Applied Science and Innovative Research
ISSN 2474-4972 (Print) ISSN 2474-4980 (Online)
Vol. 8, No. 2, 2024
www.scholink.org/ojs/index.php/asir

Original Paper

Student Classroom Behavior Detection Algorithm Based on Improved YOLOv8n

Huiming Li

School of Applied Technology, University of Science and Technology Liaoning, Anshan, 114051, China

Received: May 5, 2024 Accepted: May 25, 2024 Online Published: May 29, 2024

doi:10.22158/asir.v8n2p93 URL: http://doi.org/10.22158/asir.v8n2p93

Abstract

Currently, object detection technology is becoming more and more mature, but there are still a lot of challenges in recognizing students' classroom behaviors. In order to address the problems of inaccuracy and high computation of existing models in the process of student classroom behavior recognition, this paper adopts the improved Yolov8n model to detect and recognize student classroom behavior. Based on Yolov8n, the method adds an efficient shuffle attention to increase the ability of feature extraction and improve the model recognition accuracy; secondly, the bounding box loss function is optimized to improve the model's localization ability. The experimental results show that the mAP50 and mAP50-95 metrics of the proposed model on the student classroom behavior dataset are 99.9% and 95.4%, respectively. The proposed model can achieve the detection and identification of classroom behavior more quickly and accurately with lower computing cost, and can realize the dynamic and scientific identification of students' classroom learning.

Keywords

student classroom behavior, YOLOv8, object detection, CNNs

1. Introduction

Students' classroom performance is an important part of classroom teaching evaluation, and the identification of students' classroom behavior is of great significance for classroom teaching reform and exploration. The classroom behavior detection method based on artificial intelligence and deep learning can effectively identify the typical classroom behaviors of students and help teachers to grasp the learning status of students in a timely and effective manner, thus contributing to intelligent classroom teaching.

Traditional classroom teaching behavior analysis mainly consists of two parts: feature extraction and classification, such methods require human judgment in the recognition process and are affected by other factors such as the detection target and specific tasks, resulting in low detection accuracy, and the detection of complex backgrounds has certain limitations, which makes it difficult to promote and expand on a large scale. In recent years, with the development of deep learning, convolutional neural networks (CNNs) have become the mainstream of detection and recognition technology with excellent feature extraction capabilities. Currently, mainstream object detection and recognition algorithms are divided into two major categories: a two-stage approach based on region suggestion and a one-stage approach based on object regression. R-CNN (Girshick et al., 2014) is a two-stage object detection algorithm, which uses a region suggestion network to extract candidate regions from an image, and then recalibrates the objects in the candidate regions to obtain the detection results. Subsequently, improved two-stage algorithms such as Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2016), and Mask R-CNN (Gkioxari, 2017) have been proposed to improve detection accuracy and speed. However, two-stage target detection algorithms still have the problem of slow detection speed. One-stage object detection algorithms are significantly faster than two-stage networks because do not need region suggestion and get the detection results directly by end-to-end, typical algorithms are SSD (Liu et al., 2016) and YOLO (Redmon et al., 2016; Redmon et al., 2017) series algorithms. YOLO series detection algorithms have gone through several version updates, which ensure the detection speed of the algorithms while improving the detection accuracy and therefore widely used in the field of object detection.

The recent research of student classroom behavior detection has achieved more accurate detection goals, but there are still problems such as small sample size, weak generalization ability, and slow inference speed. In order to solve these problems, this paper is based on the newly released YOLOv8n model in YOLO series, which has higher accuracy and smaller number of parameters, and improves and optimizes the method of students' classroom behavior detection, in order to develop a detection model with high accuracy and fast detection speed for resource-constrained devices.

2. Materials and Methods

2.1 Dataset

The data in this paper is from Student Behavior Detection Dataset (Burak, 2024), we have filtered and uniformly sampled the dataset, and collated a total of 2626 pieces of student classroom behavior image, the resolution of each image is 640×640 pixels, and the number of various behavior categories are: forward, 459; hand, 495; read, 403; sleep, 423; around, 482. The annotation format is YOLO format. The dataset is divided into training set, validation set and test set in the ratio of 7:2:1. Typical student classroom behaviors are shown in Figure 1.

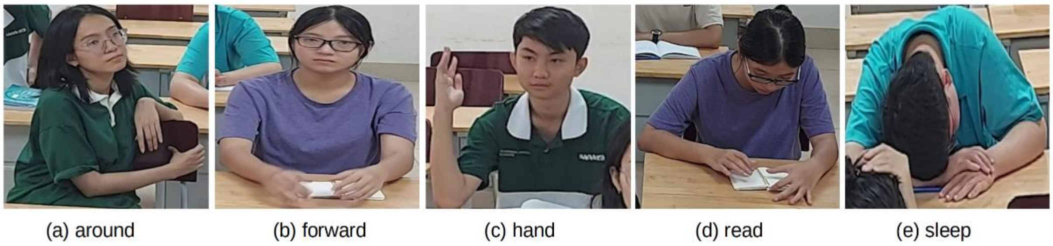
(a) around (b) forward (c) hand (d) read (e) sleep

Figure 1. Typical Student Classroom Behaviors

2.2 YOLOv8 Architecture

In the field of object detection, YOLO series models are widely used because of their high accuracy and small model size. YOLOv8 consists of three parts, Backbone, Neck and Head, and its architecture is shown in Figure. 2. Backbone is responsible for extracting high-level semantic features from the input images. Neck is located between Backbone and Head, and is used to perform multi-level feature fusion to enhance the network's ability to perceive the target. The detection head is responsible for localizing and classifying the target to generate the final detection result. Compared with YOLOv5, YOLOv8 replaces the C3 module with the C2f module, which is lighter and makes the model more adaptable to different sizes and shapes of targets, and adopts the current mainstream decoupled head structure, which effectively reduces the number of parameters and computational complexity, and enhances the model's generalization ability and robustness. In addition, YOLOv8 abandons the design of using Anchor-Base to predict the position and size of the anchor box in the previous series, and uses Anchor-Free to directly predict the center point and width-to-height ratio of the target, reducing the number of anchor boxes, which further improves the detection speed and accuracy of the model.

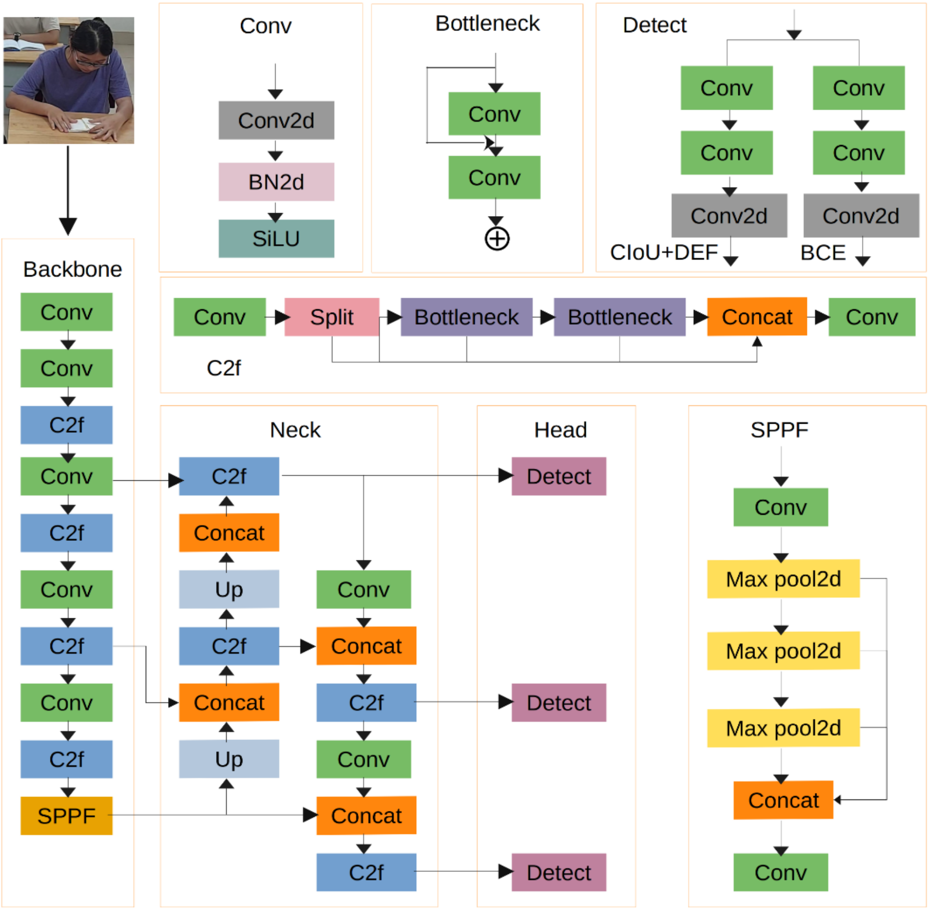

Figure 2. YOLOv8 Architecture

2.3 Network Model Improvement

2.3.1 Add Shuffle Attention

There should be significant differences in features between different targets or between targets and background, which is crucial for effectively identifying and locating targets. The object detection algorithm will recognize these features in a certain context and find regions in the image that match these features. Although the YOLOv8 model can extract detailed behavioral features of students in classroom scenarios, but classroom scenarios have complexity and variability, so how to be able to learn important features in the scenarios has become a problem that must be solved. Shuffle Attention (SA) (Zhang et al., 2021) uses replacement units to efficiently combine channel and spatial attention, to solve the problem of the high computational complexity of the combined attention. The architecture of SA is shown in Figure 3.

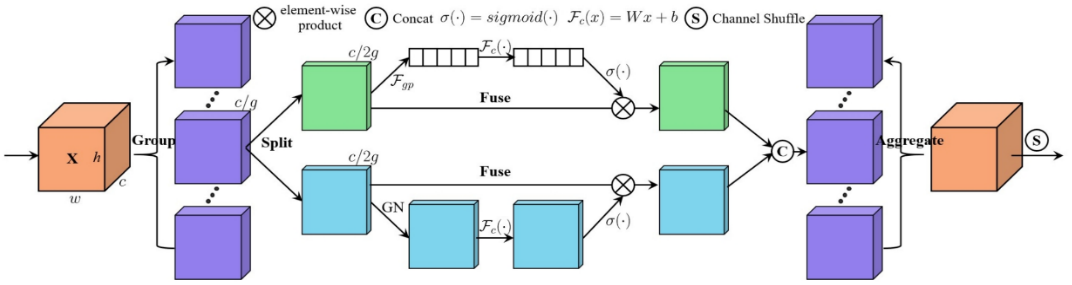

Figure 3. The Architecture of SA

Firstly, the input feature $X \in R^{(C\times H\times W)}$ divided into G groups along the channel dimension, represented as

$$X = [X_1, X_2, \dots, X_G], X_k \in C/G \times H \times W \tag{1}$$

Each sub feature X_k captures a specific semantic information. Next, X_k is divided into two branches along the channel dimension, with two sub features represented as $X_{k1}, X_{k2} \in R^{(C/2G\times H\times W)}$, X_{k1} branch captures inter channel dependencies, The X_{k2} branch captures spatial dependencies, and achieves attention to semantic and positional information through two branches. Subsequently, the two attentions are concatenated to obtain features with the same input dimension. The above operation can be achieved through the following formula:

$$s = F_{gp}(X_{k1}) = \frac{1}{H\times K}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}(i,j) \tag{2}$$

$$X'_{k1} = \sigma\big(F_c(s)\big)\cdot X_{k1} = \sigma(W_1 s + b_1)\cdot X_{k1} \tag{3}$$

$$X'_{k2} = \sigma(W_2 \cdot GN(X_{k2}) + b_2)\cdot X_{k2} \tag{4}$$

$$X'_k = [X'_{k1}, X'_{k2}] \in R_{C/G\times H\times W} \tag{5}$$

where, σ represents the activation function, GN represents group normalization. Finally, an attention map with the same dimension as the input is obtained through feature aggregation operation.

The main task of YOLOv8's backbone network is to extract features. Therefore, we add SA before the SPFF module of the backbone network to enhance the network's ability to autonomously learn key features and discard other unimportant features.

2.3.2 Optimizing the Bounding Box Loss Function

The bounding box loss function, as an important component of the object detection loss function, is crucial for the detection performance of the model. YOLOv8 uses CIoU to calculate the regression loss of bounding boxes, but it has issues such as high parameter sensitivity, computational complexity, and slow convergence speed. In order to overcome the above problems, this paper replaces the CIoU in the baseline model with WIoU (Tong et al., 2023) based on dynamic non monotonic focusing mechanism to improve the prediction speed and accuracy of the model. The calculation process is shown in equation (6):

$$L_{WIoU} = rR_{WIoU}L_{Iou}, L_{Iou} \in [0,1], R_{WIoU} \in [1,e] \tag{6}$$

where, L_{IoU} represents IoU loss, which will weaken the penalty term of high-quality anchor boxes and strengthen its focus on the centroid distance when the overlap between the anchor box and the predicted box is high; R_{WIoU} represents the penalty term of WIoU, used to strengthen the loss of ordinary quality anchor boxes; The superscript * represents not participating in backpropagation, effectively preventing the network from generating gradients that cannot converge; r dynamically allocate the gradient gain of bounding boxes and reduce harmful gradients generated by low-quality anchor boxes in the later stages of training, focusing on ordinary quality anchor boxes to improve the model's localization ability. R_{WIoU} and r are defined as shown in equations (7)-(8).

