Student Classroom Teaching Behavior Recognition Based on DSCNN Model in Intelligent Campus Education

Haiyu Zhang, Yang Li
School of Education Science, Minzu Normal University of Xingyi, Xingyi 562400, China
E-mail: 19885492560@163.com, 15510611181@163.com

Keywords: smart campus, DSCNN, students, behavior recognition, convolutional neural network

Received: February 23, 2024

Artificial intelligence technology is becoming increasingly popular. The introduction of this technology into classroom teaching becomes an important way to improve teaching quality. However, traditional methods for student behavior recognition suffer from low efficiency and insufficient accuracy. Therefore, a student classroom teaching behavior recognition scheme based on a dual stream convolutional neural network model was proposed. The research focused on the visual geometry group and the Res-Net method of convolutional neural networks and introduced knowledge distillation technology to optimize model efficiency. An attention mechanism combined with a dual stream convolutional neural network model was ultimately constructed to further improve the performance of the model. The results confirmed that the recognition accuracy of the model reached 88.1% on the UCF-101 data set and 89.4% on the STUDENT data set. The accuracy rates of classroom teaching behavior recognition for students using mobile phones, writing, chatting, raising hands, and sleeping were 97.0%, 87.9%, 90.7%, 89.2%, and 96.1%, respectively. The processing speed of this model on the UCF-101 and STUDENT data sets was more than twice and 1.5 times that of traditional DSCNN models, respectively. Therefore, the proposed attention mechanism combined with the dual stream convolutional neural network model has demonstrated excellent recognition ability. This study provides key technical support for the intelligent transformation of the education industry.

Povzetek: Članek obravnava prepoznavanje vedenja učencev med poukom z uporabo DSCNN modela in s tem podpira inteligentno preobrazbo izobraževanja.

1 Introduction

In recent years, artificial intelligence shows significant influence in various aspects of daily life. It has penetrated into every aspect of work and life, from speech recognition and image recognition to smart homes. In education, the application of artificial intelligence has also received widespread attention. Especially in the construction of smart campuses, its potential for improving teaching efficiency and quality is highly anticipated [1-3]. However, a major challenge is how to effectively apply these technologies to identify and analyze student classroom behavior. In traditional classroom teaching models, the identification and analysis of student behavior often rely on the teacher's intuitive observation. This method is not only time-consuming and labor-intensive, but also difficult to ensure the objectivity and consistency of the evaluation. With the classroom environment becoming increasingly dynamic and diverse, the limitations of traditional methods in terms of speed and accuracy become more apparent. For example, in a large classroom, it is difficult for teachers to simultaneously pay attention to the behavioral details of all students. Each student's behavioral characteristics and ways of expression are different. This diversity and complexity increase the difficulty of behavior recognition, especially when timely feedback and personalized teaching strategy adjustments are needed [4-6]. In addition, traditional image recognition techniques mainly target static images. Due to the dynamic video streaming, the processing speed and adaptability face challenges, which are particularly important in classroom teaching behavior analysis. Therefore, the study proposes a new solution. The study first explores the Visual Geometry Group (VGG) and Res-Net of Convolutional Neural Network (CNN). On this basis, Dual Stream Convolutional Neural Networks (DSCNN) are further developed, and knowledge distillation technology is introduced to improve the efficiency of this model. In addition, the study also integrates attention mechanisms into DSCNN to further enhance the performance of the model. The overall structure of the study consists of four parts. Firstly, the relevant research achievements and shortcomings of CNN and behavior recognition at home and abroad are summarized. Secondly, DSCNN is proposed and knowledge distillation techniques are introduced to improve efficiency. In addition, the study also constructs a DSCNN with integrated attention mechanisms to further enhance performance. The third part of the research experiment compares and analyzes the proposed model. Finally, the experimental results are summarized, the shortcomings of the research are pointed out, and future research directions are proposed.

2 Research backgrounds

CNN, as an advanced image processing technology, becomes a key player in the behavior recognition. This method effectively recognizes and analyzes behavioral patterns in images by simulating the processing mechanisms of the human visual system. On a global scale, many scholars are exploring the application of CNN in behavior recognition to achieve more efficient and accurate recognition results in multiple fields [7]. The following introduces some related research by scientists and scholars, and the core content of their research is shown in Table 1.

