DRN-LSTM: A Deep Residual Network Based On Long Short-term Memory Network For Students Behaviour Recognition In Education

Zhaozhen Xuan[1*]

[1] School of Automotive Studies, Henan College of Transportation, Zhengzhou, 450015 China
[*] Corresponding author. E-mail: publicgj@163.com

Received: Mar. 07, 2022; Accepted: Apr. 19, 2022

In classroom teaching, artificial intelligence technology can help automate student behavior analysis and enable teachers to master learning efficiently and intuitively provide data support for subsequent optimization of teaching design and implementation of teaching intervention, this paper proposes a residual network based on long short-term memory network. Long short-term memory network (LSTM) is introduced on the basis of deep residual network, in which LSTM can effectively capture the temporal information of students' behaviors. The Dropout layer is introduced into the residual block to improve the accuracy and convergence speed of student behavior recognition. Finally, four behaviors closely related to learning engagement state are selected for recognition: sitting, side-turning, lowering head and raising hand. The accuracy of the detection and recognition method in the verification set reaches 96.56%. The recognition accuracy of common behaviors such as playing mobile phone and writing in class is greatly improved compared with the original model.

Keywords: student behaviour recognition; Deep residual network; LSTM; dropout; education

http://dx.doi.org/10.6180/jase.202302_26(2).0010

1. Introduction

As governments, education departments and academic accreditation bodies begin to encourage schools to shape evidence-based decision-making and innovation systems, learning analysis technology has shown great advantages in decision-making assistance and teaching evaluation. With the integration of artificial intelligence and machine learning related algorithms and technologies, learning analysis achieves higher analysis accuracy [1–3].

Students' learning commitment can help the school to better understand the quality of students' learning. The core factor to evaluate the quality of a university education is the degree of students' study devotion. As an important part of learning engagement, students' classroom behavior has always been concerned by researchers. The traditional evaluation of students' classroom behavior is realized by manual observation and recording, which is inefficient. Today, with the vigorous development of artificial intelligence, trying to improve this situation with the help of artificial intelligence technology, to understand students' learning behavior and learning status in the process of classroom learning has become an important topic of current education development [4, 5], which will promote the intelligent, efficient and comprehensive development of education analysis system. In order to promote the innovation of data collection methods for students' classroom behavior, this study selected 5 classrooms equipped with camera equipment and analyzed classroom teaching videos with the support of computer vision technology, so as to provide data support for teachers to master students' learning engagement, optimize teaching design and implement teaching intervention.

In view of the fact that there is no publicly available data set of students' classroom behavior, this paper collected the video data of 5 classrooms and processed the data to make a data set. Based on computer vision technol-

ogy, this paper proposes a multi-stage method to identify students' classroom behavior. Because the range of students' classroom behaviors does not change much, and in the video images, there will be overlapping occlusion between students, which causes no small difficulty for behavior recognition [6, 7]. By using the OpenPose human key point detection algorithm, the data of students' key points are obtained and input into the deep residual network for learning, so as to obtain the attitude classifier, which can realize the recognition and classification of students' bowing, sitting, sideways and raising hands. In addition, students' classroom behaviors are often closely related to interactive objects, such as playing with mobile phones and writing. The human joint images extracted from these two behaviors are similar and cannot be directly judged by using bone boat key points. Therefore, the hand region is regarded as the most critical semantic information for behavior recognition in these two behaviors. Due to the slow loading and processing speed of existing models, it is difficult to realize the real-time detection of students' classroom behavior. In this paper, the improved LSTM algorithm of the model is used for hand detection, and the cascade classification network is carried out after the fusion of human posture information to realize the real-time detection of mobile phone playing and writing behavior. In this paper, the accuracy and processing speed of the algorithm model are evaluated based on the real video data of students' performance in classroom teaching, and good results are obtained.

This paper is organized as follows. Section 2 illustrates the proposed method. Section 3 gives the experiments. There is a conclusion in section 4.

