Computers and Artificial Intelligence；Vol.2, No.2; 2025
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.25v2n2.3743
Published by: Zeus Press

A Lightweight Object Detection Algorithm Based on Improved YOLOv8

Guochao Wan[1*], and Tao Ming[2]

[1]*Guizhou University of Finance and Economics, Guiyang, Guizhou, China*

[2]*School of Information, Guizhou University of Finance and Economics, Guiyang, China*

**Corresponding author: Guochao Wan.*

Abstract

Lightweight object detection algorithms are crucial in the field of computer vision, directly affecting whether computer vision algorithms can be deployed on resource-constrained devices and meet the real-time requirements of daily life. To address the above problems, this paper proposes a lightweight object detection algorithm LAD-YOLO based on improved YOLOv8. First, we optimize the point-wise convolution in depthwise separable convolution to enhance the model's learning ability, introduce depthwise separable convolution into the backbone network and neck network to reduce the model size, and construct a lightweight detection head. Meanwhile, the LSKA (Large Separable Kernel Attention) mechanism is introduced to help the model capture multi-scale information and achieve better detection performance. Extensive experiments conducted on the VOC dataset show that the proposed LAD-YOLO algorithm improves the precision (P) and mAP0.5:0.95 by 2.5% and 1.8% respectively compared with YOLOv8n, while maintaining lower parameters and computational complexity.

Keywords

lightweight object detection, LAD-YOLO, large separable kernel attention, lightweight convolution

1. Introduction

With the development of artificial intelligence, object detection has gradually become a hot computer vision task, which has been widely applied to many fields in daily life. The main task of object detection is to locate objects from input images and then determine the category of each object. With the increase of our needs for a better life, object detection tasks are also shining in daily life, such as drones, video cameras, food delivery robots, etc.

Traditional object detection has many disadvantages, such as slow speed, low accuracy, and inability to detect in real time; however, the emergence of deep learning has helped solve this major problem for object detection. Deep learning-based object detection is divided into one-stage detectors and two-stage detectors according to the detection process. The one-stage detector represented by YOLO (Redmon et al., 2016) perfectly solves the real-time problem, but it is difficult to handle complex image features in complex backgrounds; the two-stage detector represented by Faster-RCNN (Girshick, 2015) pursues higher detection accuracy.

With the application of object detection on constrained devices, lightweight object detection is facing severe challenges. Two-stage detectors and some large-scale one-stage detectors far cannot meet the real-time needs of constrained devices. To solve this problem, this paper proposes a lightweight object detection algorithm LAD-YOLO based on improved YOLOv8n. The algorithm improves the feature extraction ability of the model by introducing LSKA attention, optimizes the model size through depthwise convolution in the backbone network, neck network, and head network, and is committed to achieving a win-win situation between model size and model accuracy.

2. Related Work

With the development of lightweight technology, lightweight neural networks are also continuously iterating. Take MobileNet (Howard et al., 2017) and Xception (Chollet, 2017) as examples. Xception and MobileNetV1 first used depthwise separable convolution in the network, introducing depthwise separable convolution to the public. Depthwise separable convolution is divided into depthwise convolution and point-wise convolution, which effectively reduces the high parameters brought by convolution. Moreover, due to the reduction of the number of parameters and computational complexity, depthwise separable convolution also has a faster calculation speed, making it more suitable for applications in resource-constrained environments.

The attention mechanism plays a very important role in deep learning. Once proposed, it has been widely used in multiple fields such as computer vision and natural language processing. We often need to add specific attention to help the model improve its learning ability and highlight the key features of the task. Bahdanau et al. (2014) first applied the Attention mechanism to the NLP field in 2014. LSKA is a large-kernel attention mechanism, which uses a spatial separability mechanism to decompose a k×k large convolution kernel into 1×k and k×1 small convolution kernels (Lau et al., 2024). This approach increases a small number of parameters while significantly improving model performance.

3. Methodology

3.1 YOLOv8

As an excellent one-stage object detection algorithm, YOLOv8 (Jocher et al., 2023) has quickly gained widespread attention since its release and has been used to implement various tasks. The YOLOv8n network architecture is divided into three core parts: the feature extraction backbone network (Backbone), the feature fusion neck network (Neck), and the detection head network (Head). Compared with YOLOv5 (Lisa & Bot, 2017), YOLOv8 has significant improvements in detection accuracy and speed-accuracy. YOLOv8 includes five versions: n, s, l, m, and x. This paper selects YOLOv8n as the baseline model for improvement to achieve lightweight object detection.