Table 1: Results and key content of references

Reference	Title	Key content	Finding
[8]	Human activity recognition and embedded application based on convolutional neural network	The author designed a hybrid improved CNN and random gradient descent method to extract human activity features from images	The performance of the improved algorithm in indicators such as loss function value and confusion matrix is better than that of the comparative model.
[9]	Shallow convolutional neural networks for human activity recognition using wearable sensors	The author used various common machine learning frameworks to compare their designed Live One Subject Out Cross Validation (LOS OCV)	The method designed by the author performs well in dealing with new theme images that are not present in the data set.
[10]	3D dual-stream convolutional neural networks with simple recurrent unit network: A new framework for action recognition	The author proposed an improved image recognition algorithm based on 3D CNN and dual stream deep fusion framework.	The designed algorithm can effectively identify video information from the selected data set.
[11]	A dual stream model for activity recognition: exploiting residual-CNN with transfer learning	The author constructed an improved image pre-training model using deep residual networks	The image quality annotated by the design model is higher than that of the pre-improved model
[12]	Learning spatio-temporal representations with a dual-stream 3-D residual network for non-driving activity recognition	The author used a spatial residual network for structural optimization to enhance the spatial feature learning ability of the model.	The model designed by the author performs better than all comparison algorithms on data sets with complex spatial structures.
[13]	Dual-stream structured graph convolution network for skeleton-based action recognition	The author designed a bone action recognition model based on CNN.	The designed model performs significantly better than the basic algorithm before improvement on five multimodal benchmark data sets.
[14]	Linear dynamical systems approach for human action recognition with dual-stream deep features	The author improved the CNN algorithm using linear dynamic systems and pre-training methods and applied it to human action recognition.	The test results on three public data sets show that the designed algorithm has a higher recognition accuracy than the latest image recognition algorithms.
[15]	Hybrid handcrafted and learned feature framework for human action recognition	The author combined discrete wavelet transform technology with dense trajectory models to design an accelerated action recognition method.	The test results show that the design model can accurately recognize the acceleration actions of objects.

Specifically, Y. Xu et al. proposed CNN in deep learning, which effectively recognized behaviors in human daily life. The network parameters were optimized using a random gradient descent algorithm. The training model was compressed on the STM32CubeMX-AI platform. Six common activities were identified using neural networks on embedded devices: sitting, standing, walking, jogging, going upstairs, and going downstairs. These results confirmed that the model had accurate recognition ability for these daily activities [8]. Huang et al. proposed a shallow CNN suitable for human activity recognition. This method utilized graph neural networks to achieve interaction between different channels. This method promoted the effective deployment of lightweight deep models by eliminating redundant information between channels. Experiments on multiple standard data sets such as UCI-HAR showed that this method enabled the shallow CNN to aggregate more useful information. The performance of this method surpassed the performance of multiple baseline deep networks [9]. Zhao et al. proposed a framework that combined dual stream deep fusion to effectively utilize the long-term information of videos. A spatiotemporal feature stream containing time series was obtained by preprocessing the video into static frames and optical flow maps and applying 3D CNN. Subsequently, these feature flows learned time-series features through a simplified recurrent network and were classified using a SoftMax classifier. These results confirmed that the model outperformed existing methods in terms of performance [10]. Singh et al. proposed a dual stream pre-training model by fusing the deep residual networks to perform behavior recognition on video streams. The standard video action benchmarks of UCF-101 and HMB-51 were utilized for further training and evaluation. The performance of depth-based residual network variants was analyzed in the experiment. This method not only provided competitive results, but also better utilized pre-trained models and annotated image data [11].

Yang et al. proposed a dual stream 3D residual network aimed at enhancing the learning of spatiotemporal features and improving behavior recognition performance. This network adopted a parallel dual stream structure, focusing on learning short-term spatial representation and small-scale temporal representation, respectively. Adual feed driver behavior monitoring framework was constructed based on this network, which could classify driver head and hand movements. The results indicated that this method was at least 5% higher than the other three advanced methods [12]. Xu et al. proposed a new method called dual flow structure graph convolutional network to solve the skeleton-based behavior recognition problem. This method integrated the spatiotemporal coordinates and appearance context of skeletal joints into the graph convolutional learning. The dynamic interaction of different body parts was simulated by constructing local inter graph. The end-to-end prediction of human behavior categories was performed by integrating two graph convolutional response flows. These results confirmed that the framework performed well on behavior recognition tasks [13]. Du et al. proposed a CNN-based dual stream deep feature extraction framework for processing temporal relationships. This framework modeled the temporal relationship between adjacent video slices through a linear dynamic system. They were manifested as linear dynamic background features and linear dynamic differential features, respectively. Experiments on multiple data sets validated the effectiveness and robustness of this method, demonstrating its competitive performance with current advanced methods [14]. Zhang et al. proposed a DSCNN-RNN that combined learned features with handcrafted features to meet strict spatiotemporal analysis requirements. This hybrid feature framework generated efficient FISHER vectors through a novel temporal feature packet scheme. This framework encoded video events and accelerated action recognition in real-world applications. The evaluation of the design indicated that its recognition performance was superior to existing benchmark methods [15].

