9% and is the first ever near real-time deep learning object detector. At this point, the basic architecture of the two-stage detector is determined.

This paper first introduces the YOLO series algorithm, including the principle, innovation and advantages and disadvantages of various algorithms, then introduces the application field of YOLO series, and finally analyzes its future development trend. [7]

2. YOLO Series Algorithm Development Process

2.1. Introduction to YOLO

In 2016, the one-stage object detection network was proposed. When tested in the same configuration, it was found to be able to process 45fps images per second while easily running the detection in real time. Because of its speed and its special use method, the authors give it the name YOLO (You Only Look Once). [8]

This method completely abandons the detection mode of "region candidate + regression" in the two stages, scales the image to be measured into a uniform size and divides it into multiple grids, then predicts the target category according to the grid where the target center is located, and outputs the detection results on the last convolution layer. The core idea of YOLO is to transform target detection into a regression problem, using the whole image as the input of the network, and just through a neural network, the location of the bounding box and its category can be obtained. [9]

First, the input image is uniformly adjusted to 448×448 pixels, and then the image is divided into S×S grid. If the center of the detection object falls into a grid cell, the grid cell is responsible for detecting the object. Each grid cell scores its bounding boxes to predict the likelihood of detection objects in the grid cells. If the predicted value is 0, it means no detection objects are present in the grid cells. Each of YOLO's grid cells can provide an envelope of complex numbers, but an envelope only selects the highest scoring, most likely object for prediction. Since each grid cell can only

make local predictions, it can avoid falling between several cells with good confidence at the same time, but it is not a matter of grabbing the things needed. [10] With the further development of neural networks, the YOLO series of object detection shows good performance, and many branches are derived, as shown in Figure 1 below.

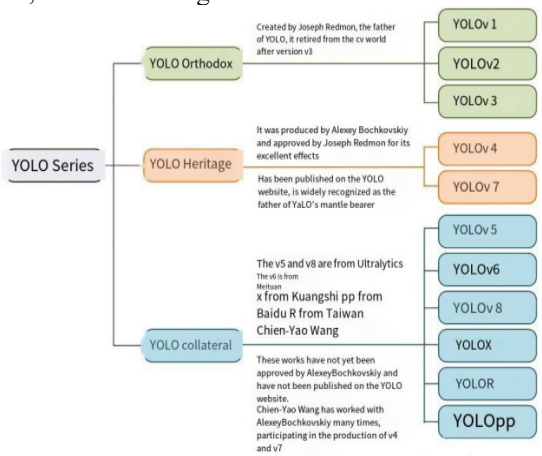

Figure 1. Part of the YOLO series products

2.2. YOLOv2

Compared with YOLOv1 algorithm, YOLOv2 adopts Darknet19 network, which includes 19 convolution layers and 5 max pooling layers. 3x3 and 1x1 convolution are mainly adopted, which are two kinds of convolution layers. The 1x1 convolution here can compress the channel number of feature map. In this way, model computation and parameters can be reduced; The introduction of Anchor and K-means clustering improved the recall rate; In order to prevent overfitting, BN layer is used after each convolutional layer to speed up model convergence. Finally, global avg pool is used for prediction. The feature fusion module (passthrough) is introduced to fuse fine-grained features. By using YOLOv2, the mAP value of the model is not significantly improved, but the calculation amount is reduced. However, YOLOv2 will lose small targets, because the resolution of the feature map is subsampled during the design, YOLOv2 cannot detect small targets well. [11] The CNN infrastructure used in YOLOv2 is relatively simple and does not use RPN network, which leads to its accuracy gap compared with some advanced object detection algorithms.

2.3. YOLOv3

Compared with YOLOv2 algorithm, YOLOv3 adopts Darknet-53 as the backbone network. Compared with Darknet-19, YOLOv3 has deeper network, more parameters and more adequate training, so the detection performance is better. YOLOv3 adopts multi-scale prediction, which can improve the detection accuracy of small targets by detecting objects on different scales. Feature pyramids with different convolution kernel sizes are used to detect objects of different sizes, which overcomes the shortcomings of YOLOv2 in small target detection. YOLOv3 uses three anchor frames of different sizes, which can better adapt to objects of different sizes and aspect ratios; And the introduction of a residual network structure, which is able to learn features better and speed up training. [12] However, YOLOv3's inference is slower because the detection process needs to run multiple convolutional layers; YOLOv3 has problems with inaccurate object positions.

