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

Original Research Article

Optimization of YOLO for real-time object detection in robotic applications

Ran Wang

University of Shanghai for Science and Technology, Shanghai, 200093, China

Abstract: With the rapid development of robotics technology, real-time object detection has become increasingly crucial in robotic application scenarios. YOLO (You Only Look Once), as an efficient object detection algorithm, has been widely applied in the field of robotics. However, in practical applications, it still faces issues such as the balance between speed and accuracy, and poor performance in detecting small objects. This paper delves deeply into the optimization strategies for YOLO real-time object detection in robotic applications. By improving the network structure, optimizing the training algorithm, and adopting multi-modal data fusion and other methods, it aims to enhance the performance of the YOLO algorithm in robotic applications, providing strong support for robots to achieve accurate perception and decision-making in complex environments.

Keywords: Robotics; YOLO algorithm; Object detection; Optimization strategies

1. Introduction

Robots are widely used in numerous scenarios such as industrial production, logistics and warehousing, the service sector, security monitoring, etc., which places higher demands on their environmental perception and target recognition capabilities. Real-time and accurate object detection is the foundation for robots to achieve autonomous decision-making and intelligent operation. The YOLO algorithm holds an important position in the field of real-time object detection due to its fast detection speed. It can directly predict the category and location of objects on a single image, greatly improving the detection efficiency. However, in the actual application scenarios of robots, such as under complex lighting changes, occlusion situations, and when detecting small objects, the performance of the YOLO algorithm needs to be further enhanced. Therefore, optimizing the YOLO real-time object detection has important theoretical significance and practical application value.

2. YOLO algorithm and its advantages in real-time object detection

The YOLO algorithm is a deep learning-based real-time object detection method that transforms object detection into a regression problem, enabling target localization and classification through a single forward pass. Compared to traditional two-stage detection methods, YOLO offers significant advantages. First, YOLO achieves end-to-end detection, greatly improving detection speed and meeting real-time requirements. Second, YOLO utilizes global image information for prediction, reducing background false positives and enhancing detection accuracy. Additionally, YOLO's relatively simple network structure makes it easy to implement and deploy.

In robotic applications, real-time object detection is crucial for environmental perception and decision-making. YOLO's speed and accuracy make it an ideal choice for robotic vision systems. For example, in autonomous navigation, YOLO can detect obstacles in real time, providing reliable information for path planning. In target tracking, YOLO can quickly identify and locate targets, improving tracking efficiency. Thus, optimizing YOLO for real-time object detection in robotics has significant practical implications.

3. YOLO algorithm principle

The YOLO algorithm transforms the object detection problem into a regression problem, predicting the category and position of the object directly on the image through a forward propagation. Specifically, the YOLO algorithm divides the input image into S×S grids, with each grid responsible for predicting B bounding boxes and their corresponding confidence and category probabilities. Confidence indicates the likelihood of the bounding box containing the target and the degree of matching between the predicted box and the real box, while category probability indicates the category to which the target belongs within the bounding box[1].

The network structure of YOLO algorithm usually consists of convolutional layers, pooling layers, and fully connected layers. Convolutional layers are used to extract features from images, pooling layers are used to reduce the size of feature maps, and fully connected layers are used to output the final detection results. Through end-to-end training, the YOLO algorithm can learn the mapping relationship between image features and target categories and positions.

4. Analysis of real-time object detection requirements in robotic applications

Robotic applications impose multiple requirements on real-time object detection, including speed, accuracy, and resource constraints. First, real-time performance is a fundamental requirement. Robots must react quickly in complex and dynamic environments, necessitating extremely high processing speeds for object detection algorithms. Although YOLO already offers fast detection speeds, it may still struggle to meet real-time requirements when processing high-resolution images or complex scenes.

Second, detection accuracy is critical for ensuring the effectiveness of robotic tasks. Robots must accurately identify and locate targets to avoid misjudgments and missed detections. While YOLO achieves high accuracy in general scenarios, its performance may degrade when detecting small objects, occluded targets, or in complex backgrounds. Therefore, improving accuracy while maintaining speed is a key direction for optimizing YOLO.

Finally, resource constraints pose another challenge for robotic applications. Most robotic platforms have limited computational resources and storage, making it difficult to support complex and large deep learning models. Thus, optimizing YOLO also requires considering model complexity and resource consumption to ensure efficient operation on resource-constrained platforms[2].

