

Improved YOLO Method Based on Attention Mechanism

Yizhe Zheng

*College of Computing, City University of Hong Kong, Hongkong, China
yizzheng3-c@my.cityu.edu.hk*

Abstract: Object detection, as one of the core tasks of computer vision, possesses extensive implications in domains such as autonomous driving and video surveillance. Single stage detectors represented by the YOLO series have attracted attention for their efficient inference speed, but their detection accuracy still faces challenges in complex scenarios. The traditional YOLO model extracts feature through stacked convolutional layers, but lacks explicit modeling of inter channel dependencies, resulting in insufficient ability to distinguish key features. Attention mechanisms have significantly improved model performance by dynamically adjusting feature weights. The SE module proposed by SENet enhances the response of important channels through channel recalibration, but existing works often embed it directly into the Backbone network, which may cause redundant calculations due to module stacking and affect real-time performance. In addition, YOLO's Neck part is responsible for multi-scale feature fusion, and its optimization is crucial for detection performance, but there is little research on attention design for this region. This article proposes a lightweight SE-YOLO model that integrates the SE module into the Neck section of YOLO. Specifically, a channel attention mechanism is introduced after the SPP layer to capture channel level statistical information through global average pooling, and adaptive weights are generated using fully connected layers to enhance target related features while suppressing background noise. Experiments on the PASCAL VOC dataset show that the improved model performs well mAP@0.5 By improving the indicators by 1%, This research offers fresh concepts for the lightweight improvement of object detection models.

Keywords: YOLO, Attention mechanism, Deep learning.

1. Introduction

Improved YOLO method based Object detection is one of the core technologies in the field of computer vision[1], providing critical support for applications such as autonomous driving, video surveillance, and medical image analysis. In autonomous driving, real-time detection of pedestrians, vehicles, and traffic signs directly affects system safety; In video surveillance, abnormal behavior recognition relies on efficient target tracking; In medical imaging[2], the accuracy of lesion localization is crucial for diagnostic efficiency. The common requirement of these scenarios is that algorithms need to balance high accuracy and real-time performance, while traditional detection models often struggle to balance the contradiction between the two.

The attention mechanism[3] provides a solution to the above problem by simulating the selective attention characteristics of the human visual system. This mechanism can dynamically adjust the weights of different channels or spatial positions in the feature map, enhance the feature response of

target related regions, and suppress redundant background noise. In recent years, classic attention modules such as SENet[4] and CBAM[5] have shown outstanding performance in classification tasks, but their applications in object detection are still limited to simple stacking of Backbone networks, which may compromise real-time performance due to computational redundancy, and have not been optimized for the multi-scale feature interaction requirements unique to detection tasks.

As the YOLO series iterates from v1 to subsequent versions[6], the model continues to improve in feature extraction, loss function, and network structure, but there is still a lack of systematic exploration of fusion strategies for attention mechanisms[7]. Existing research has mostly focused on improving the Backbone section, while neglecting the optimization potential of the Neck and Head[8]. In fact, the Neck part plays a key role in aggregating deep and shallow features, and its design directly affects the sensitivity of the model to multi-scale targets.

This article proposes a lightweight channel attention enhancement method for YOLO models[9]. This article embeds an improved SE module in the Neck section of YOLO to optimize multi-scale feature fusion through channel recalibration; Meanwhile, ResNet-18[10] is used instead of Darknet-53 to reduce the number of parameters while maintaining the discriminative ability of the attention mechanism; This study is based on the PASCAL VOC dataset to validate the effectiveness and generalization of the model. This study seeks to promote the efficient application of attention mechanisms in object detection through structural innovation and experimental verification, providing technical references for balancing accuracy and speed in practical scenarios.

The following are this article's primary contributions.

(1) This article proposes a lightweight attention guided by Neck. Integrate the SE module into the Neck region of YOLO, and improve the model's capacity to concentrate on key targets through channel adaptive weighting in the feature fusion stage.

(2) This article uses ResNet-18 instead of Darknet-53, which reduces computational complexity while retaining strong feature extraction capabilities, making it more suitable for lightweight deployment.

(3) This article systematically verifies the adaptability of the attention mechanism to detection tasks. Through comparative experiments, it has been proven that the attention enhancement in the Neck section has a significant effect on detecting small and occluded targets, providing a new architectural design perspective for optimizing real-time detection models.on attention mechanism.