3.2 LAD-YOLOv8

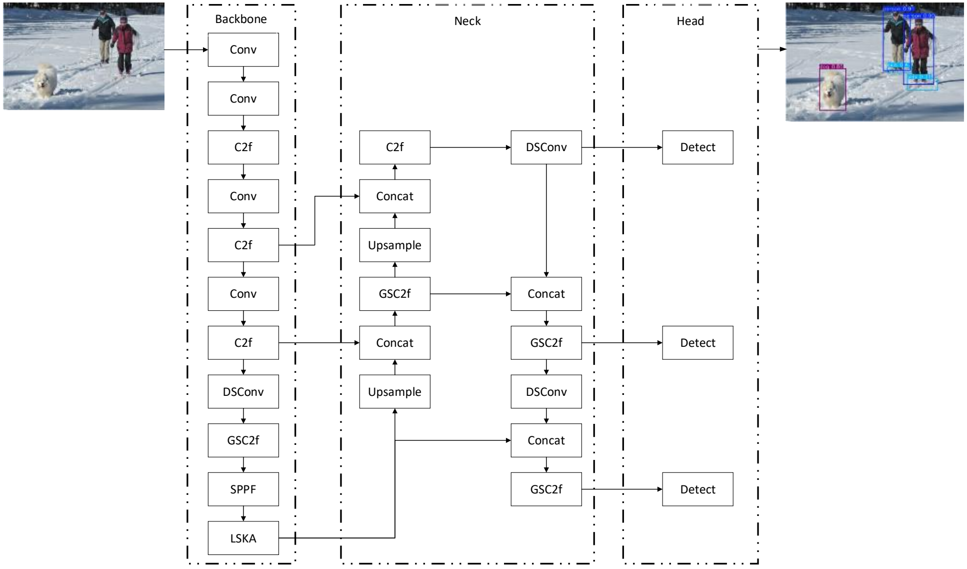
This paper is committed to achieving a balance between reducing computational complexity and improving accuracy, that is, ensuring the improvement of algorithm accuracy on the basis of reducing the computational complexity, parameters, and complexity of the original algorithm. The model structure is shown in Figure 1.

With the deepening of the backbone network, the depth of the model is also increasing. The high parameters brought by redundant channels are the main problem of the model size. Therefore, the introduction of depthwise separable convolution will greatly reduce the model size. The calculation amount ratio is shown in formula (1):

$$Ratio = \frac{\mathrm{D}_K \cdot \mathrm{D}_K \cdot \mathrm{M} \cdot \mathrm{D}_F \cdot \mathrm{D}_F + \mathrm{M} \cdot N \cdot \mathrm{D}_F \cdot \mathrm{D}_F}{\mathrm{D}_K \cdot \mathrm{D}_K \cdot \mathrm{M} \cdot N \cdot \mathrm{D}_F \cdot \mathrm{D}_F} = \frac{1}{N} + \frac{1}{{D_K}^2} \tag{1}$$

Where DK is the size of the convolution kernel, and M and N are the input and output channels respectively.