In summary, although previous research has made significant progress in image and behavior recognition, there are still many limitations. For example, the application of traditional models in complex scenes is limited. Because these studies only focus on analyzing static images rather than dynamic videos and they fail to effectively combine spatial and temporal information features. To this end, the study proposes CNN's VGG and Res-Net network methods and further develops DSCNN to improve the accuracy of student behavior recognition. This study introduces knowledge distillation technology and integrates attention mechanisms to improve speed and efficiency. This study provides innovative solutions for improving the effectiveness of smart campus education, especially in enhancing the accuracy of behavior recognition and enhancing the quality of teaching interaction.

3 Student classroom teaching behavior recognition based on dscnn model in intelligent campus education

The study preliminarily explored the VGG and Res-Net methods of CNN. DSCNN is further developed to improve the accuracy of student behavior recognition. Although this model is enhanced in functionality, it still faces challenges in processing speed, especially in extracting optical flow from dynamic videos. To address this issue, knowledge distillation technology is introduced to optimize the efficiency of the model. In addition, a DSCNN that combines attention mechanisms is formed to

further improve the performance of the model.

3.1 Construction of VGG and Res-Net network methods based on convolutional neural networks

The application of CNN is increasing in many application fields of behavior recognition. CNN mainly includes input, output, convolutional, and pooling layers in Figure 1. Convolutional and pooling layers are closely connected and arranged alternately. CNN converts the features of an image into the output of the network. Then these features are transmitted to the fully connected layer to achieve effective image classification.

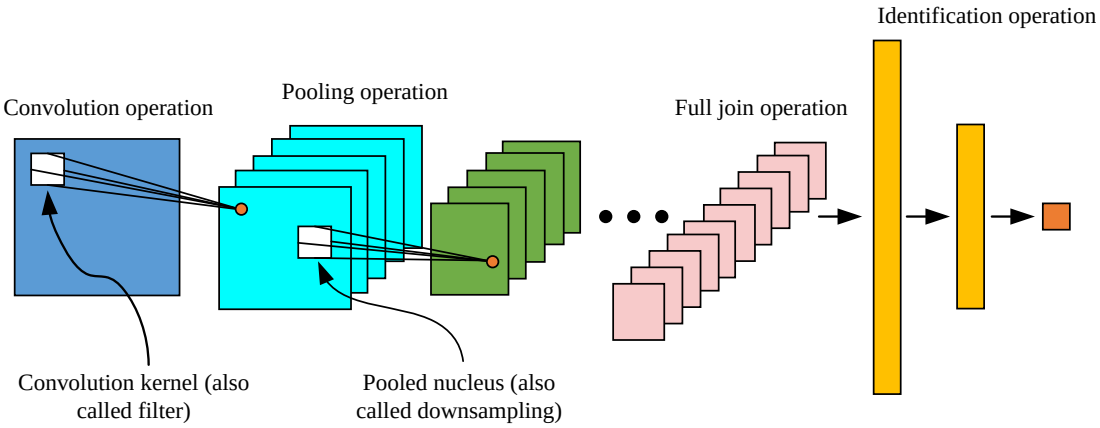


Figure 1: CNN structure diagram

The convolutional layer structure consists of a set of convolutional kernels, each responsible for capturing different features of the image and forming a unique feature map. These convolutional kernels have specialized weights and bias values, and their working principle is similar to that of neurons in traditional neural networks [16-18]. Within the convolutional layer, neurons are connected to multiple neurons in the upper layer. The range of neuron connections is determined by the size of the convolutional kernel, forming a receptive field. The key parameters that affect the performance of convolutional layers include the size of the convolutional kernel, the step size of movement, and the edge filling method. The pooling layer, located between continuous convolutional layers, mainly functions to reduce parameters by reducing the size of feature maps. Pooling operations include max pooling and average pooling. These operations are performed without parameters and are usually performed at specific window sizes to improve processing efficiency and control overfitting. The specific calculation process is shown in equation (1).