$$R_{WIoU} = exp\left(\frac{\left(x-x_g\right)^2+\left(y-y_g\right)^2}{\left(W_g^2+H_g^2\right)^*}\right) \tag{7}$$

$$r = \frac{\beta}{\delta\alpha_{\beta-\delta}} \tag{8}$$

where the parameters (x, y）, (x_g, y_g）, (W_g, H_g) represents the center position of the predicted box, the center position of the ground-truth box, and the width and height of the minimum rectangle containing the predicted box and the ground-truth box, respectively. β is the outlier with a small gradient gain, which can reduce the impact on bounding box regression. α, δ denote hyperparameters.

3. Experiment

3.1 Experimental Environment and Datasets

This experiment was conducted under the 64-bit operating system Ubuntu 23.10, the GPU was NVIDIA RTX 3060 with 12 GB of video memory, the host memory was 16 GB, the programming language was Python 3.11.6, the GPU was accelerated using CUDA v12.2, and the training was based on the deep learning framework Pytorch 2.1.0. The SGD optimizer was used, the initial learning rate was set to 0.01, the batch size was 32, and the input image size was 640×640.

3.2 Metrics

This experiment uses precision, recall, and mean average precision (mAP) to evaluate the model. The mAP50 is the average of the precision of all the samples with threshold 0.5, which reflects the trend of the model's precision with the recall rate, and the higher the value means that the model is more likely to maintain a high precision with a high recall rate. The mAP50-95 represents the average precision at different IoU thresholds (from 0.5 to 0.95 in 0.05 steps).

3.3 Experimental Results and Analysis

Figure 4 shows the loss curve and validation performance of the model during training and validation. From the graph, it can be seen that as the number of training iterations increases, the training and validation losses of the model gradually decrease, indicating that the model continuously learns more accurate features. From the perspective of validation effectiveness, the model performs well in terms of precision rate, recall rate, mAP50 and mAP50-95. Figure 5 shows the prediction results of the model on the test set. It can be seen from the figure that the model correctly predicted the classroom behavior of students with high confidence.

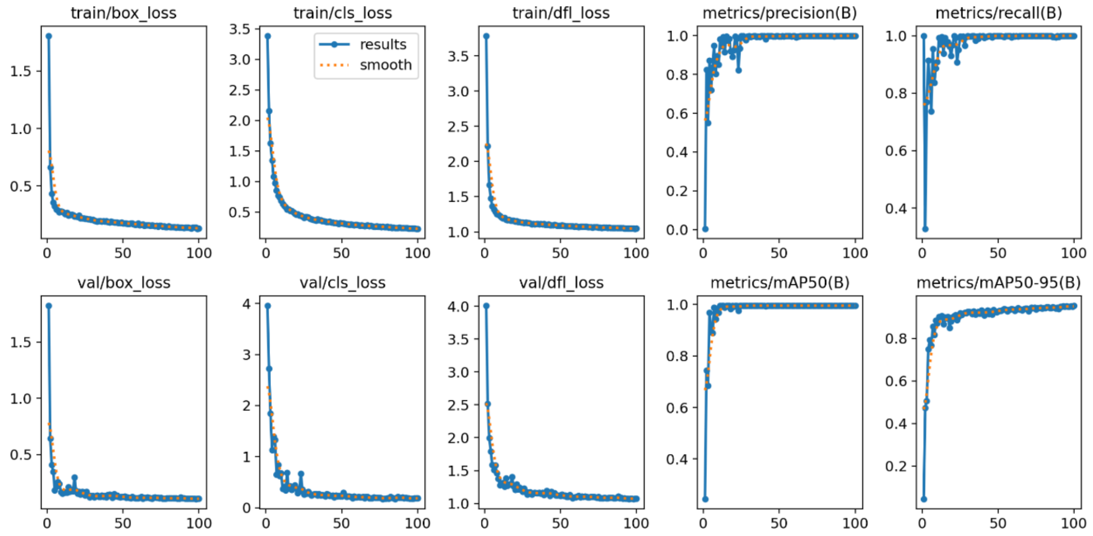

Figure 4. Loss, Accuracy, and Recall of the Model in the Training and Validation Sets

