ISSN: 2424-8460(O)
2251-2608(P)

Original Research Article

A comprehensive review of YOLO object detection algorithms

Haonan Tian

Hunan University of Science and Technology, School of Computer Science and Engineering, Xiangtan, Hunan, 411100, China

Abstract: Object detection technology plays a crucial role in the field of computer vision and has made significant progress in recent years. Among them, the You Only Look Once(YOLO) object detection algorithms have attracted attention due to their speed, accuracy, and end-to-end characteristics. In order to promote the development and practical application of object detection, this paper provides a comprehensive review of the development history, technical principles, strengths, weaknesses, and future development directions of the YOLO object detection algorithms. Firstly, the technical principles of the YOLO algorithms are detailed, including the end-to-end detection frameworks based on a single neural network and the concept of dense prediction. Then, this paper reviews the evolution of YOLO algorithms, from YOLOv1 to the latest version, and combs their key technical innovations and performance improvements.Finally, we discuss the potential research directions for the future.

Keywords: Deep learning; Computer vision; Object detection; YOLO

1. Introduction

In recent years, with the rapid development of deep learning technology, especially the emergence of Convolutional Neural Networks (CNNs), object detection technology has made great progress and shown remarkable performance in numerous practical applications. Object detection aims to identify the locations of specific objects in images or videos and classify them[1]. This technology holds vast potential applications in fields such as autonomous driving, video surveillance, and medical image analysis.

In the developmental trajectory of object detection techniques, the You Only Look Once(YOLO) object detection algorithm[2] stands as a significant milestone. The YOLO algorithms garner considerable attention in academia and industry due to their efficient real-time detection speed and excellent performance. By treating the object detection task as a single regression problem, the YOLO algorithms employs a fully convolutional neural network to directly detect objects on images, achieving higher detection speed and good detection accuracy.

This paper aims to provide a comprehensive overview and analysis of the YOLO object detection algorithms. Firstly, the basic principles of the YOLO algorithm are elaborated in detail.Subsequently, starting from its development history, the core technologies and key components, including network architecture, loss functions, training strategies, etc., are systematically introduced. Finally, the future development directions of the YOLO algorithm are explored in depth.

2. The principle of YOLO algorithms

The anchor-based YOLO object detection algorithms set dense anchor boxes on the image, and then uses the information learned by the neural network to confirm whether these boxes contain objects or not and to adjust the positions of the anchor boxes. As shown in **Figure 1**, the input image is divided into a grid, with k anchor

boxes of different sizes and ratios set around each grid center to cover as much of the entire image as possible. The image is then fed into a convolutional neural network for feature extraction. During this process, due to the convolution operation's ability to learn feature information from surrounding pixels, the grids on the feature map have a larger receptive field, gradually incorporate information from other positions on the image and are no longer confined to the current grid, thereby providing conditions for detecting larger objects. The size of the feature map output by the convolutional neural network is consistent with the grid partition. A feature vector on the feature map contains image information within the receptive field centered on the corresponding grid position in the original image, thus enabling detection of objects at that position. After passing through the detection head for classification and bounding box regression, each anchor box on the corresponding grid learns its 4 position parameters, confidence score, and class probability. Confidence score represents the confidence of whether the anchor box detects an object. When it falls below a set threshold, it is considered as not detecting an object, and the corresponding anchor box is removed. If an object is detected, the class with the highest probability is taken as the object's class, and the anchor box is adjusted by the learned position parameters to generate accurate detection boxes.

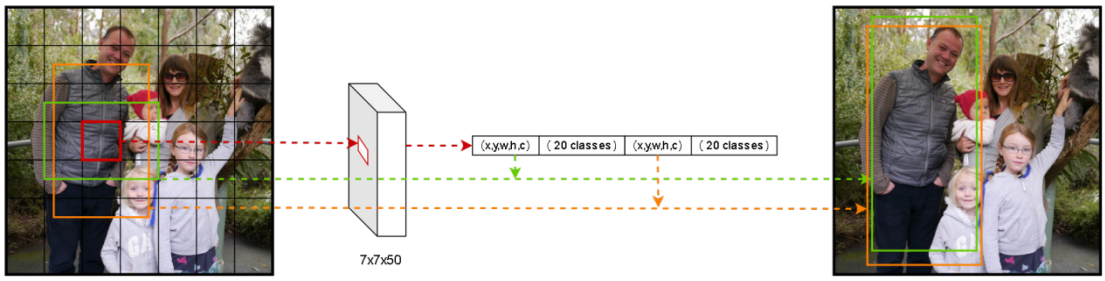

Figure 1. The schematic diagram of YOLO algorithms.