5. Optimization strategies for YOLO-Based real-time object detection in robotic applications

5.1. Improving network structure

Introducing attention mechanisms, such as the SE (Squeeze-and-Excitation) module from SENet or the CBAM (Convolutional Block Attention Module), into the YOLO network. For example, CBAM calculates attention from both channel and spatial dimensions. In the channel dimension, global average pooling extracts global features for each channel, followed by a squeeze-and-excitation operation through fully connected layers to enhance focus on important channel features. In the spatial dimension, convolutional operations generate spatial attention maps, highlighting features in target regions. This enables the network to focus more on target objects, especially small or occluded ones, thereby improving detection accuracy.

To better detect objects of varying sizes, multi-scale feature fusion techniques are employed. Drawing inspiration from FPN (Feature Pyramid Network), a feature pyramid is constructed within the YOLO network.

By fusing high-resolution features from shallow layers with high-semantic features from deep layers, the network captures both detailed and semantic information. For instance, outputs from different convolutional layers in YOLO are upsampled or downsampled to the same size and then concatenated. The fused feature maps contain rich multi-scale information, enabling better detection of objects of different sizes. This is particularly useful for robots detecting distant pedestrians or nearby obstacles[3].

5.2. Optimizing training algorithms

In addition to traditional data augmentation methods like random cropping, flipping, and rotation, specific augmentation techniques tailored to robotic applications are introduced. For example, to account for varying lighting conditions, training data is processed to simulate different light intensities and angles, helping the network learn target features under diverse lighting. Additionally, artificial occlusions of varying degrees and shapes are added to training images to simulate real-world scenarios, enabling the network to detect targets accurately despite occlusions. These diverse augmentation strategies enhance the network's generalization capabilities.

To address YOLO's limitations in detecting small objects and localization accuracy, the loss function is optimized. The original loss function is modified to increase the weight of small object localization. The localization loss weight is dynamically adjusted based on object size, with higher weights assigned to smaller objects to prioritize their position prediction. Furthermore, variants of IoU (Intersection over Union) loss, such as GIoU (Generalized IoU), DIoU (Distance-IoU), or CIoU (Complete-IoU), are introduced. These improved loss functions consider not only the overlap between predicted and ground-truth boxes but also their distance, scale, and aspect ratio, guiding the network more effectively in localization learning and improving detection accuracy.

5.3. Multi-modal data fusion

In robotic applications, combining visual images with LiDAR data provides more comprehensive environmental information. LiDAR offers precise distance measurements, excelling in obstacle detection, while visual images contain rich texture and color information, aiding object classification. By fusing these data types, the network leverages both visual and distance information for more accurate object detection. For instance, LiDAR point cloud data is projected onto visual images, and the fused data is fed into the YOLO network for training and detection. This approach enhances detection performance in complex environments[4].

Beyond visual and LiDAR data, other sensor data, such as ultrasonic or infrared, can be integrated. Ultrasonic sensors provide accurate short-range distance measurements, while infrared sensors excel in detecting heat-emitting objects. Fusing these sensor data with visual inputs enriches the network's information, improving detection capabilities in special environments. For example, in dark environments, infrared data complements visual images, aiding target detection.

6. Conclusion

The optimized YOLO algorithm significantly improves detection speed and reduces resource consumption while maintaining high accuracy. These enhancements make YOLO more suitable for resource-constrained robotic platforms, offering new solutions for the development of robotic vision systems. As deep learning and robotic applications continue to evolve, the optimization and application of YOLO in real-time object detection will have even broader prospects.

About the author

Wang Ran (2004.01-), male, Han ethnicity, Baotou, Inner Mongolia Autonomous Region. Research direction: Artificial Intelligence,robot.

Nationality: Han nationality,

Native place: Weifang City, Shandong Province,

Education: senior high school, professional title: none,

References

[1] Zhao P ,Chen J ,Li J , et al. Design and Testing of an autonomous laser weeding robot for strawberry fields based on DIN-LW-YOLO [J]. Computers and Electronics in Agriculture, 2025, 229 109808-109808.

[2] Ahmad A ,Y. E E ,Jafar A , et al. A Real-Time Olive Fruit Detection for Harvesting Robot Based on YOLO Algorithms [J]. Acta Technologica Agriculturae, 2023, 26 (3): 121-132.

[3] Harada R ,Oyama T ,Fujimoto K , et al. Trash Detection Algorithm Suitable for Mobile Robots Using Improved YOLO:Regular Papers [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2023, 27 (4): 622-631.

[4] Reis D ,Welfer ,Cuadros L S D , et al. Mobile Robot Navigation Using an Object Recognition Software with RGBD Images and the YOLO Algorithm [J]. Applied Artificial Intelligence, 2019, 33 (14): 1290-1305.