International Journal of Computer Applications Technology and Research
Volume 14–Issue 03, 107 – 110, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1403.1009

Explore Performance Improvements for YOLOv8_CBAM Models

Wei Ma
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Yan Chen
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Jiacui Tang
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Meiqin Wu
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Peng Xiao
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Cengyu Hou
College of Communication Engineering,
Chengdu University of Information Technology,
Chengdu, China

Abstract: In recent years, the innovative development of attention mechanism modules has provided new ideas for algorithm optimization, including large-scale separable kernel attention (LSKA), efficient multi-scale attention (EMA) and dilated multi-scale attention (MSDA). The impact of these attention mechanism modules on the performance improvement of the YOLO model remains to be explored. In this experiment, the Traffic Sign Localization and Detection dataset is used to explore how CBAM can improve the object detection performance of the yolov8 model. Experimental results show that the improved YOLOv8-CBAM model shows significant performance improvements, with a single-frame inference time increase of 0.6 ms, an average accuracy (mAP@50) of 2.1%, and a recall rate of 9.2%. Comparative experiments further reveal that the CBAM module strengthens the feature selection ability through the attention mechanism, especially in complex background or small target detection.

Keywords: Attention mechanism; YOLOv8; traffic sign detection; feature calibration; real-time perception

1. INTRODUCTION

Deep learning models are typically constructed based on multi-layer Convolutional Neural Networks (CNN), with training paradigms encompassing various modes such as supervised learning, semi-supervised learning, and unsupervised learning (Schmidhuber, 2015)[1]. CNNs exhibit powerful image feature extraction capabilities due to their unique local perception and weight sharing mechanisms (LeCun et al., 2015)[2]. The core feature extraction process is achieved through the sliding operation of convolutional kernels across the image spatial domain, and this end-to-end learning approach allows CNNs to significantly surpass traditional image processing methods in image understanding tasks (LeCun et al., 2015). The successful application of this technology has extended to multiple fields: achieving high-accuracy facial recognition in the field of computer vision (Guo et al., 2016); advancing intelligent monitoring in agriculture and medical image analysis across disciplines (Gawehn et al., 2016); and playing a critical role in environmental perception in autonomous driving systems[3].

As an innovative detection paradigm within the CNN framework, YOLO (You Only Look Once) adopts a single-stage detection strategy, achieving real-time detection by jointly predicting the coordinates of target bounding boxes and class probabilities (Redmon et al., 2016). Compared to traditional two-stage detectors, the YOLO series models exhibit two notable advantages: first, they enhance detection speed to the millisecond level through a fully convolutional network architecture; second, they effectively address the issue of missed detections in scenarios with overlapping targets by employing a dense prediction mechanism (Bochkovskiy et al., 2020)[4]. Since the introduction of the initial model by the Redmon team in 2016, the YOLO architecture has continuously evolved—YOLOv4 incorporates the CSPDarknet backbone network to strengthen feature representation (Bochkovskiy et al., 2020); YOLOv5 optimizes training strategies to enhance model generalization (Jocher, 2020); YOLOv6 and YOLOv7 achieve breakthroughs in accuracy through reparameterization design and dynamic label assignment, respectively (Li et al., 2022; Wang et al., 2023)[5]. The recently released YOLOv8, as the most representative algorithm in this series, achieves a new height in the balance between detection accuracy and inference speed (Diwan et al., 2023). Its innovative improvements include the use of mosaic data augmentation to enhance few-shot learning capabilities, the design of the C3 module to reduce computational complexity, and the introduction of an anchor-free detection mechanism to strengthen scale adaptability (Sohan et al., 2024). These technological innovations collectively drive the marginal improvement of object detection performance. YOLOv8 can not only perform object detection tasks but also simultaneously support instance segmentation and pose estimation tasks. This means users can utilize the same model to accomplish various types of computer vision tasks, thereby reducing the complexity of model development and deployment[6]. Additionally, it can handle a variety of common data formats and supports exporting trained models into different formats, such as ONNX and TensorRT, facilitating deployment across various hardware platforms and frameworks. Ultralytics offers a range of pre-trained models for YOLOv8, which have been trained on large-scale datasets, demonstrating good generalization capabilities. Users can select appropriate pre-trained models for fine-tuning based on their needs, significantly shortening the training time and development cycle of the models[7].

The main purpose of this experiment is to analyze the role of the CBAM module in enhancing object recognition when integrated with yolov8. The experimental results indicate that CBAM primarily focuses on the interdependency relationships among the various channels of feature maps during the training process on the Traffic Sign Localization and Detection dataset, generating a weight for each channel to represent its importance, emphasizing significant channel features while suppressing those of lesser importance. Compared to the yolov8 model, the yolov8-CBAM model has achieved a 9.2% increase in recall and a 2.1% improvement in mean average precision (mAP@50), along with an enhancement in inference speed.

Finally, this paper summarizes the performance improvement of the CBAM module under the structural framework of yolov8. The structure of the paper is as follows: Section 2 reviews the work of the yolov8 algorithm with CBDAM, Section 3 describes the working methods of the CBAM module, Section 4 describes the experimental setup and results in detail, and Section 5 summarizes the research results.

2. RELATED WORK

This chapter describes the current research direction: Firstly, YOLOv8 is a model for object detection, instance segmentation, and pose estimation developed by Ultralytics on the basis of the YOLO series of algorithms, representing the crystallization of YOLO's development. YOLOv8 adopts an improved version called CSPDarknet as its backbone network. It is optimized based on the inherited CSP (Cross Stage Partial) structure, enabling more efficient extraction of image features[8]. The neck network employs a combined structure of PAN (Path Aggregation Network) and FPN (Feature Pyramid Network). This hybrid structure allows for efficient information transfer and fusion between feature maps at different scales, leveraging the top-down path of FPN for high-level semantic information and the bottom-up path of PAN to supplement low-level detail information. Additionally, CBAM was proposed in 2018 by Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon[9]. Through two modules, channel attention and spatial attention, it adjusts the feature map from both channel and spatial dimensions. The channel attention module emphasizes important feature channels while suppressing less important ones, thus enabling the model to focus more on discriminative features; the spatial attention module can focus on the spatial areas where the target objects are located, enhancing the feature response in the target regions. Furthermore, CBAM can adaptively learn attention distributions across different tasks and datasets, demonstrating excellent generalization capabilities[10]. YOLOv8 and CBAM are particularly favored in deep learning, especially in image detection.

3. SYSTEM MODEL

This section first briefly introduces the working principles of yolov8 and CBAM, and then describes the process of integrating yolov8 with CBAM.

3.1 Yolov8 structure

The YOLOv8 is primarily composed of three main components: the backbone network, the neck network, and the head network. The backbone network is responsible for extracting basic features from the input images; the neck network further processes and fuses the features output by the backbone network to enhance feature representation; the head network decodes the features based on different tasks (object detection, instance segmentation, pose estimation) and outputs the final prediction results. The structural diagram is shown in Figure 1 below.

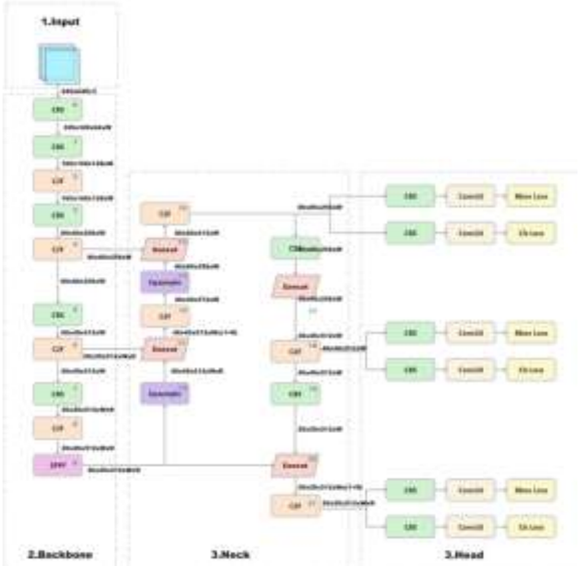

Figure 1: YOLOv8 Architecture Diagram

The backbone network is primarily composed of the CSPDarknet and the SPPF (Spatial Pyramid Pooling - Fast) module. CSPDarknet serves as the core of the YOLOv8 backbone network, where CSP reduces computational load while enhancing the model's feature extraction capability by splitting and reorganizing the feature maps along the channel dimension. The SPPF module is introduced at the end of the backbone network, allowing for pooling operations on feature maps at different scales, and subsequently merging the features of varying scales to strengthen the model's perception of objects of different sizes. The neck network PANet (Path Aggregation Network) integrates both top-down and bottom-up feature fusion paths. The top-down path facilitates the transfer of high-level semantic information to lower-level feature maps, thereby enhancing the semantic representation of lower-level features; conversely, the bottom-up path transmits low-level localization information to higher-level feature maps, improving the model's localization accuracy for objects. The detection head separates the classification and regression tasks, utilizing different branches for processing.

3.2 CBAM Principle

First, there is an original feature map in CBAM, which is the input feature map. Next, the input feature map will be sent to a Channel Attention Module and a Spatial Attention Module, and ultimately the refined feature map will be obtained.

3.2.1 Channel Attention Module

First of all, we will do a global maximum pooling downsampling and global average pooling downsampling for the input feature map F, and F will change from the original H

× W × C to two 1 × 1 × C feature maps, and then we will send these two feature maps into two fully connected layer [MLP], and finally output two 1 × 1 × C feature maps. After obtaining the two 1 × 1 × C feature plots, we add them together and limit their values to 0-1 through the sigmoid activation function, which gives us the final Channel Attention, which is M-C. and its dimensions are 1 × 1 × C. This process is represented by Equation 1 as follows:

$$M_c = \sigma(MLP(AugPool(F)) + MLP(MaxPool(F))) \qquad (1)$$

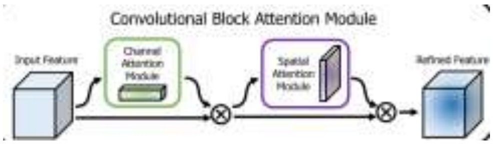

Figure 2: · Channel Attention Module

3.2.2 Spatial Attention Module

First of all, we will get a feature map of H × W × C size F ′, and the spatial attention will also perform a global maximum pooling downsampling and global average pooling downsampling respectively, but at this time we do it in the channel dimension, and the global maximum pooling will get the blue feature map in the above figure, its size is H × W × 1, and the global average pooling downsampling will get the orange feature map in the above figure, and its size is H × W × 1. Then we stitch the orange and blue feature maps in the channel dimension to get the H × W × 2 size feature maps. A convolution is then performed to turn the resulting H × W × 2 feature map into an H × W × 1 feature map. Finally, a sigmoid activation function restricts the value of the eigengram to 0-1, that is, the final M-s. 。 Its dimensions are 1 × 1 × C. This process is represented by Equation 2 as follows:

$$M_s = \sigma(f^{7\times7}([AugPool(F);MaxPool(F)])) \qquad (2)$$

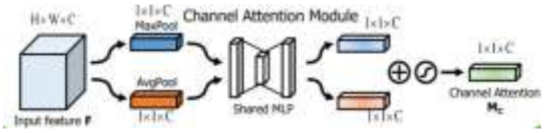

Figure 3: Spatial Attention Module

3.3 Yolov8 and CBAM Integration

A CBAM module is added behind all C2f modules in the original neck network of yolov8. First, the input feature map is processed through the backbone network and then fed into the neck network. In the neck network, the C2f module performs feature extraction and fusion, outputting a feature map rich in feature information. The output feature map of the C2f module is used as the input for the CBAM module, which is then processed sequentially through the channel attention module and the spatial attention module, resulting in a feature map weighted by the attention mechanism. The feature map processed by the CBAM module can then be passed to subsequent network layers for further feature extraction and object detection tasks. The structural diagram after the combination of Yolov8 and CDAM is shown in Figure 4:

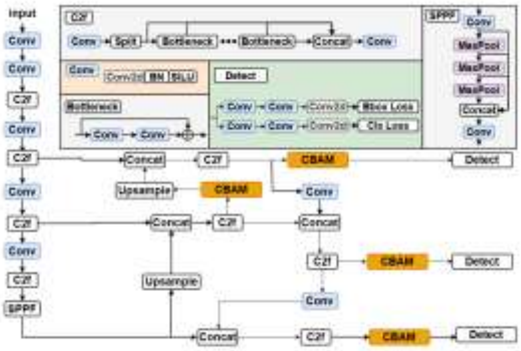

Figure 4: The structural diagram of Yolov8 combined with CDAM.

4. COMPARISON EXPERIMENT

We used the Traffic Sign Localization and Detection dataset for testing, which consists of 6164 images of traffic signs. The training set includes 4170 images, while the test set contains 1994 images, categorized into 29 classes of traffic signs. Testing was conducted for 300 rounds on both the yolov8 model and yolov8_CBAM, and it was found that the yolov8 model converged in 182 rounds, while yolov8_CBAM converged in 217 rounds. The experimental data is shown in Figures 5 and 6 below:

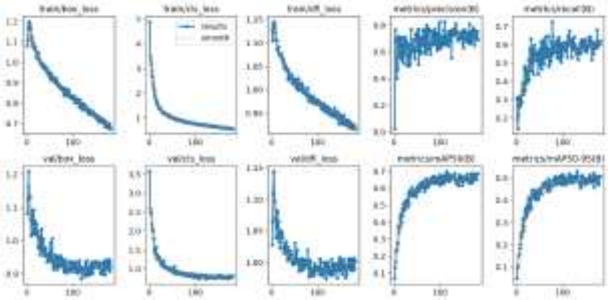

Figure 5: Yolov8 model result chart

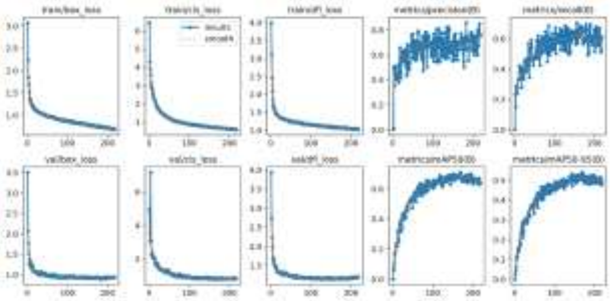

Figure 6: Yolov8_CBAM model result chart

The experimental results indicate that the yolov8_CBAM model shows faster fitting speed and better stability compared to the yolov8 model. Although the initial loss is higher, the optimization magnitude is greater. The val/box_loss of the yolov8_CBAM model decreased from 6 to 1, representing a reduction of 83.3%. The validation loss decreases in parallel with the training loss, demonstrating good generalization capability of the model. The comparison results of the experiment are shown in Table 1 below:

Table 1 Comparison test results

model	Box(precision)	recall,	mAP50	mAP(50-95)	LOPs (G)	Infer Time (ms)
Yolov8	0.775	0.579	0.688	0.529	8.1	1.2
Yolov8_CBAM	0.799	0.671	0.708	0.543	8.2	1.6

The CBAM module can improve the network's mAP_50 by 0.02, but it also increases the model's parameters and computational load, leading to an increase in inference time by 0.4ms, which can be considered negligible. This demonstrates that the CBAM module enhances the feature selection capability through the attention mechanism, particularly exhibiting superior performance in complex backgrounds or small object detection.

5. CONCLUSION

This research incorporates the CBAM module into the C2f layer of the YOLOv8 neck network, achieving dual-dimensional dynamic calibration of feature responses through channel-space. The YOLOv8-CBAM model enhances the average precision (mAP@50) by 2.1% and a recall rate increase of 9.2% while maintaining the image preprocessing time at 0.2 ms/frame. Comparative experiments further reveal that the spatial attention module can improve the accuracy of YOLOv8 in object detection.

6. REFERENCES

[1] Sapkota, R., Ahmed, D., & Karkee, M. (2024). Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. *Artificial Intelligence in Agriculture*, *13*, 84-99.

[2] Sironmani, P. P., & Augasta, M. G. (2024). A novel CNN architecture with an efficient channelization for histopathological medical image classification. *Multimedia Tools and Applications*, *83*(6), 17983-18003.

[3] Lee, S., & Rhee, J. (2024). Improving Test Accuracy on the MNIST Dataset using a Simple CNN with Batch Normalization. *Journal of The Korea Society of Computer and Information*, *29*(9), 1-7.

[4] Hussain, M. (2024). Yolov1 to v8: Unveiling each variant–a comprehensive review of yolo. *IEEE access*, *12*, 42816-42833.

[5] Dewi, Christine, et al. "Deep convolutional neural network for enhancing traffic sign recognition developed on Yolo V4." *Multimedia Tools and Applications* 81.26 (2022): 37821-37845.

[6] Moussaoui, H., Akkad, N. E., Benslimane, M., El-Shafai, W., Baihan, A., Hewage, C., & Rathore, R. S. (2024). Enhancing automated vehicle identification by integrating YOLO v8 and OCR techniques for high-precision license plate detection and recognition. *Scientific Reports*, *14*(1), 14389.

[7] Bakana, S. R., Zhang, Y., & Twala, B. (2024). WildARe-YOLO: A lightweight and efficient wild animal recognition model. *Ecological Informatics*, *80*, 102541.

[8] Karthika, B., Dharssinee, M., Reshma, V., Venkatesan, R., & Sujarani, R. (2024, June). Object Detection Using YOLO-V8. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-4). IEEE.

[9] Wang, C. (2024, October). Vehicle Target Detection Algorithm Based on CBAM-YOLOv5s. In *2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (pp. 1662-1669). IEEE.

[10] Lu, X., Jiang, Q., Shen, Y., Lin, X., Xu, F., & Zhu, Q. (2024). Enhanced residual convolutional domain adaptation network with CBAM for RUL prediction of cross-machine rolling bearing. *Reliability Engineering & System Safety*, *245*, 109976.