In neural networks, the nonlinear activation functions are introduced to handle complex models. These functions, such as Sigmoid and ReLU, enable neurons to convert input signals into outputs and pass them on to the next layer. The fully connected layer closely follows the convolutional and pooling layers. The main function of the fully connected layer is to integrate the features extracted by the previous layers. The fully connected layer achieves higher-level learning through nonlinear fusion. Each neuron establishes connections with all neurons in the fully connected layer, thereby integrating information from each layer. Next, the VGG-Net model is proposed. VGG-Net consists of five layers of convolution and two layers of pooling, plus a fully connected layer. The network depth is enhanced and performance is improved by using a small 3×3 convolution kernel [19-21]. VGG-Net demonstrates excellent generalization ability on multiple data sets through this design, especially in extracting deep features from images. Figure 2 is a schematic diagram of the VGG-Net network structure.

$$\begin{cases} a_i \in \{0,1\}^k, a_{i,j} = 1 if \quad j = \arg\min \| x_i - d_k \| \\ h_m = \dfrac{1}{|N_m|} \sum_{i \in N_m} a_i \\ a_i = \arg\min L(a, D) \square \| x_i - Da \| + \gamma \| a \| \\ h_{m,j} = \max_{i \in N_m} a_{i,j}, \ for \quad j = 1, ...k \end{cases} \tag{1}$$

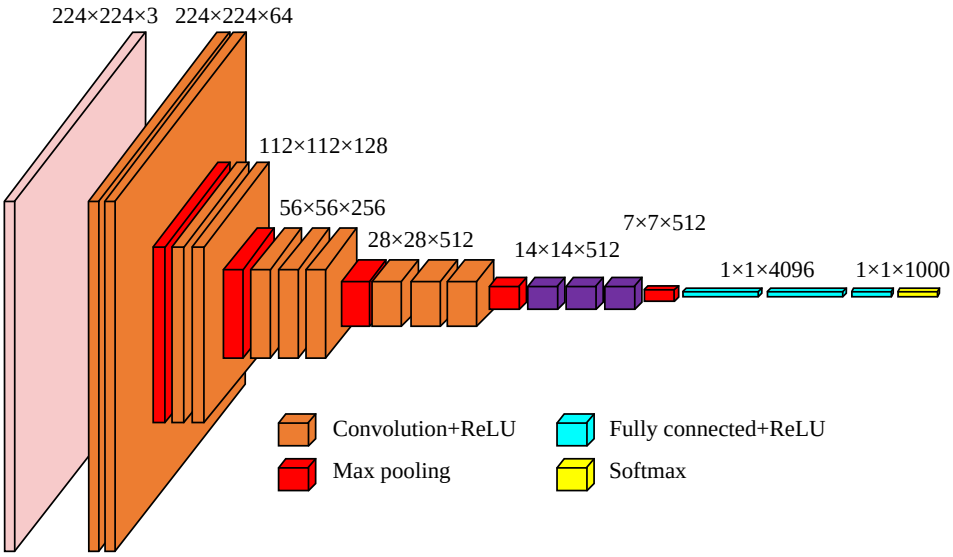


Figure 2: VGG-Net network structure diagram

An increase in the layers in deep neural networks may lead to increased training difficulty, therefore a Res-Net residual network model is proposed. The design of this model allows the network to optimize the training by learning the differences between input and output. Learning residual $F(x)-H(x)-x$ in the Res-Net architecture simplifies the learning objectives. This method reduces the complexity of model learning, especially when dealing with deeper networks. Res-Net significantly improves the training efficiency and accuracy of deep networks through this structure. The specific calculation of residual units is represented by equation (2).

$$\begin{cases} y_l = h(x_l) + F(x_l, W_l) \\ x_{l+1} = f(y_l) \end{cases} \quad (2)$$

In equation (2), f represents the ReLU activation mechanism. F refers to the residual mapping process. $h(x_l)$ is used to describe identity transformations. x_l represents the initial data entering the residual unit. x_{l+1} is the processed output data. Based on these definitions, the transformation of features between shallow and deep networks is represented by equation (3).