3. YOLO series

3.1. YOLOv1-v5

The YOLO algorithms are one-stage object detection algorithms. They skips the step of region proposals and accomplish object classification and localization simultaneously, greatly simplifying the detection process, thus resulting in higher detection speeds. Since their release, YOLO series have undergone several significant version updates. Each brings technological advancements and performance improvements. YOLOv1 regards object detection as a regression problem. It resizes the input image to a fixed resolution, then divides it into dense grids, with each grid responsible for detecting objects whose center falls within it. The input image of YOLOv1 passes through 24 convolutional layers to extract image features, then through two fully connected layers and adjustment operations to produce prediction feature maps with the same resolution as the grid. Each feature vector on the feature map contains detection information for the corresponding grid, including category information and position information for multiple detection boxes.

YOLOv2[3] introduces the anchor box strategy, which sets up a group of anchor boxes with different sizes and ratios on each grid to cover different positions and scales of the entire image. Each anchor box on the grid can predict a target, so they can detect multiple objects within the same grid. YOLOv3[4] evolves the backbone network from Darknet-19 used in YOLOv2 to Darknet-53, deepens the number of network layers, and introduces the residual structure in ResNet, which strengthens the feature extraction capability of the network. The algorithm also draws inspiration from the idea of a feature pyramid network and thus use multiple feature maps of different scales to detect objects of different sizes. YOLOv4[5] adopts CSPDarknet53 as the backbone network and uses the FPN+PAN structure for feature fusion in the neck network. It transmits the position informa-

tion of low-level feature maps back to high-level ones, improving the network's localization ability. YOLOv5 makes few improvements on YOLOv4 and provides versions with different scales. These versions have the same network structure, and they control the depth of the model and the number of convolution kernels through depth coefficients and width coefficients to realize models of different complexity.

3.2. YOLOv7

YOLOv7[6] is the latest anchor-based version in the YOLO series, which surpasses previous object detection algorithms in both detection speed and accuracy within the range of 5 FPS to 160 FPS. It is overall similar to YOLOv5, and the main improvements are the network module, auxiliary training head, and label assignment strategy, etc. YOLOv7 scales the images to a uniform size, inputs them into the backbone network for feature extraction and downsampling at the same time to obtain the feature maps that contain information of different scales. It then realizes feature fusion through the structure of the FPN+PAN network. The fused feature map is sent to the detection head to obtain the detection result.

The key modules of the YOLOv7 detection network include the MP-1 module and the ELAN module. Their structures are illustrated in **Figure 2**. The MP-1 module is a downsampling module with two branches. The first branch undergoes max-pooling followed by convolution to halve the number of channels, while the second branch initially transforms the channel dimension with convolution, followed by downsampling using a convolutional layer with a stride of 2. Finally, the feature maps from both branches are concatenated to maintain the total number of channels unchanged. The ELAN module is a network structure designed for efficient feature extraction. By controlling the shortest and longest gradient paths, it enables the network to learn more features and exhibit stronger robustness. ELAN comprises two branches: the first branch reduces the channel number through convolution, while the second branch passes through four convolutional layers after downsampling with convolution. Subsequently, three feature maps from the second branch and the feature map from the first branch are concatenated, and convolutional modules are utilized to fuse information from different channels.

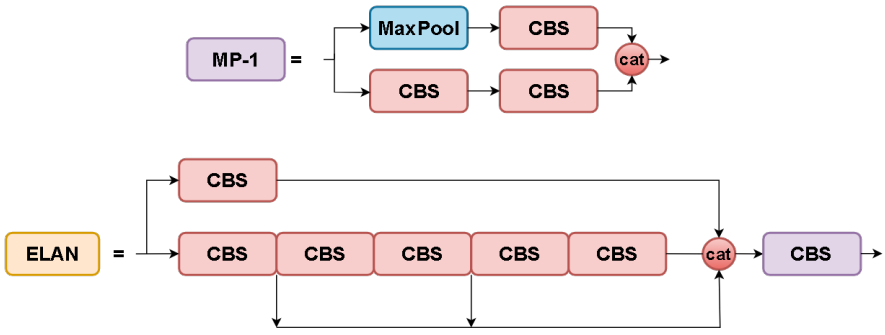

Figure 2. Structure of YOLOv7 network modules.

3.3. YOLOv8

YOLOv8 is a significant update version of Ultralytics' YOLOv5, which introduces new improved techniques to further enhance the algorithm's detection performance. These improvements include a new backbone network, anchor-free detection heads, and loss functions. Similar to YOLOv5, YOLOv8 also provides models of different scales (n/s/m/l/x) based on scaling factors, suitable for detection on devices with varying performance.

The primary innovative modules of YOLOv8 are the C2f module and the detection head, whose structures are illustrated in **Figure 3**. The C2f module draws inspiration from the ELAN module in YOLOv7, which allows YOLOv8 to maintain lightweight while obtaining richer gradient flow information. The detection heads employ the prevalent decoupled head structure, which separates classification and bounding box regression. YOLOv8 abandons the anchor box strategy in favor of anchor-free design.