2. Proposed behaviour recognition method

Zhu et al. [8] proposed deep residual network for the first time to solve the problem of gradient disappearance in convolutional neural network with stacking of layers, and applied it to image recognition task, showing high classification accuracy. Deep residual networks are composed of multiple residual blocks with jumping connections. The introduction of residual idea greatly alleviates the gradient disappearance problem of deep neural networks. The structure of residual block is shown in Fig. 1.

Residual block contains two kinds of mappings, one is identity mapping; The other is residual mapping. Suppose the desired optimal solution is H(X)=X, and the residual mapping refers to the residual value of the mapping H(X) and X, expressed by F(X), that is, F(X)=H(X)-x. When F(X) is infinitely close to 0, the network will reach the optimal state and if it continues to deepen the network depth, and

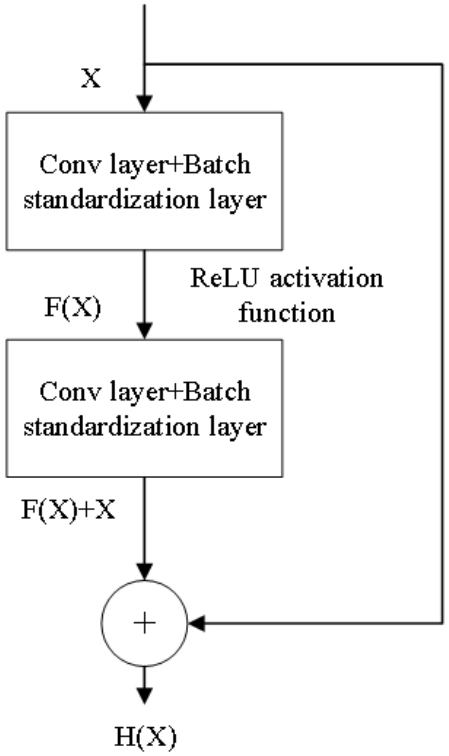

Fig. 1. Structure diagram of residual block

the network will always be in the optimal state [9]. When the input of residual block is X_n, the calculated output is,

$$X_{n+1} = f\left(X_n + F\left(X_n, W_n\right)\right) \tag{1}$$

Where $F(\cdot)$ is the residual mapping, W_n is the corresponding weight parameter, and $f(\cdot)$ is the activation function. It can be seen from Fig. 1 that the dimension mismatch may exist between different residual blocks. In this case, a linear transformation W_s is required for the identity mapping X_n.

$$X_{n+1} = f\left(W_s X_n + F\left(X_n, W_n\right)\right) \tag{2}$$

Where W_s is the weight parameter.

The residual structure has two structural advantages: first, the characteristics of shallow layer can be reused in deep layer when the network propagates forward; When the second network propagates back, the gradient in deep layer can transmit directly to shallow layer. Therefore, when there is a large reconstruction error between the input and output of the network, the residual block with fast connection can directly feed back the error information to the front network layer through the fast connection. This structural design not only improves the model training speed, but also effectively alleviates the problem of network degradation.

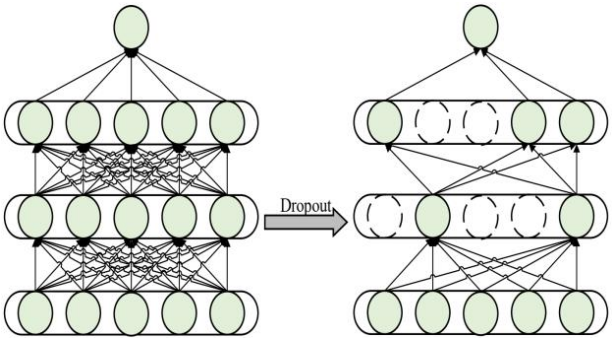

Fig. 2. Schematic diagram of Dropout layer

2.1. Dropout layer

Dropout is a method [10] to deal with the fitted problem, which is often used in the optimization of deep networks. The theoretical process from the deep feed-forward network before Dropout to the deep feed-forward network after Dropout is shown in Fig. 2. The dotted circle represents the deleted neuron, and the edge connected to it has also been deleted.

The Dropout layer works as follows. Firstly, some hidden layer neurons in the network are randomly discarded to construct a new hidden layer, while keeping the input and output neurons unchanged. Small batch training input samples are propagated forward through the newly constructed hidden layer. It then propagates back according to the result of the returned loss function. Unhidden neuron parameters are updated through optimization algorithms, and the "dropped" neurons are finally recovered and the Dropout process is repeated until training is completed.

After adding the Dropout layer, the calculation formula for the network can be expressed as:

$$r_j^{(l)} \sim \text{Bernoulli}(p) \tag{3}$$

$$\widetilde{y}^{(l)} = r^{(l)} * y^{(l)} \tag{4}$$

$$z_i^{(l+1)} = w_i^{(l+1)} * \widetilde{y}^{(l)} + b_i^{(l+1)} \tag{5}$$

$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right) \tag{6}$$

Where $r_j^{(l)}$ is the random coefficient and $y^{(l)}$ is the hidden layer neuron. $\widetilde{y}^{(l)}$ is the neuron after dropout. $z_i^{(l+1)}$ is the layer $l+1$ neuron to be activated. $y_i^{(l+1)}$ is $l+1$ layer output neuron. F() is the activation function. $w_i^{(l+1)}$ and $b_i^{(l+1)}$ are the weights and biases of layer $l+1$, respectively. By adding a Dropout layer, the number of intermediate features can be reduced, thereby reducing redundancy and complex co-adaptation relationships between neurons, increasing orthogonality between features at each layer and avoiding over-fitting [11].

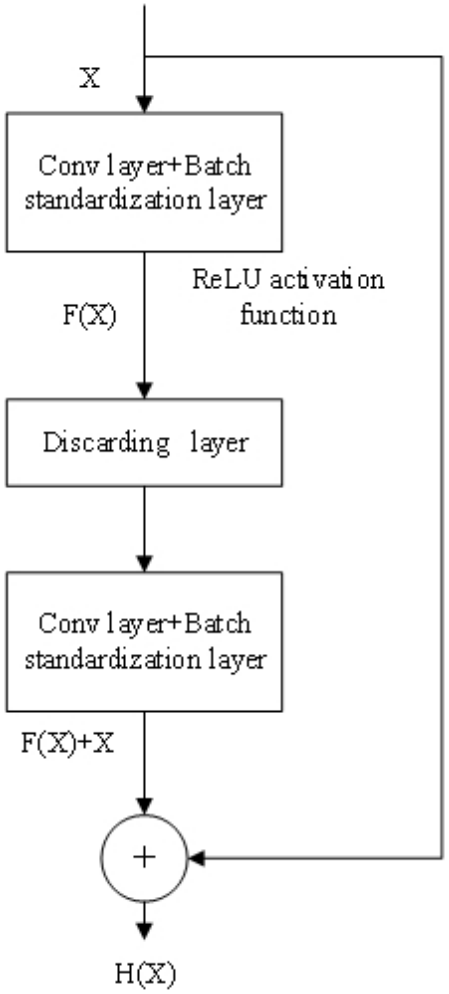

Fig. 3. Structure diagram of improved residual block

2.2. Improved residuals

In order to avoid the phenomenon of over-fitting when the number of layers is deep, the structure of residual block is improved. The standard residual block structure is shown in Fig. 1, and the improved residual block structure is shown in Fig. 3.

2.3. Combining with LSTM

Convolutional neural network (CNN) has strong feature extraction capability, and it contains convolutional layer, activation layer, pooling layer, batch standardization layer, full connection layer and classification layer, among which the classification layer is composed of multi-layer perceptron [12]. The convolutional layer carries out deep feature ex-

traction, where the convolutional process can be described as follows:

$$y_i^{l+1}(j) = K_i^1 \cdot x^l(j) + b_i^l \tag{7}$$

Where, K_i^l represents the weight of the i-th filter in layer l. b_i^l represents the deviation of the i-th filter in layer l. $x^l(j)$ represents the j-th local region in the l layer. $y_i^{l+1}(j)$ represents the input of the j-th neuron in frame i at layer $l+1$.

The pooling layer is usually connected after the convolution layer, which reduces the spatial size of features and network parameters by means of down-sampling operation. The maximum pooling operation can be described as follows:

$$P_i^{l+1}(j) = \max_{(j-1)W+1 \le t \le jW} \left\{ q_i^l(t) \right\} \tag{8}$$

Where, $q_i^l(t)$ represents the value of the t-th neuron in frame i-th of layer l, $t \in [(j-1)W+1, jW]$. W is the width of the pooled area. $P_i^{l+1}(j)$ is the corresponding value of neurons in layer $l+1$ of the pooling operation.

LSTM is a variant of recurrent neural network (RNN), and its structure is improved to alleviate the gradient problem of RNN [13]. The structure of LSTM mainly includes forgetting gate, input gate and output gate. The forgetting gate of LSTM determines the amount of information passed, and its calculation process is as follows:

$$f_t = \sigma\left(w_{L1} \cdot [h_{t-1}, x_t] + b_{L1}\right) \tag{9}$$

Where σ is the sigmoid function. w_{L1} and b_{L1} are weight and bias respectively. h_{t-1} is the output of the previous unit. x_i is the current input.

The input gate of LSTM determines whether new information can be remembered by cell units, and its calculation process is as follows.

$$i_t = \sigma\left(w_{L2} \cdot [h_{t-1}, x_t] + b_{L2}\right) \tag{10}$$

$$\widetilde{C}_t = \tanh\left(w_{L3} \cdot [h_{t-1}, x_t] + b_{L3}\right) \tag{11}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t \tag{12}$$

Where w_{L2}, w_{L3} and b_{L2}, b_{L3} are the weight and bias of input gate and memory gate units respectively. *tanh* is the activation function. f_t and i_t are the outputs of the forgetting gate and the input gate respectively. C_t is the output of the memory unit. C_{t-1} is the output of the previous memory unit. $\widetilde{C}_t$ is the output of the activation function *tanh*.

The final output h_t of LSTM unit is determined by the output o_t of the output gate and the output C_t of the memory unit. The specific calculation is as follows:

$$o_t = \sigma\left(w_{L4}\left[h_{t-1}, x_t\right] + b_{L4}\right) \tag{13}$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{14}$$

Where w_{L4} and b_{L4} are the weight and bias of the output gate respectively.

2.4. Modified Deep Residual Network Model

Considering that different models have their own advantages, model fusion can complement their strengths and learn from weaknesses from different models. Convolutional neural network (CNN) has strong feature extraction capability, which can extract deeper fault features and capture attribute information of fault occurrence. Short and long-term memory network (LSTM) has the ability of short and long-term memory and can capture the time sequence information of fault occurrence. Therefore, in order to retain the timing sequence characteristics of vibration signals to the maximum extent [14], this paper designs an improved deep residual network integrating improved residual block and LSTM layer as a student behavior recognition model. It can not only deal with the original behavior features directly, but also integrate the time sequence information of the student behavior into the model while extracting the student behavior information.

The structure of the improved residual network is shown in Fig. 4, which includes two initial convolution layers, a pooling layer, a short and long memory network layer and three improved residual blocks. Then the behavior classification is carried out by convolution layer, global average pooling layer, expansion layer, full connection layer and classification layer. The initial convolution layer is taken as an example to describe the parameters. (Conv1D,32,3) indicates that the model is an one-dimensional convolution neural network with a filter size of 32 and a convolution kernel size of 3. (MaxPooling,3) represents the maximum pooling operation, and the pooling block size is 3. (LSTM,32) represents that the model is a long and short memory network, and the number of neurons is 32. (Dropout,0.25) represents the drop layer, and the drop rate is 0.25. (Dense,64) is the fully connected layer with 64 neurons. Flatten stands for unfolded layer. ReLU represents the linear unit activation function after rectification. Softmax represents the classification layer activation function. LSTM layer is designed to carry out time sequence feature extraction, multilayer residual block is designed to carry out deep behavior feature extraction, the global average pooling layer is designed to

deal with the learning features, the features of each image as a pool area to perform the operation, and the output size equal to the number of feature maps classification layer number of neurons in the same number and behavior category [15–18].