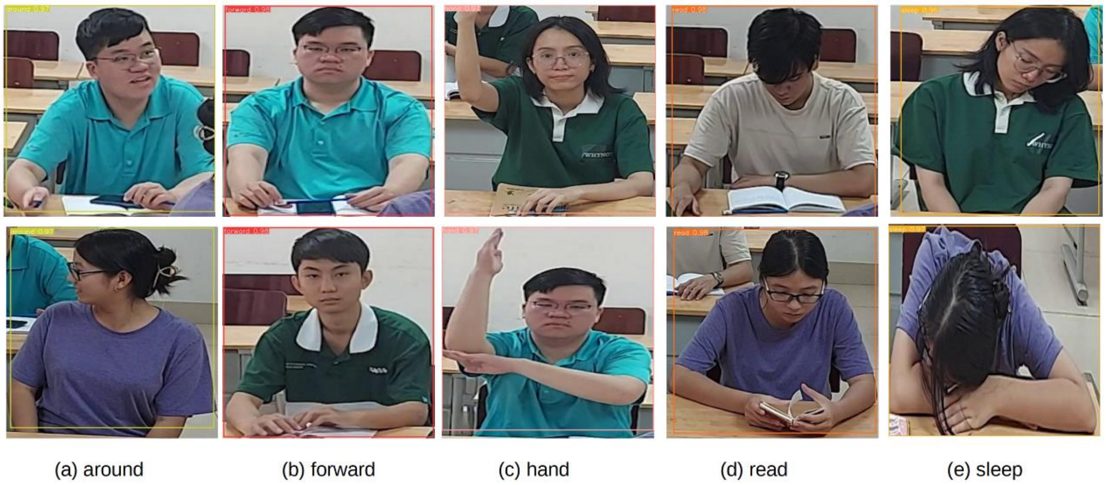
(a) around (b) forward (c) hand (d) read (e) sleep

Figure 5. Prediction Results of the Model on the Test Set

3.4 Ablation Experiment

To verify the optimization effect of each module, ablation experiments were conducted in this paper, and the results are shown in Table 1. According to Table 1, after adding SA to the baseline model, MAP50-95 increased by 0.06. After introducing the regression loss function WIoU, MAP50-95 increased by 0.02. The proposed model shows higher accuracy in detecting student classroom behavior compared to the YOLOv8n baseline model, proving the effectiveness and feasibility of the optimization module.

Table 1. The Ablation Experiment Results

Model	precision	recall	mAP50	mAP50-95
Baseline	0.999	1	0.995	0.946
Baseline+SA	0.999	1	0.995	0.952
Baseline+SA+ WIoU	0.999	1	0.995	0.954

4. Conclusion

Students' behavior in the classroom reflects the effect of classroom teaching, and also reflects their participation in the course, and the research and analysis of classroom behavior can provide direction for the evaluation of classroom teaching. In this paper, we propose a classroom behavior detection algorithm based on the improved YOLOv8n model, which improves the feature learning ability of the model by adding the attention mechanism to the baseline model, and improves the accuracy and speed of the model by introducing the WIoU bounding box loss function. The experimental results show that the proposed algorithm can effectively identify and locate students' classroom behaviors, which has certain teaching aid value.

References

Burak. (2024). *Student Behavior Detection Dataset*. https://universe.roboflow.com/burak-koyfx/student-behavior-detection

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448). https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587). https://doi.org/10.1109/CVPR.2014.81

He, K., Gkioxari, G., Doll ár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969). https://doi.org/10.1109/ICCV.2017.322

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

Redmon, J., & Farhadi, A. (2018). *Yolov3: An incremental improvement*. arxiv preprint arxiv:1804.02767.

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). https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence, 39*(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Tong, Z., Chen, Y., Xu, Z., & Yu, R. (2023). *Wise-IoU: bounding box regression loss with dynamic focusing mechanism*. arxiv preprint arxiv:2301.10051.

Zhang, Q. L., & Yang, Y. B. (2021 June). Sa-net: Shuffle attention for deep convolutional neural networks. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2235-2239). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414568