ISSN: 2424-8460(O)
2251-2608(P)

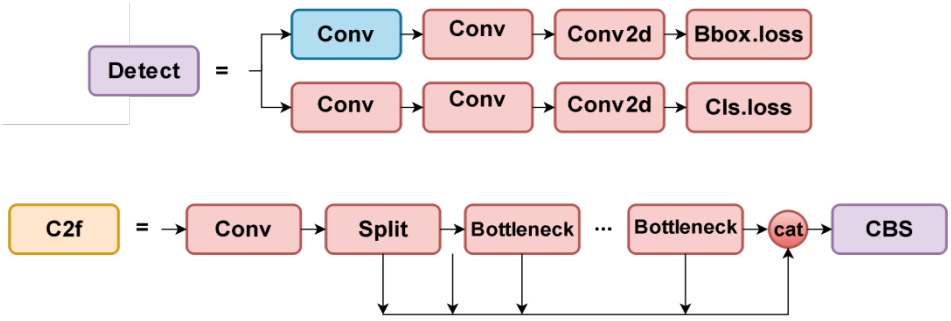

Figure 3. Structure of YOLOv8 network modules.

YOLOv8 adopts a dynamic positive and negative sample allocation strategy. It selects positive samples based on the weighted results of classification and regression scores, which enables the network to dynamically focus on high-quality samples after training. The loss function of YOLOv8 consists of two branches: the classification branch and the bounding box regression branch, without using a commonly employed foreground branch. The classification branch employs Binary Cross Entropy (BCE) loss, while the bounding box regression branch utilizes Distribution Focal Loss and CIoU loss. The total loss is a weighted sum of these three losses according to certain proportions. Mosaic data augmentation is disabled in the last ten epochs of model training in YOLOv8 to prevent regression in detection performance.

4. Conclusion

Object detection, as a crucial research area in the field of computer vision, has made significant progress in recent years. The YOLO series algorithms have garnered considerable attention due to their concise and efficient design.By transforming the object detection problem into a single regression task and implementing end-to-end detection using convolutional neural networks, YOLO has greatly enhanced the speed and accuracy of object detection. From YOLOv1 to YOLOv8, this series of algorithms has continuously evolved in terms of accuracy and speed, providing better solutions for real-time object detection applications.

Although the YOLO object detection algorithms have achieved remarkable success, there are still some challenges and areas for improvement. Future research directions and development focuses include the following aspects:

(1) Accuracy Improvement: While the YOLO algorithms have made significant breakthroughs in speed, there is still room for improvement in detection accuracy compared to some higher-accuracy object detection algorithms. Future research could focus on how to enhance detection accuracy while maintaining high speed, such as by introducing more complex network structures, refining loss functions, optimizing post-processing steps, and other methods.

(2) Real-time Performance and Efficiency: YOLO algorithms are renowned for their real-time performance and efficiency. However, as application scenarios continue to expand and hardware technology advances, the demand for more efficient real-time detection becomes increasingly pressing. Future research could further enhance the algorithm's real-time performance and efficiency through additional optimization of network structures, designing more effective computational mechanisms, leveraging hardware acceleration, and other methods.

(3) Interpretability and Security: In critical application areas such as autonomous driving and medical diagnosis, the interpretability and security of algorithms are crucial. Future research could explore methods to improve the interpretability of the YOLO algorithms, reduce the risk of misjudgment, and thereby enhance their reliability in safety-critical applications.

(4) Small Object Detection: The YOLO algorithms face certain challenges in detecting small objects, as these objects often lose detailed information on low-resolution feature maps. Future research can explore how to improve network structures or feature extraction methods to enhance the detection capability for small objects. This could involve incorporating techniques such as multi-scale feature fusion and attention mechanisms.

(5) Detection in Complex Scenes: The performance of the YOLO algorithms may decline when handling complex scenes, such as those with dense objects, occlusions, or complex backgrounds. This is due to interference between objects and background noise. Future research could focus on improving object representation, incorporating scene semantic information, and enhancing object correlation to improve detection performance in complex scenes.

In summary, the future research directions for the YOLO object detection algorithms include but are not limited to improving detection accuracy, enhancing the ability to detect small objects and objects in complex scenes, increasing real-time performance and efficiency, and exploring the interpretability of neural networks. With continuous technological advancements and evolving application needs, the YOLO algorithms will experience more innovations and breakthroughs in the future.

About the author

Haonan Tian is a master's student at the School of Computer Science and Engineering, Hunan University of Science and Technology. His research focuses on object detection in crowded scenes.

References

[1] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2016: 779-788.

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

[4] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[5] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[6] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.