2.5. Student Behavior Recognition Process

In order to solve the problem of difficult behavior feature extraction caused by the complex and changeable learning environment of students, a DRN-LSTM method for student behavior recognition is proposed in this paper. The student behavior recognition process based on DRN-LSTM is shown in Fig. 5. First, in the data processing stage, data sampling and data standardization are required, and then the proportion of training set, validation set and test set is divided. Finally, the data label is one-hot coded. In the model training phase, network parameters are initialized first, and then the pre-processed training set data is input into the improved deep residual network that integrates the LSTM layer and Dropout layer for training, and parameter adjustment is performed using the validation set. After the training, the behavior recognition model can be obtained. In the recognition stage, the test data is input into the behavior recognition model and the recognition results are output.

3. Experiments and analysis

Daily classroom videos of 5 classrooms are collected as data sources. The actions with recognition and classification in the video are manually screened and annotated. The original video is edited in 10s to generate the original video data set. Considering that students' common images in the classroom scene can be divided into bowed head, sideways, sitting up and raising hands, these four types of videos are selected from the original video database and segmented according to frames. Each image has a resolution of 256×256 pixels. Considering the balance of training data, the video library is screened and a total of 5500 images are finally obtained, including 1600 sitting, 1400 sideways, 1400 bowed head and 1100 hands raised.

A total of 4000 images are collected in the hand movement data set, including 3200 training images and 800 testing images. In order to realize the classification of hand movements, the data set of hand movements is divided into three subcategories, including 2400 conventional behavior images, 800 mobile phone playing images and 800 writing images.

This data set is input into the classification network model proposed above for training. The accuracy of the model on the test set is compared with the accuracy of the direct training of the images collected in this paper, the accuracy of the images processed by LSTM algorithm and the accuracy of the images processed by Residual network. The results are shown in Table 1.

Table 1. Comparison of accuracy

Model	Accuracy/%
LSTM	86.32
Residual network	88.45
DRN-LSTM	96.56

LSTM algorithm is used to extract the hand region, and prune the model based on the operation speed and the deployment of the model. First, the mean Average Precision mAP (mAP) is 0.8206 after normal training for the hand data sets. Then, global sparse training is used to perform attenuation with a learning rate of 0.1 at 0.7 and 0.9 stages of the total cycle, and the default scale parameter is 0.001.

After sparse training, channel pruning is carried out, and the width value of channel pruning is set to 0.84. The ratio of minimum number of maintained channels in each layer is 0.01. Then fine-tune the pruned model is used to improve the accuracy. Table 2 shows that the performance of the proposed network is greatly improved after pruning on this data set. The number of parameters in the model is 16.61%, and the compression ratio of the model is 83.4%. On Titian V, the processing time is reduced by 47.38%, while the maps of all classes remains basically unchanged. Therefore, the pruned model can be used as the reference network of the action classification module in the new framework of this paper.

After screening the processed hand area, the images of playing mobile phone and writing are obtained. The training data set is constructed. Hand behaviors are divided into three categories: playing with mobile phone, empty-handed and writing. So hand movement classifier is obtained by training in the network mentioned above. The parameters of the training are set to learn rate=10-4, epoch=200, and batch size=128. The entire network is trained by stochastic gradient descent. In the final test set, the detection accuracy of mobile phone playing and writing is 93.2% and 88.7% respectively.

In order to verify the accuracy and efficiency of classroom behavior recognition based on the combination of hand regional features and global motion features, and to prove the effectiveness of the proposed method, 200 groups of experiments are conducted on the newly recorded classroom video data set of ordinary classes. The experimental results of its final recognition are shown in Table 3. FPS (Frames Per Second) refers to the average number of image frames that a model can process per second. The experi-