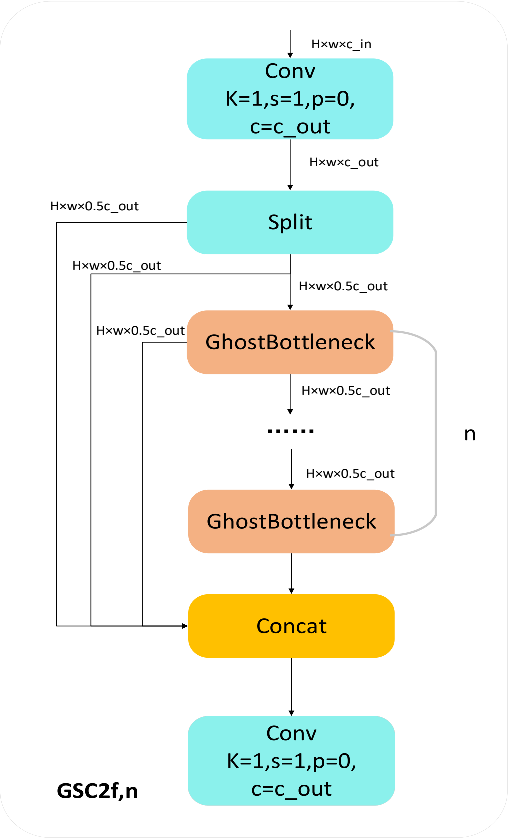
Figure 1: LAD-YOLO model diagram



According to formula (1), it can be obtained that: with the increase of the number of channels, the calculation efficiency of depthwise separable convolution is higher; therefore, we add depthwise convolution at the end of the backbone network to reduce the model size, and introduce depthwise separable convolution in the neck network to solve the problem of high channels caused by the Concat operation.

At the same time, depthwise separable convolution and Ghost convolution are introduced into the C2f module to obtain the lightweight GSC2f module, realizing the lightweight feature fusion network. The specific model structure diagram is shown in Figure 2.

Figure 2: Structure diagram of GSC2f



The detection head of YOLOv8 is a decoupled head and is divided into two branches. The Box branch is responsible for outputting the coordinates of the target, and the Cls branch is responsible for outputting the category of the target. The detection head is often bloated due to the large number of categories in the dataset, which affects the detection speed. Therefore, this paper uses depthwise separable convolution to introduce the classification head of YOLOv8, replacing the standard convolution to realize the lightweight detection head and improve the model performance.

Depthwise separable convolution is divided into depthwise convolution and point-wise convolution. However, depthwise convolution will lead to the disconnection of information in all channels. Although the 1×1 point-wise convolution can improve part of the situation, there are still defects. Therefore, the LSKA large kernel separable attention is introduced into the feature extraction network to improve the feature extraction ability of the model. The large kernel separable attention decomposes the large convolution kernel of the depthwise separable convolution layer into a combination of multiple 1D convolution kernels through the spatial separability mechanism, effectively reducing the number of parameters and computational complexity, and can maintain efficient calculation even for very large convolution kernels. By using a larger convolution kernel to capture multi-scale features, it helps the model extract richer features, and reduces the number of parameters and computational complexity of the convolution block through separable convolution, significantly improving model performance.

4. Empirical Analysis

4.1 Experimental Platform and Parameters

All experiments in this paper are based on the Linux operating system, with the CPU model being Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz, configured with an NVIDIA GeForce RTX 3090 graphics card with a total memory of 24GB, and the deep learning framework is PyTorch2.7.0+cu126.

Model training parameters: the image input size is 640, the batch size is 64, the training is 220 rounds, the optimizer is SGD, and the initial learning rate and final learning rate are 0.01 and 0.001 respectively.

4.2 Dataset

This paper uses the PASCAL VOC (2007+2012) dataset for ablation experiments (Everingham et al., 2010), which contains 20 categories such as aeroplane, bicycle, bird, boat, car, etc. Since the PASCAL VOC2007 and PASCAL VOC2012 datasets are mutually exclusive and incompatible, this paper uses VOC2007train_val (5,011) + VOC2012train_val (4,495), a total of 9,506 images as the training set, and VOC2007test with a total of 4,952 images as the validation set.

4.3 Evaluation Indicators

The performance of the model is evaluated by comparing the differences in detected images between YOLOv8 and the improved method. In terms of evaluation indicators, the main indicators selected are: Precision (P), Recall (R), and Mean Average Precision (mAP). The specific calculation formulas are shown in equations (2-5).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$AP = \int_{0}^{1} P(r)\ \mathrm{d}(r) \tag{4}$$

$$\text{mAP} = \frac{1}{\text{C}} \sum_{\text{i}}^{C} AP_i \tag{5}$$

4.4 Ablation Experiments

This paper takes YOLOv8n as the baseline model, and conducts ablation experiments on the PASCAL VOC dataset with the lightweight detection head, lightweight module, and LSKA attention respectively.

The first group is the YOLOv8 baseline model, the second to fourth groups are the mutual ablation of the lightweight detection head, GSC2f and DSConv, and the fifth group is the effect of the LAD-YOLO algorithm proposed in this paper. According to Table 1, the above-mentioned modules significantly improve the model algorithm.