2. Dataset and methods

2.1. Dataset

This study is based on the PASCAL VOC 2007 dataset for experimental validation. PASCAL VOC is a seminal benchmark dataset in the domain of object detection, jointly released by institutions such as Oxford University and the University of Edinburgh. It is widely used for algorithm performance evaluation due to its high annotation quality and rich scene diversity. VOC2007 contains 9963 natural scene images, covering 20 common object categories, with a total of 24640 annotated instances. The target scale in the dataset varies significantly, and there are complex situations such as partial occlusion and lighting changes, which can effectively test the robustness of the model in real scenes.

2.2. Overall structure

The model in this article is improved based on the YOLOV3 framework, and the overall structure is shown in Figure 1. The main parts of the improvement include Backbone, Neck, and Head. The model optimizes the feature fusion ability of the Neck part by introducing a lightweight channel attention mechanism, which improves detection accuracy while maintaining real-time performance.

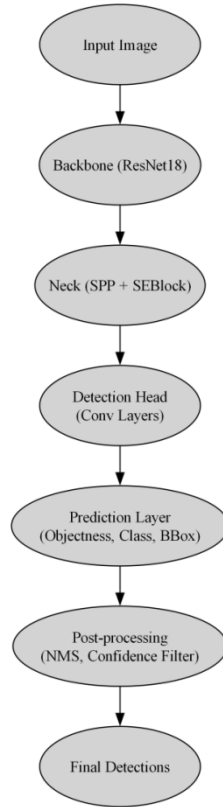


Figure 1: Overall structure diagram (picture credit: original)

2.3. Module analysis

2.3.1. Backbone

This model adopts ResNet-18 as the backbone network, and its core design is based on a residual learning framework, which alleviates the gradient vanishing problem of deep networks by introducing skip connections. ResNet-18 consists of four residual stages, each consisting of multiple residual blocks stacked together. Subsequent to traversing the initial 7×7 convolutional layer and the 3×3 max pooling layer, the input image is sequentially processed through four residual stages, with output feature map steps of 4, 8, 16, and 32, respectively. Consistent with the classic ResNet design, each residual block is composed of two 3×3 convolutional layers, and the channel dimension is adjusted through 1×1 convolution in skip connections. The final output deep feature map has a size of $1/32$ of the input image (i.e. a stride of 32) and a channel count of 512. It is worth noting that although ResNet-18 has the ability to extract multi-level features, this model only uses the high-order semantic features of the last layer without explicitly fusing multi-scale information. This may result in insufficient sensitivity for detecting small-scale targets, but it is beneficial for reducing computational complexity.

During the feature extraction process, ResNet-18 gradually downsamples to increase the receptive field and find global contextual information. Its residual structure allows the network to maintain stable gradient propagation during deep training, effectively extracting the texture, shape, and semantic features of the target. However, due to the use of only a single scale feature map, the adaptability of the model to scale changes depends on the multi-scale feature enhancement ability of the subsequent Neck module. Experiments have shown that this design has a high efficiency accuracy balance in medium-sized object detection tasks, but may limit the detection performance of extreme scale objects such as extremely small or large objects.

2.3.2. Neck

The Neck module is composed of spatial pyramid pooling and compression excitation network in series, aiming to enhance the expressive ability of the output features of the backbone network.

The SPP module adopts a parallel multi branch maximum pooling strategy, using three different scales of pooling kernels, 5×5 , 9×9 , and 13×13 , to perform spatial pyramid pooling on the input feature map. After each branch pooling operation, the feature map is restored to its original spatial size through interpolation or cropping, and then concatenated along the channel dimension to form multi-scale fused features. This process expands the number of channels in the feature map to four times the original input, and then compresses it back to 512 channels through a 1×1 convolutional layer to achieve a compact representation of cross scale features. The SPP module significantly enhances the model's ability to represent multi-scale targets, especially partially occluded or deformed targets, by fusing contextual information from different receptive fields.

SE Block dynamically adjusts the weight distribution of feature channels through the channel attention mechanism, and its structure includes two key stages:

Compression: Compressing geographical data via global average pooling dimensions into 1×1 , generate channel description vectors, and capture global spatial information.

Motivation: Learn the nonlinear relationship between channels through Two layers that are completely joined and apply the Sigmoid function to generate channel weight coefficients of 0-1. Finally, the original feature map is multiplied by the weight coefficients channel by channel to achieve adaptive recalibration of the feature channels.

The combination strategy of SPP and SE Block combines the dual advantages of spatial multi-scale fusion and channel dimension feature selection. SPP enhances the adaptability of the model to target scale changes through spatial pyramid structure, while SE Block suppresses irrelevant background features through an attention mechanism, highlighting key channels related to target detection. Experimental results have shown that this design can improve the average accuracy of mAP by about 1% while maintaining computational efficiency.

3. Experimental result

3.1. Experimental configuration

This experiment uses a stochastic gradient descent optimizer (SGD) for model training, with the basic learning rate of 1×10^{-3} , momentum factor of 0.9, and weight decay of 5×10^{-4} . An equivalent batch size of 64 is achieved through 8 gradient accumulations. The learning rate strategy includes a fourth power progressive warm-up in the first training cycle, and a step wise decay to 10% of the current value in the 90th and 120th cycles, with a total training cycle of 150 rounds. The data augmentation scheme integrates multi-scale training, random horizontal flipping, color jitter, and standardization processing.

3.2. Experimental results and comparative analysis

In order to validate the performance of the algorithm, the commonly used VOC2007 dataset was selected as samples for training, as shown in Figure 2. In the VOC2007 dataset, the training and testing sets were pre divided and did not require manual proportional partitioning. Then, the training set is fed into the improved YOLOv3 network for training. When the trained model achieves the best results on the validation set, the model is saved and tested on the test set.



Figure 2: Original data (picture credit: original)

The results are shown in Figure 3. Figure 3 shows the detection result of the image, indicating that both the person and the horse in the image have been successfully detected. The confidence level of this model for "person" is 0.83, and the confidence level for "horse" is 0.67, indicating that the model has a certain level of credibility for the identified target. It can be seen that the detection box accurately frames the positions of people and horses, indicating that the model has good performance in object localization and can meet the basic needs of object detection for this type of image.



Figure 3: Training effect (picture credit: original)

In order to further validate the efficiency of the method suggested in this research, comparative experiments were undertaken with the native YOLOv3 and the algorithm in this article. The experimental findings are presented in Table 1.

Table 1: Experiment results

	This article's algorithm	YOLOV3 algorithm
mAP@0.5	70.6	69.6
Precision	68.2	65.3
Recall	83.3	73.1

From the experimental results, it can be seen that the recognition accuracy of our algorithm is as high as 70.6%, which is higher than the native YOLOV3 algorithm. Other performance indicators such as accuracy and recall are also higher than the native YOLOV3 algorithm, making it more accurate for image detection

4. Conclusions

This article uses an improved YOLOV3 model based on an attention mechanism to train and test the performance indicators of the model using the VOC2007 dataset. ResNet-18 is used instead of the original Darknet-53 to reduce computational complexity while retaining strong feature extraction capabilities, making it more suitable for lightweight deployment. At the same time, add the SE module to the Neck area to improve the model's ability to focus on key targets. The fundamental contribution of this work is to test the application dimensions of attention processes in detection models and reveal the optimization potential of Neck networks; Establish a new paradigm for the collaborative design of lightweight networks and attention mechanisms, and break through the performance bottleneck of traditional models through structural innovation; The experimental verification results show that the improved method has a certain degree of improvement in the performance of the model, specifically in $\text{map}@0.5$. There have been varying degrees of improvement in accuracy and recall. Subsequent work will continue to adjust the structure and parameters of the algorithm to achieve better testing performance and speed.

References

- [1] Li, W., Feng, X. S., Zha, K., et al. (2021). Summary of target detection algorithms. In *Journal of Physics: Conference Series* (Vol. 1757, No. 1, p. 012003). IOP Publishing.
- [2] Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology*, 10(3), 257–273.
- [3] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).
- [4] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141).
- [5] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3–19).
- [6] Terven, J., Córdova-Esparza, D. M., & Romero-González, J. A. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, 5(4), 1680–1716.
- [7] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- [8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (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).
- [9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).
- [10] Gao, C., Meng, D., Yang, Y., Wang, Y., Zhou, X., & Hauptmann, A. G. (2013). Infrared patch-image model for small target detection in a single image. *IEEE Transactions on Image Processing*, 22(12), 4996–5009.