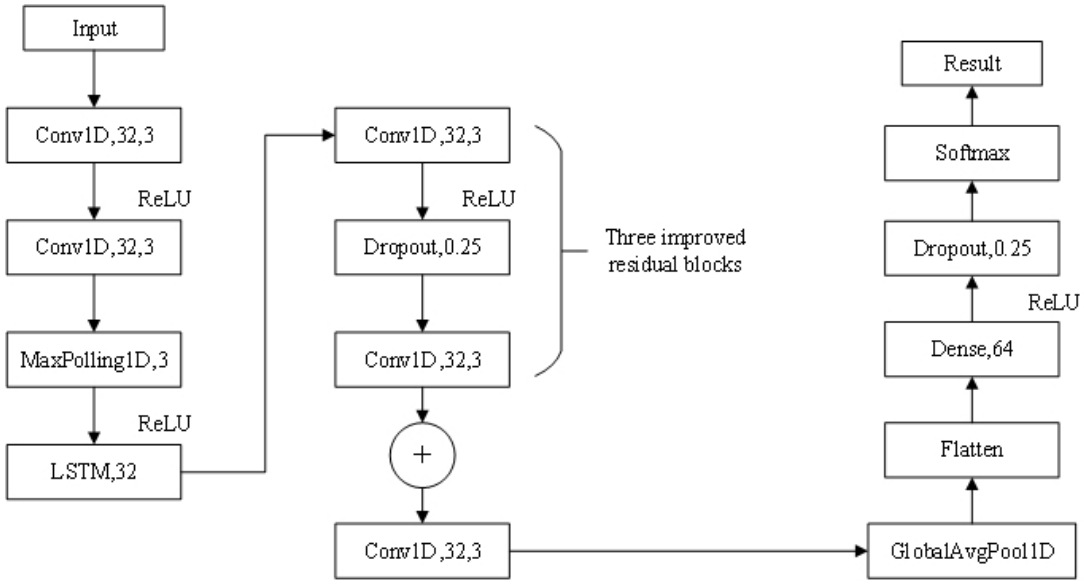

Fig. 4. Structure diagram of improved residual block

Table 2. Changes in parameters after pruning

Model	mAP/%	Parameter size/MB	Reasoning time/s
LSTM	0.8206	59.5	0.0122
DRN-LSTM	0.8241	9.4	0.0076

ment shows that the accuracy of behavior recognition based on the fusion of residual network depth feature information is improved compared with the image classification of hand region, and the student classroom behavior recognition algorithm adopted fully meets the requirements of accuracy and efficiency.

The data labels of classroom behavior recognition may have ambiguity, which brings great problems to the convergence of model training. The loss function represents the difference between the model's prediction label and the real label for a sample, which can measure the model's learning of the sample. Checking the sorting of loss function loss. If the loss value is large, the sample label may be mislabeled or the behavior of its own picture is ambiguous. The best way to deal with such data is to remove it manually. Thus, the interference of similar behaviors can be eliminated and the accuracy of recognition can be improved.

4. Conclusion

In this paper, deep learning and computer vision technology are used to study the method of students' classroom behavior recognition, and the data set of students' classroom behavior in real scenes is created. The model of student behavior classification is obtained by global posture recognition and local posture recognition. LSTM has the ability of short and long term memory, which can capture the time sequence information of students' behavior and obtain more abundant feature representation. The Dropout layer introduced into the residual block can discard some redundant information in the network, prevent the network from over-fitting, and improve the identification efficiency. The model is compressed to obtain an efficient and high-precision behavior recognition system. Through the test, the system has obtained good results, and can realize the automatic detection of classroom behavior, which is of great significance to measure students' learning commitment, teachers' optimization of teaching design and implementation of teaching intervention, and students' self-adaptive learning. In the future, graph convolutional neural network will be used to process the video time and the picture space dimension of each meal, so as to further improve the accuracy of class behavior classification of students.

References

[1] S. Li and T. Liu, (2021) *"Performance Prediction for Higher Education Students Using Deep Learning"* **Complexity *2021***: DOI: 10.1155/2021/9958203.

[2] S. Mitchell, C. West, and S. Lockwood, (2021) *"Community dental health coordinators' role in community-*

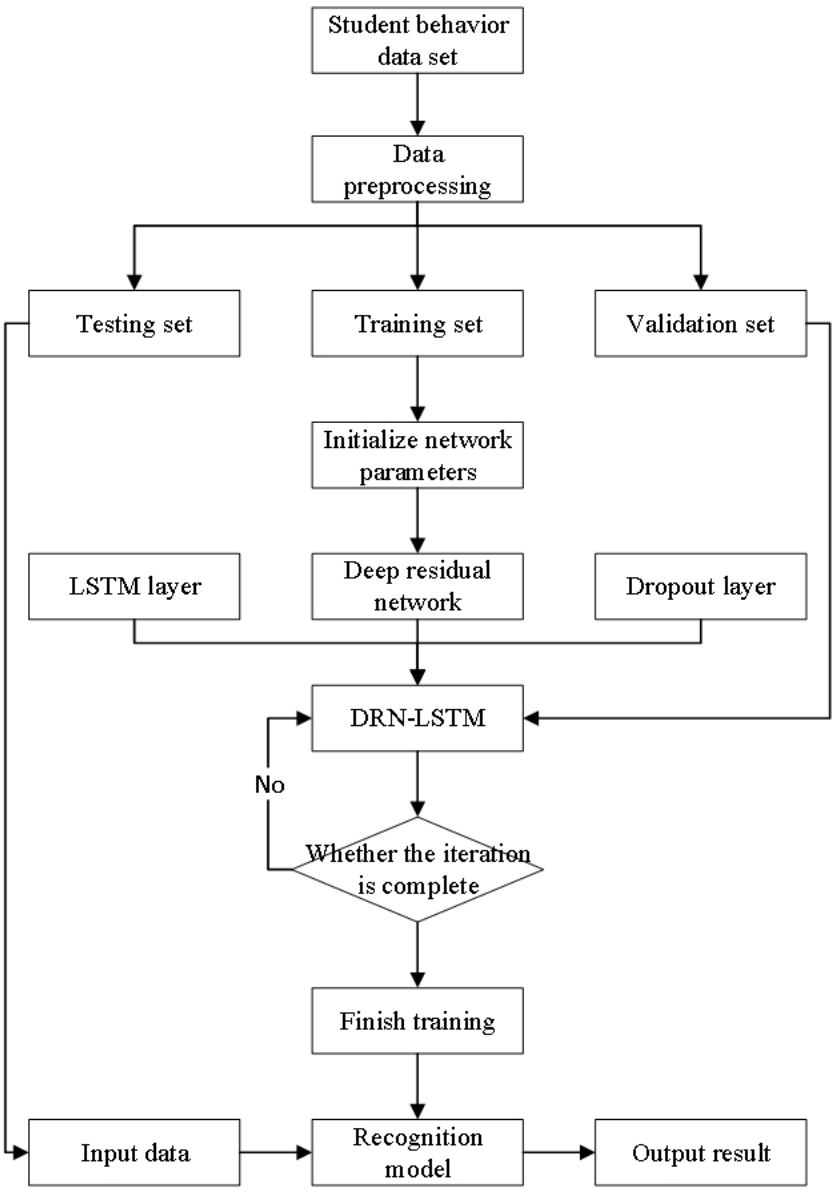

Fig. 5. Structure diagram of improved residual block

Table 3. Accuracy and efficiency results

Behaviour	Testing number	Correct recognition times	Accuracy/%	FPS
look down at the phone	100	95	95.0	13.9
Lower the head to write	100	96	96.0	10.5
Total number/Average	200	191	95.5	12.1

based pediatric dental education" **Journal of Dental Education** ***85***(7): 1217–1222. DOI: 10.1002/jdd.12593.

[3] E. Crisol Moya, V. Gámiz Sánchez, and M. Romero López, (2021) *"University students' emotions when using E-portfolios in virtual education environments"* **Sustainability (Switzerland)** ***13***(12): DOI: 10.3390/su13126973.

[4] A. Raheemullah, N. Andruska, M. Saeed, and P. Kumar, (2020) *"Improving Residency Education on Chronic Pain and Opioid Use Disorder: Evaluation of CDC Guideline-Based Education"* **Substance Use and Misuse** ***55***(4): 684–690. DOI: 10.1080/10826084.2019.1691600.

[5] N. Dash, M. Taha, S. Shorbagi, and M. Abdalla, (2022) *"Evaluation of the integration of social accountability values into medical education using a problem-based learning curriculum"* **BMC Medical Education** ***22***(1): DOI: 10.1186/s12909-022-03245-6.

[6] L. Lu, X. Meng, Z. Mao, and G. Karniadakis, (2021) *"DeepXDE: A deep learning library for solving differential equations"* **SIAM Review** ***63***(1): 208–228. DOI: 10.1137/19M1274067.

[7] M. Jiang and S. Yin, (2022) *"Facial expression recognition based on convolutional block attention module and multi-feature fusion"* **International Journal of Computational Vision and Robotics** ***1***(1): 1. DOI: 10.1504/IJCVR.2022.10044018.

[8] S. Zhu, Q. Zhang, W. Zhai, Z. Yuan, and C. Cai, (2021) *"Sensor deploying for damage identification of vibration isolator in floating-slab track using deep residual network"* **Measurement: Journal of the International Measurement Confederation** ***183***: DOI: 10.1016/j.measurement.2021.109801.

[9] R. Rajeswari, R. Ganeshan, B. Maram, and R. Cristin, (2021) *"Brain MRI Images Classifications with Deep Fuzzy Clustering and Deep Residual Network"* **International Journal of Computational Methods**: DOI: 10.1142/S021987622142007X.

[10] M. Josefsson, M. J. Daniels, and S. Pudas, (2020) *"A Bayesian semi-parametric approach for inference on the population partly conditional mean from longitudinal data with dropout"* **Biostatistics**: DOI: 10.1093/biostatistics/kxab012.

[11] Y. Ba and L. Qi, (2021) *"Construction of WeChat Mobile Teaching Platform in the Reform of Physical Education Teaching Strategy Based on Deep Neural Network"* **Mobile Information Systems** ***2021***: DOI: 10.1155/2021/3532963.

[12] A. Jisi and S. Yin, (2021) *"A new feature fusion network for student behavior recognition in education"* **Journal of Applied Science and Engineering (Taiwan)** ***24***(2): 133–140. DOI: 10.6180/jase.202104_24(2).0002.

[13] F. Kong, J. Li, B. Jiang, H. Wang, and H. Song, (2021) *"Integrated Generative Model for Industrial Anomaly Detection via Bi-directional LSTM and Attention Mechanism"* **IEEE Transactions on Industrial Informatics**: DOI: 10.1109/TII.2021.3078192.

[14] H. Nie, R. Peng, J. Ren, and Y. Qu, (2021) *"A through-focus scanning optical microscopy dimensional measurement method based on deep-learning classification model"* **Journal of Microscopy** ***283***(2): 117–126. DOI: 10.1111/jmi.13013.

[15] M. Narwaria, (2021) *"The Transition from White Box to Black Box: Challenges and Opportunities in Signal Processing Education"* **IEEE Signal Processing Magazine** ***38***(3): 163–173. DOI: 10.1109/MSP.2021.3050996.

[16] Z. Li, Z. Zheng, F. Lin, H. Leung, and Q. Li, (2019) *"Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN"* **Multimedia Tools and Applications** ***78***(14): 19587–19601. DOI: 10.1007/s11042-019-7356-3.

[17] D. Bai, S. Chen, and J. Yang, (2019) *"Upper Arm Motion High-Density sEMG Recognition Optimization Based on Spatial and Time-Frequency Domain Features"* **Journal of Healthcare Engineering** ***2019***: DOI: 10.1155/2019/3958029.

[18] X. Wang, S. Yin, K. Sun, H. Li, J. Liu, and S. Karim, (2020) *"GKFC-CNN: Modified gaussian kernel fuzzy C-means and convolutional neural network for apple segmentation and recognition"* **Journal of Applied Science and Engineering** ***23***(3): 555–562. DOI: 10.6180/jase.202009_23(3).0020.