Table 1: Ablation Experiments

A	B	C	P/%	R/%	mAP50/%	mAP50:95/%	GFLOPs	Parameters (M)
			0.645	0.51	0.56	0.369	8.2	3
√			0.64	0.501	0.55	0.365	7	2.7
	√		0.631	0.494	0.543	0.351	7	1.93
		√	0.664	0.522	0.562	0.374	8.3	3
√	√	√	**0.67**	**0.516**	0.57	**0.387**	**5.7**	**1.68**

4.5 Comparative Experiments

In order to further verify the detection performance of the algorithm proposed in this paper, excellent one-stage lightweight detectors are selected for comparison on the PASCAL VOC dataset, and the experimental environment is kept consistent. The comparative algorithms include different lightweight object detection algorithms of the YOLO series, such as YOLOv3-tiny (Redmon & Farhadi, 2018), YOLOv5n, YOLOv6n (Li et al., 2022), YOLOv8n, YOLOv9t (C.-Y. Wang et al., 2024), and YOLOv10n (A. Wang et al., 2024). It can be seen from Table 2 that compared with the lightweight detection algorithms in the table, LAD-YOLO leads in multiple indicators, while maintaining a lower model size (floating-point operations and parameters).

Table 2: Ablation Experiments

	P/%	R/%	mAP50/%	mAP50:95/%	GFLOPs	Parameters（M）
YOLOv3tiny	0.606	0.505	0.534	0.302	19.1	12.1
YOLOv5n	0.638	0.513	0.56	0.355	7.1	2.5
YOLOv6n	0.655	0.506	0.558	0.368	11.8	4.23
YOLOv8n	0.645	0.51	0.56	0.369	8.2	3
YOLOv9t	0.667	0.513	**0.576**	0.381	7.6	1.97
YOLOv10n	0.627	0.494	0.541	0.362	8.3	2.7
LAD-YOLO	**0.67**	**0.516**	0.57	**0.387**	**5.7**	**1.68**

5. Conclusion

Aiming at the lightweight problem faced by constrained devices in the field of object detection, this paper is committed to achieving a win-win situation between model size and accuracy, so an LAD-YOLO algorithm model based on improved YOLOv8 is proposed. The experimental results show that compared with the baseline model, this algorithm has achieved better results in various indicators, with a 2.5 percentage point increase in precision and a 1.8 percentage point increase in mAP50:95, a 31% reduction in computational complexity, and a 44% reduction in parameters. In the future, it is expected to obtain a self-built dataset through drone shooting for specific tasks, carry out experiments around this dataset, and finally deploy it on the drone for downstream object detection tasks.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint*, arXiv:1409.0473. https://doi.org/10.48550/arXiv.1409.0473

Chollet, F. (2017, 21-26 July 2017). *Xception: Deep learning with depthwise separable convolutions* [Paper presentation]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303-338. https://doi.org/10.1007/S11263-009-0275-4

Girshick, R. (2015). *Fast R-CNN* [Paper presentation]. Proceedings of the IEEE international conference on computer vision, Santiago, Chile.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*, arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861

Jocher, G., Qiu, J., & Chaurasia, A. (2023). *Ultralytics YOLO (version 8.0.0) [computer software]*. https://github.com/ultralytics/ultralytics

Lau, K. W., Po, L. M., & Rehman, Y. A. U. (2024). Large separable kernel attention: rethinking the large kernel attention design in CNN. *Expert Systems with Applications, 236*, Article 121352. https://doi.org/10.1016/J.ESWA.2023.121352

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., & Nie, W. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint*, arXiv:2209.02976. https://doi.org/10.48550/arXiv.2209.02976

Lisa, M., & Bot, H. (2017). *My Research software (version 2.0.4) [computer software]*. https://doi.org/10.5281/zenodo.1234

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). *You only look once: Unified, real-time object detection* [Paper presentation]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint*, arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., & Han, J. (2024). Yolov10: Real-time end-to-end object detection. *arXiv preprint*, arXiv:2405.14458. https://doi.org/10.48550/arXiv.2405.14458

Wang, C.-Y., Yeh, I. H., & Liao, H.-Y. M. (2024). *YOLOv9: Learning what you want to learn using programmable gradient information* [Paper presentation]. Computer Vision – ECCV 2024, Milan, Italy.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights