International Core Journal of Engineering Volume 8 Issue 5, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

Analysis of Classroom Teacher-Student Interaction based on Yolo

Chenyu Bai, Ziyuan Feng, and Mengyuan Yang

North China University of Science and Technology, Tangshan 063200, China

Abstract

The classroom behavior of teachers and students is an important basis for classroom teaching evaluation. Classroom behavior analysis aims to study the internal mechanism of teachers'teaching activities and students' academic development in the classroom, to help teachers and students reflect on their own classroom performance, and to promote the improvement of classroom teaching quality. With the continuous deepening of the integration of artificial intelligence technology and education, it is of great significance to change the previous methods of teacher-student classroom behavior analysis for classroom evaluation.At present, the field of Intelligent Teaching is in a fast developing period. Thanks to the continuous progress of deep learning algorithms in recent years, the method of classroom teaching evaluation based on video image processing has also emerged. Following the trend of informationization and intellectualization reform in the field of education, this project intends to build an intelligent evaluation system for teacher-student interaction behavior in classroom teaching based on YOLO algorithm for object recognition and detection.

Keywords

Intelligent Evaluation; YOLO Algorithm; Deep Learning; Target Detection; Computer Vision.

1. Introduction

Artificial intelligence is a subject that enables computers to simulate human thought processes or intelligent behaviors. In recent years, the application of artificial intelligence technology in the field of education has flourished. The Notice on the Development Plan of New Generation of Artificial Intelligence[1] issued by the State Council highlights the importance of building a new education system including intelligent learning and interactive learning, seizing important strategic opportunities for the development of artificial intelligence, building the pioneering advantages of the development of artificial intelligence in China, and accelerating the construction of an innovative country and a strong scientific and technological country in the world.

In recent years, with the rapid development of computer hardware technology and the theoretical knowledge of target detection algorithm, the target detection algorithm based on deep learning convolution neural network has gradually replaced the traditional behavior detection algorithm. The 1980s was the embryonic stage of convolution neural network. On the basis of the concept of receptive field, Fukushima, a famous Japanese scholar, proposed Neocognitron (neurocognitive machine), which laid a solid theoretical foundation for the future research of convolution neural network. With the development of in-depth learning, a large number of classical and excellent detection algorithms have appeared, including mainstream network SSD, YOLO, YOLOv2, YOLOv3, etc.

Classroom behavior analysis aims to study the internal mechanism of teachers'teaching activities and students' academic development in the classroom, to help teachers and students reflect on their own classroom performance, and to promote the quality and improvement of classroom teaching. The

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

traditional methods of classroom teaching behavior analysis mostly collect and analyze data through self-evaluation, manual supervision, manual coding, etc. There are some drawbacks such as strong subjectivity of coding, small sample size, time-consuming and laborious, which lead to its low explanability and scalability.[2] The popularization of artificial intelligence technology has brought an opportunity to improve these shortcomings. Intelligent technology can be used to collect and analyze data, to identify classroom behavior more comprehensively and timely, to gain insight into classroom teachers'teaching and students' learning status, and to provide strong support for the improvement of teaching quality.

This experiment takes teachers'classroom teaching behavior and students' classroom performance as the object of study. Video images are analyzed through in-depth learning, and an intelligent evaluation system for teachers'and students' interactive behavior is built to achieve the behavior analysis and wisdom management of classroom teaching. The application of the system in teaching practice can provide some reference for the study of classroom teaching based on artificial intelligence technology, as well as provide some technical support for classroom teaching behavior, teachers'professional development and the improvement of teaching quality.

At present, the field of Intelligent Teaching is developing rapidly. Thanks to the continuous optimization of deep learning algorithms in recent years, the classroom teaching evaluation system based on video image processing has also emerged. Following the trend of informationization and intellectualization reform in the field of education, this project intends to build an intelligent evaluation system for teacher-student interaction behavior in classroom teaching based on YOLO algorithm for object recognition and detection. It refers to the analysis of classroom teaching behavior through computer visual detection and recognition algorithm based on in-depth learning, on the basis of real-time recording and sampling of a large number of teachers and students'behavior with video monitoring.

The purpose of this study is to achieve the following functions: 1) Teacher action behavior recognition; (2) Examination of students'head-up rate and head-down rate; (3) Classroom interaction analysis. This system analyzes classroom teaching behavior from three aspects: teachers, students and teacher-student interaction. It can effectively help teachers to further understand the situation of classroom teaching, reduce the teaching pressure, and help students improve the efficiency of classroom learning, and ultimately achieve the improvement of the quality of classroom teaching. The system strives to perfectly create an intelligent evaluation system of teacher-student interaction behavior, so that science and technology can be perfectly integrated with education, and the evaluation of learning can be traced.

2. Introduction to the Principle of Yolo v3 Network

In this experiment, the YOLOv3 network is used to detect teachers'behavior and students' head-up and head-down rates, and its network structure is shown in Figure 1. YOLO series network model refers to the GoogLeNet structure to achieve end-to-end learning process[3]. YOLOv3 network is further improved and perfected based on YOLOv2, which has three advantages over YOLOv2. (1) Using the new trunk feature extraction network Darknet-53 for feature extraction; (2) Build feature pyramids to extract more effective features through multiscale feature fusion; (3) Replace Softmax with Logistic in the final object classification.

The above three improvements make YOLOv3 a target detection network with balanced speed and accuracy, on the basis of high detection speed and real-time detection. The overall detection effect of YOLOv4 is better than that of YOLOv3, but because it increases the complexity of the network structure, resulting in the overall detection speed being lower than that of YOLOv3, it is difficult to meet the requirements of real-time detection in this experiment. Therefore, the YOLOv3 network is used for the detection of teachers'and students' behavioral movements. The network structure of yolov3 is shown in Figure 1.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

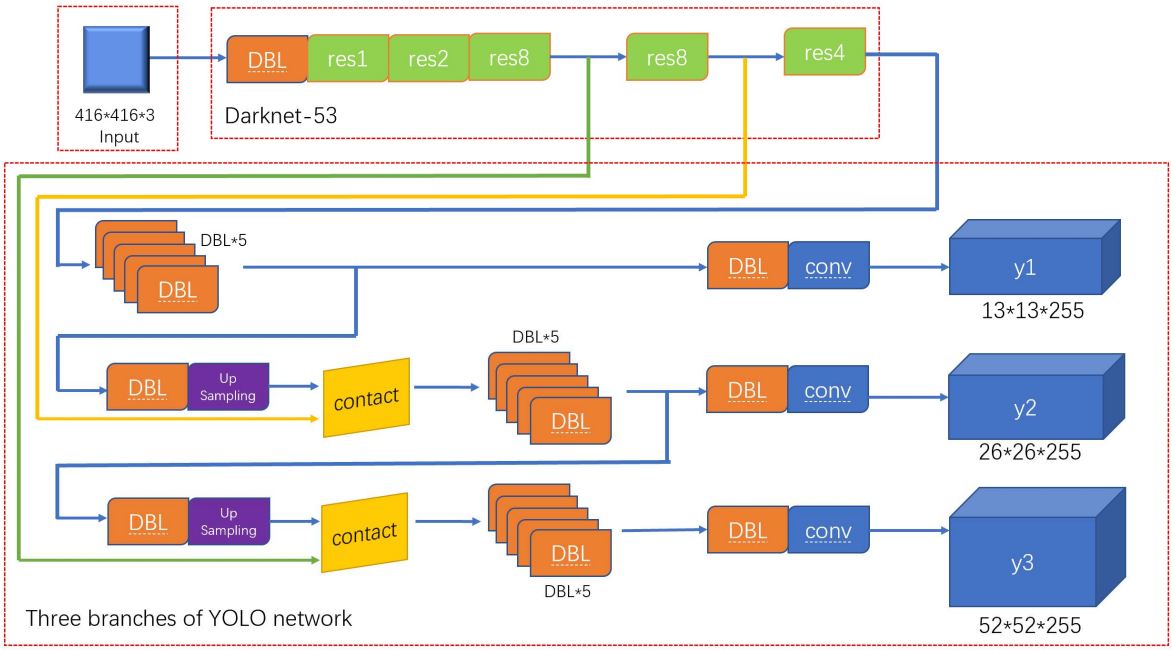

Figure 1. Yolov3 network structure diagram

YOLOv3 backbone feature extraction network Darknet-53 network is shown in Figure 2. Darknet-53 contains 53 convolution layers, which can extract rich feature information. The backbone network of YOLO-V3 consists of the first 52 convolution operations of the Darknet-53 network (excluding the last average pooling, full connection layer, and Softmax portion of the Darknet-53 network), in which the Darknet-53 network contains 26 residual blocks, six convolution layers, and one full connection layer, with two convolution operations in each residual block and the last full connection layer passing through 1 ×× 1 Convolution implements 53 convolution operations, so it is called Darknet-53.

	Type	Filters	Size	Output
	Convolutional	32	3×3	256×256
	Convolutional	64	3×3/2	128×128
1	Convolutional	32	1×1	
	Convolutional	64	3×3	
	Resnet unit			
	Convolutional	128	3×3/2	64×64
2	Convolutional	64	1×1	
	Convolutional	128	3×3	
	Resnet unit			
	Convolutional	256	3×3/2	32×32
8	Convolutional	128	1×1	
	Convolutional	256	3×3	
	Resnet unit			
	Convolutional	512	3×3/2	16×16
8	Convolutional	256	1×1	
	Convolutional	512	3×3	
	Resnet unit			
	Convolutional	1 024	3×3/2	8×8
4	Convolutional	512	1×1	
	Convolutional	1 024	3×3	
	Resnet unit			
	Avgpool		Global	
	Connected		1 000	
	Softmax			

Figure 2. Darknet-53 network

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

An important feature of Darknet53 is the use of Residual, which is structured as shown in Figure 3. The backbone network of YOLO-V3 introduces the residual unit [4], which makes the mapping after introducing the residual more sensitive to errors between predicted and real values, avoids the disappearance of gradients, the explosion of gradients, and makes the deeper network converge. At the same time, on the premise that the neural network can converge, as the network depth increases, the performance reality of the network gradually increases to saturation, and then decreases rapidly. This is caused by the fact that the identical mapping of the neural network is not easy to fit. By introducing residual units, the identical mapping can be well fitted, which ensures that the performance of the network will not decrease with the increase of the number of network layers after the optimal number of network layers is reached, so a deeper neural network can be designed.

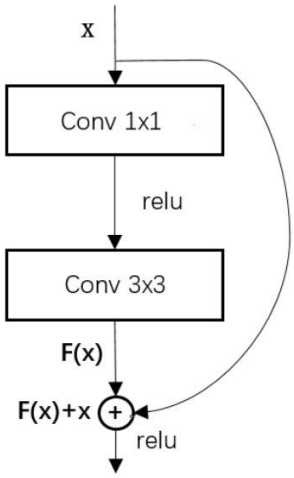

Figure 3. Residual structure

Sensory fields of different scales have different levels of semantic information. To achieve different fine-grained detection, YOLOv3 uses FPN[5] (feature pyramid networks) structure to extract feature maps of different scales for object detection. 416 for input × 416 pictures, network 32 times down-sampling at 79 layers 13 × 13-size feature map with a large field of perception and abstract semantic information is suitable for detecting large-scale targets. To detect intermediate targets, 79-layer feature maps were sampled on top and concatenated with 61-layer feature maps, resulting in 16 times down sampling of 26 × 26-size feature map. This feature map has a medium-scale sensory field and is suitable for detecting medium-scale objects. Similarly, to detect small targets, for 26 × The 26-size feature map was resampled and concatenated with the 36-layer feature map, resulting in an 8-fold down sampling of 52 × 52-size feature map. This feature map has the smallest receptive field and is suitable for detecting small-scale targets.

3. Experimental Process

3.1 Experimental Environment

The original data of the teacher's teaching video used in this experiment comes from the intelligent classroom which is collected real-time based on the camera. The smart classroom is a multi-function classroom with multimedia. There are two cameras above the classroom. The camera in the middle records the video of the teacher's behavior on the platform, while the camera in the front of the classroom records the video of the student's behavior when walking in the classroom. We use the cameras in the middle of the teacher and the cameras in front of the teacher to record behavioral videos as the source of the datasets of teacher behavior and student behavior, respectively.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

3.2 Experimental Platform

The model training and testing hardware platform uses Intel Corei7-9700K processor, NVIDIA GeForce GTX 1080 Ti11G graphics card, 16G memory, and the software platform is Windows 10 system, using CUDA 9.0 for GPU acceleration training.

3.3 Experimental Data

The data set of this experiment comes from the behavior action videos of teachers and students collected by the smart classroom camera. The behavior videos recorded by the camera in the middle of the teacher and the camera in front of the teacher are used as the source of the behavior data set of teachers and students respectively. There are 501 training set pictures and 495 test set pictures, totaling 996. There are two types of behavioral actions for teachers, face_student and face_blackboard. There are also two types of behavioral actions for students, face_up and face_down. For these types, 2440 tags are labeled in the dataset, and then pictures containing these types are divided into training and test sets at a ratio of 8:2. Some of the picture samples are shown in Figure 4.

Figure 4. Sample pictures

3.4 Model Training

The initial learning rate is set to 0.0001 during training. The decrease of learning rate uses a polynomial attenuation learning rate scheduling strategy. Bachsize is set to 8, momentum and weight attenuation coefficients are set to 0.9 and 0.0005, respectively. The YOLOv3 detector is designed to be more suitable for training with a single GPU. Therefore, only one NVIDIA GeForce GTX 1080Ti graphics card is selected for model training in this paper. The loss curves of training loss and validation loss are shown in Figure 5. It can be seen that when Epoch is greater than or equal to 10, the loss curves of both have converged to 0 and the difference between them is very small. At the end of the training process at the 50th round, the model fits very well as a whole.