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \quad (3)$$

As mentioned above, the construction of Res-Net residual network is based on the optimization of VGG-Net architecture. The core innovation of Res-Net lies in the introduction of a fast connection mechanism and the addition of residual units to enhance network performance. Res-Net specifically adopts advanced convolution strategies to implement down-sampling operations. A key point of Res-Net design philosophy is that as the feature map size is reduced to half of its original size. Meanwhile, the number of feature maps doubles accordingly to maintain the comprehensive processing capability of the network.

3.2 Identification of student classroom teaching behavior based on DSCNN

The recognition of student classroom behavior is particularly important in intelligent campus education. DSCNN is proposed to further improve the accuracy of student behavior recognition. The innovation of DSCNN lies in the parallel analysis of spatial and temporal information in videos, which can effectively improve the accuracy of behavioral analysis. However, the model faces challenges in processing speed, especially when extracting optical flow from dynamic videos. Therefore, the study introduces knowledge distillation technology, which transfers key knowledge from the teacher network to a simpler student network. This not only optimizes the deployment process of behavior recognition, but also reduces the need for complex computing. The knowledge distillation is similar to teaching activities. The teacher network has a huge structure and parameter quantity, while the student network is relatively simplified. Student networks can achieve considerable recognition results while maintaining a small scale through the guidance of teacher networks. Knowledge distillation essentially combines transfer learning and model compression, which can remove unnecessary complexity while retaining decisive predictive information. Efficient and accurate knowledge transfer can be achieved through this distillation process. Therefore, the student networks can maintain high performance even under a simplified architecture in Figure 3 [22-24].

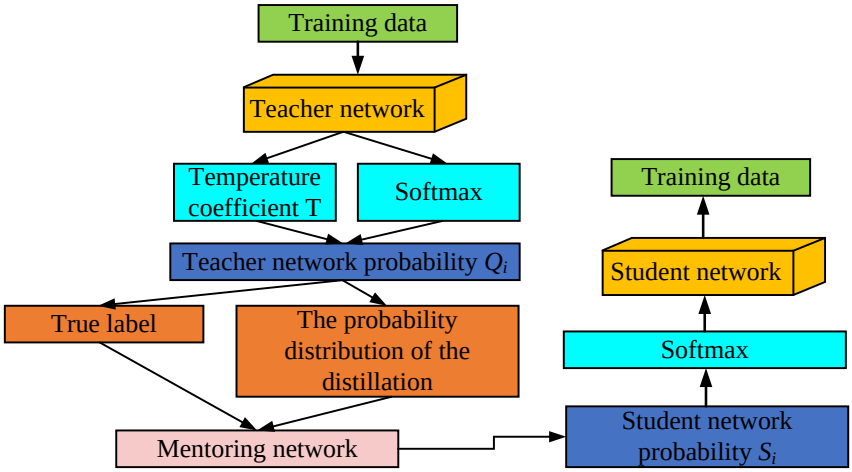


Figure 3: Schematic diagram of knowledge distillation architecture

In Figure 3, the teacher network optimizes the predicted probability distribution by adjusting a specific parameter T. Through this method, the student network maintains effective recognition ability while simplifying its structure. In this framework, Q_i refers to the prediction probability of the teacher network for each category. S_i is the class prediction probability formed by the student network under the guidance of the teacher network. The Q_i of teacher networks is seen as a soft target for richer information. Compared to hard targets, soft targets can provide more information on the relationships between categories. The specific calculation of Q_i is represented by equation (4).

$$Q_i = \frac{\exp(z_i / T)}{\sum_j \exp(z_j / T)} \tag{4}$$

In equation (4), z_i is the probability score for each category. When adjusting the parameter T to a higher value, the probability distribution of the obtained soft target becomes more average. In standard training scenarios, i.e. using hard targets, T is set to 1. The training of student networks can be effectively monitored through KL divergence, represented by equation (5).

$$KL(Q_i \| P_i) = \sum Q_i \log \frac{Q_i}{P_i} \tag{5}$$

In the knowledge distillation, the similarity between teacher and student networks is measured by KL divergence. The decrease of KL indicates that the probability distribution of the two networks tends to be consistent. A cross entropy loss function is used to further supervise the training of student networks, represented by equation (6).

$$L_{CrossEntropy}(y, \hat{y}) = -[y \log \hat{y} + (1-y)\log(1-\hat{y})] \tag{6}$$

In equation (6), y and $\hat{y}$ represent the true and predicted values. It is crucial to correctly extract features in the task of teaching behavior recognition in student classrooms, as it directly affects the accuracy of recognition results. Firstly, a teacher network with optical flow as input is developed. Subsequently, the network is used to guide another student network using RGB images as input. This method allows student networks to learn relevant features even without performing optical flow calculations [25-27]. The study introduces motion simulation RGB stream and motion enhanced RGB stream. The motion enhanced RGB flow combines mean square error and cross entropy loss to transfer knowledge between the teacher and student networks in Figure 4.