2.4. YOLOv4

Compared with YOLOv3 algorithm, YOLOv4 combines many previous research techniques, combines them with appropriate innovative algorithms, and achieves a perfect balance between speed and accuracy compared with the previous YOLO series. However, YOLOv4 is not much different from YOLOv3 in essence. YOLOv4 uses CSPDarknet53 as the backbone network for feature extraction, which reduces the consumption of computing resources. SPP structure and residual connection network are introduced to accelerate and optimize the convolution process, thus improving the prediction performance and speed; post-processing optimization measures such as dynamic adjustment of IoU threshold and Logistic Activation instead of SoftMax were added to optimize the accuracy and robustness of detection results. The OpenCV DNN module is used to optimize the forward calculation performance, and the detection speed is greatly improved on the premise of ensuring the detection accuracy. However, the computing resource consumption is high, because the YOLOv4 model is relatively large and requires a lot of calculation. Compared with some other networks, YOLOv4 is more complex. Although YOLOV4 has great advantages in performance, it is also more complex, requiring more training time and more network parameters. Poor effect on s mall target detection: Because YOLOv4 uses a larger anchor box and higher prediction resolution, its effect on small target detection is worse than that of other algorithms.

2.5. YOLOv5

Compared with YOLOv4 algorithm, YOLOv5 is smaller than YOLOv4 model, so it has higher computing efficiency and lower video memory occupation. Some new evaluation indexes are added to YOLOv5, such as mAP@.5, mAP@.75, etc., which help to evaluate the model performance in a more comprehensive way. YOLOv5 enhances the learning ability of CNN, which makes it lightweight while maintaining its accuracy in the detection process. YOLOv5 can deduce effectively from single image, batch image, video and even webcam port input directly. YOLOv5s's object recognition speed of up to 140FPS is impressive. From YOLOv5n to YOLOv5x, the detection accuracy of these five YOLOV5S models gradually increased, while the detection speed gradually decreased. [13] However, the biggest drawback of YOLOv5 is its weak target detection capability. In some complex scenes, the effect will be weak. For non-standard scenes, the performance is poor. If the scene is special or the target distance is relatively far, the accuracy of YOLOv5 will be decreased and the performance is poor. Shallow network depth. Compared with some other target detection algorithms, YOLOv5 has a shallow network depth, which will affect its performance.

2.6. YOLOv7

at present, the model accuracy and inference performance are more balanced is YOLOv7 model (corresponding to the open-source git version 0.1). YOLOV7 is the most advanced algorithm of YOLO series at present, surpassing the previous YOLO series in accuracy and speed. YOLOv7 introduces the ResNet50 deep residual network, replacing the CSPDarkNet53 of YOLOv5; To improve the generation of Anchor frame, YOLOv7 introduces Anchor Free frame, which does not require preset anchor frame, and can adapt to various scales and aspect ratio targets; The multi-scale

training strategy can improve the detection ability of the model for s mall targets and better adapt to the change of image resolution. Mosaic data enhancement method is introduced, which can randomly combine images to increase the diversity and quantity of training samples. The CUDNN library is used for deep learning calculation, which improves the training and reasoning speed of the model; YOLOv7 algorithm is built based on deep learning technology, which can improve its detection performance through continuous learning and support transfer learning of deep models. [14] However, higher computing resources, more layers and parameters are required, so higher computing resources are required, and more efficient Gpus are required for training and reasoning. The detection effect of small size and crowded scenes is not good. Although YOLOv7 has improved and optimized it, the detection effect is still poor.

3. YOLO Algorithm Application Field based on Deep Learning

YOLO series involves many application fields. For YOLOv7, which has the best performance at present, the following application examples are introduced:

1. Real-time target detection: YOLOv7 can be used for real-time target detection, such as automatic driving, video surveillance and security systems. It needs to process high-resolution images and high-speed video, and can correctly detect targets, thus helping monitors and decision makers to react quickly. [15]- [16]

2. Object recognition: YOLOv7 can also implement object recognition. For example, in a large warehouse, there are many different types of products and equipment. Each object can be accurately identified and classified using YOLOv7, allowing for better management and control of inventory and production processes.

3. Face recognition: YOLOv7's face detection function can be used for applications such as security access control, intelligent gate, party check-in, etc. By using YOLOv7 for face detection and recognition, it can ensure that only authorized personnel can enter the area, and the activities of participants can be tracked.

4. Natural Language Processing: YOLOv7 can also be used in conjunction with natural language processing (NLP) technology to enable automatic text extraction and classification. For example, monitoring specific keywords on social media to see what people are saying about an event. This can help businesses and institutions understand the views of opinion leaders and consumers to make better marketing strategies and decisions.

5. Robot control: YOLOv7 can be used as a vision engine in robot control. Robots can use YOLOv7 to detect and track objects, such as humans, other robots or obstacles. This can help the robot avoid collisions, locate the target correctly, and react quickly in case of an emergency.

4. Look to the Future

The current YOLO algorithm based on deep learning has a great improvement in performance compared to traditional detection algorithms. Although the detection speed and accuracy are greatly improved, there are still some unsolved problems, such as incomplete data set, less data for small target detection, low detection performance, and failure to achieve real-time online detection without reducing detection accuracy. At present, in the development of society, country and other aspects, target detection is a big topic, and real-time online detection is the top priority, which is worthy of continued research and development. This paper puts forward several prospects for the future development direction of YOLO:

1. Data enhancement: Adding more data can improve the accuracy of the model. Data enhancement can be achieved by rotating, scaling, cutting, flipping, color transformation, etc., on the training image. Consider using more data enhancement techniques to augment the data set.