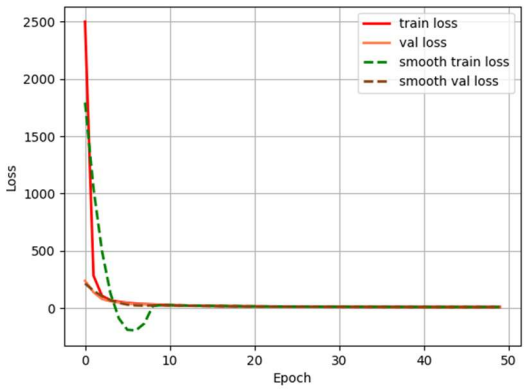

Figure 5. Loss curve

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

3.5 Model Test

Both the accuracy rate and recall rate are commonly used evaluation indicators in machine learning classification problems. For positive and negative samples P and N, there are two cases of identifying correct T and identifying error F. The positive and negative sample recognition table is shown in Table 1.

Table 1. Identification of positive and negative samples

	Positive sample	Negative sample
Retrieved	TP	FP
Not retrieved	FN	TN
Retrieved	TP	FP

In order to evaluate the behavior recognition algorithms proposed in this paper in the teaching scenarios, we compute some common indicators in the field of computer vision in the above behavior recognition algorithms. Confusion matrix is a specific matrix used to visualize the performance of deep learning algorithms. Each column represents the predicted value, and each row represents the actual category. In the confusion matrix below, the diagonal accuracy represents the accuracy of the model prediction. The confusion matrix data of teacher-student behavior and action in this experiment is shown in Figure 6.

Confusion Matrix		Predicted	
		Face Blackboard	Face Student
Actual class	Face Blackboard	0.921	0.079
	Face Student	0.036	0.964

Confusion Matrix		Predicted	
		Head Up	Head Down
Actual class	Head Up	0.886	0.091
	Head Down	0.082	0.843

Figure 6. Teacher student behavior action confusion matrix

The precision rate is the ratio of the number of correctly retrieved samples to the total number of retrieved samples, that is, the calculation formula of precision rate P is shown in formula (1):

$$P=TP/(TP+FP). \quad (1)$$

The recall rate is the ratio of the number of correctly retrieved samples to the number of samples that should be retrieved. The recall rate R is calculated as shown in formula (2):

$$R=TP/(TP+FN). \quad (2)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

The calculation results of the above indicators are shown in Table 2:

Table 2. statistical indicators of teacher and student behavior recognition

All actions	Precision p	Recall rate R	mAP
Face Student	0.984	0.965	0.975
Face Blackboard	0.926	0.933	0.931
Head Up	0.905	0.912	0.909
Head Down	0.875	0.866	0.883

It can be seen that in terms of precision rate and recall rate, the behavior recognition framework has achieved excellent performance in identifying teachers' and students' classroom behavior. Some prediction effect diagrams are shown in Figure 7.

Figure 7. Prediction effect diagram

4. Conclusion

In this study, a teacher-student interactive behavior evaluation system based on the deep learning optimization model YOLOv3 algorithm is proposed, and the data set of teacher-student interactive behavior is trained and tested by real-time monitoring video of classroom teaching. The average recognition accuracy of multiple classroom behaviors on the self-built data set of teacher-student classroom behaviors is more than 90%, which verifies the effectiveness of this algorithm in evaluating teacher-student interaction behavior. In the future, we will improve the generalization of classroom behavior assessment methods and try to expand the behavioral dataset of teachers and students to provide more diverse data for smart classroom analysis.

References

[1] Zhang Lishan, Feng Shuo, Li Tingting Formal modeling and intelligent computing for classroom teaching evaluation [J] Modern distance education research, 2021,33 (01): 13-25.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202205_8(5).0024

[2] Wu Qiushi Model driven interval estimation method for R & D cost of IT project [J] Science and technology communication, 2020,12 (11): 127-130 DOI:10.16607/j.cnki. 1674-6708.2020.11.058.

[3] Wang Zipeng, Zhang Rongfen, Liu Yuhong, Huang Jihui, Chen Zhixu Improved yolov3 garbage classification and detection model for edge computing equipment [J] Progress in laser and optoelectronics, 2022, v.59; No.711(04):291-300.

[4] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[5] Lin T Y,Dollar P,Girshick R,et al.Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Honolulu:IEEE,2017:936-944.