2. Better pre-trained model: Consider using stronger pre-trained models to initialize the YOLOv7 network to improve the accuracy and robustness of the model. For example, larger and deeper models such as ResNet, EfficientNet, etc., can be used.

3 Adaptive learning rates: To better optimize the model, you can use the adaptive learning rate to adjust the learning rate of each layer to better balance speed and precision during training.

4. Improvement of activation function: It is possible to consider using other activation functions such as Swish, Mish, etc. to replace LeakyReLU used in YOLOv7 to improve the performance of the model.

5. Improvement of target detection loss function: The cross-entropy loss function used by YOLOv7 is one of the common target detection loss functions, but it may not be optimal in some cases. You can try to use other target detection Loss functions, such as Focal Loss, IoU Loss, etc.

6. multi-task learning: In addition to target detection, other tasks such as semantic segmentation and instance segmentation can be included in the training to improve the comprehensive performance of the model.

7.Model compression: Some model compression techniques, such as pruning and quantization, can be considered to reduce the model size and inference time of YOLOv7, so as to adapt to the deployment on low-power devices. These are just some ideas, and how to combine and implement them still needs specific experiments and debugging. How to achieve these improvements while maintaining the performance of the model also needs to be considered.

5. Conclusion

Based on the development history of YOLO, this paper introduces the principle, innovation points, advantages and disadvantages, application fields and future development trend of YOLO series algorithms, and shows the effect that YOLO series algorithms can achieve at present, which can greatly improve the efficiency of production and life in the application field. It has great application prospects in the field of safety monitoring with high flexibility and timeliness. This paper looks forward to YOLO series and provides ideas for the future development of YOLO series.

References

[1] LIU Li, OUYANG Wanli, WANG Xiaogang, et al. Deep learning for generic object detection: A survey[J]. International Journal of Computer Vision, 2020, 128(2): 261 -- 318. doi: 10. 1007/ s11263-019-01247-4.

[2] ZOU Zhengxia, SHI Zhenwei, GUO Yuhong, et al. Object detection in 20 years: A survey[J]. arXiv preprint arXiv: 1905.05055, 2019.

[3] DALAL N and TRIGGS B. Histograms of oriented gradients for human detection[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005:886-893. doi: 10.1109/ CVPR.2005.177.

[4] Liu Pu, Zhang Xing-Hui, Zhang Zhi-Li et al. Overview of Object Detection from RCNN to YOLO [C]// China High-tech Industrialization Research Association Intelligent Information Processing Industrialization Branch. The sixteenth national signal and intelligent information processing and application of academic conference proceedings. [publisher unknown], 2022:8. DOI: 10.26914 / Arthur c. nkihy. 2022.053359.

[5] Liu Yang. Masks worn under the dense crowd scene detection algorithm research [D]. Hebei university of engineering, 2022. The DOI: 10.27104 /, dc nki. Ghbjy. 2022.000365.

[6] Shi Duan-Yang, Lin Qiang, Hu Bing, ZHANG Xin-Yu. Radar Science and Technology,2022,20(06):589-605.

[7] Zhou Shujuan, Zhang Yu, Zhang Heng, Wang Qi, Liu Guangjie, Zhu Jinlong. Object recognition of TEM nanoparticle structures based on deep learning [J]. Journal of Changchun Normal University,2022,41(12):35-40.

[8] Zhang Shan, Lu Yujiao, Luo Dawei. Overview of object Detection Algorithms based on Deep Learning [C]// Proceedings of the 12th Annual Conference on New Network Technologies and Applications, Network Application Branch, China Computer Users Association, 2018.

[9] Zhou Xiaoyan, Wang Ke, Li Lingyan. An overview of object detection algorithms based on Deep Learning [J]. Electronic Measurement Technology, 2017, 40(11):5.

[10] Zhang Qi, ZHANG Rongmei, Chen Bin. A review of image recognition technology based on deep learning [J]. 2019.

[11] Zheng Wei-cheng, LI Xue-Wei, LIU Hong-zhe. Overview of object Detection Algorithms based on Deep Learning [C]// The 22nd Annual Conference on New Network Technologies and Applications, 2018, Network Application Branch of China Computer Users Association. 0.

[12] Liu Yan-Qing. Improved Object Detection Algorithm Based on YOLO Series [D]. Jilin University.

[13] Zheng Weicheng, LI Xuewei, LIU Hongzhe. An overview of object detection algorithms based on Deep Learning [J]. China Broadband, 2022(3):3.

[14] Li Bingzhen, Jiang Wenzhi, Gu Guide, et al. Review of object detection algorithms based on Convolutional neural Networks [J]. Computer and Digital Engineering, 2022(005):050.

[15] Guo Zefang. Overview of Deep Learning Algorithms for Image Object Detection [J]. Mechanical Engineering & Automation, 2019 (1):4.

[16] Wang Yan. Research on visual detection and tracking technology of surgical instruments based on deep learning.