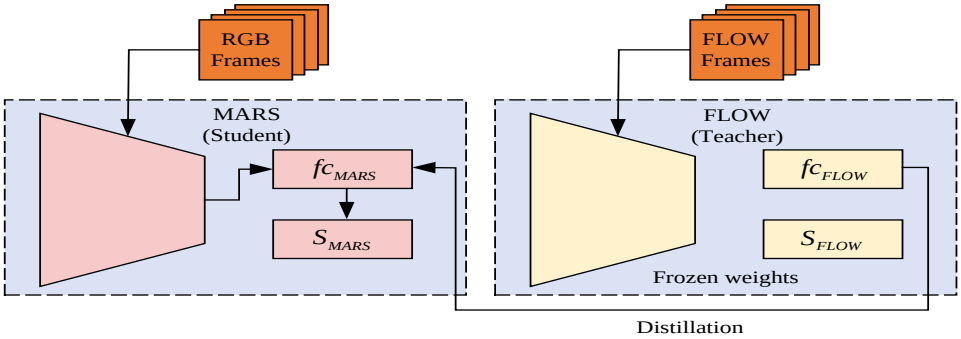


Figure 4: Motion augmented RGB stream network structure diagram

The loss calculation of the entire network depends on the backpropagation algorithm in the implementation of motion-enhanced RGB streams. A linear mixture combining mean square error and cross-entropy loss is mainly used to evaluate the difference between the predicted and actual values of the model, represented by equation (7).

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

In equation (7), N refers to the number of samples. In motion-enhanced RGB streams, the loss calculation uses a combination of two loss functions, namely mean square error and cross-entropy loss, represented by equation (8).

$$L_{MARS} = L_{CrossEntropy}(S_{MARS}, \hat{y}) + a \| f_{cMARS} - f_{cFLOW} \|^2 \quad (8)$$

In equation (8), S_{MARS} is the predicted probability output of the motion-enhanced RGB stream processed by Softmax. f_{cMARS} and f_{cFLOW} represent the feature outputs of motion enhanced RGB flow and optical flow network, respectively. a refers to adjusting the weights of motion features. The network feature learning adopts a hierarchical average absolute error loss function. Therefore, the deep information of the trained optical flow model can be extracted. This loss function is represented by equation (9).

$$L_{MAE} = \frac{1}{n-1}\sum_{i=1}^{n-1} \| f_{DS(i)} - f_{FLOW(i)} \|_1 \quad (9)$$

In equation (9), i refers to the hierarchy in the network. $f_{DS(i)}$ and $f_{FLOW(i)}$ are the feature outputs of the dual acting flow and optical flow networks at the i th layer, respectively. By using this equation, each convolutional layer of the dual action flow can match the corresponding layer's features of the optical flow network during training and backpropagation. This effectively mimics the motion information captured by optical flow networks.

3.3 Recognition of student classroom teaching behavior using attention mechanism combined with DSCNN model

Two independent networks are combined for processing RGB and optical flow images in advanced research on human behavior recognition. Ultimately, the outputs of these networks are integrated to obtain more accurate recognition results. Although DSCNN has significantly outperformed traditional models in this field, the effectiveness of traditional DSCNN is becoming limited as the demand for recognition accuracy increases. Attention mechanism is introduced in the study to improve the effectiveness. Attention mechanism is initially applied in natural language processing and is now effectively integrated into DSCNN. Based on the previous two sections, the study further proposes the combination of attention mechanism and DSCNN. This model can focus on key features more effectively while ignoring irrelevant information by adaptively adjusting the weights of spatial and temporal feature vectors. Therefore, the performance of the model can be improved. DSCNN is jointly constructed by spatial flow CNN and temporal flow CNN. Space networks process video frame images to extract spatial attributes of actions. Time networks preprocess optical flow images to capture temporal features of actions. The features extracted by the two networks are processed using the Softmax function to calculate the final classification score. The scores of these two streams are fused to form the final recognition output of the model by training a classifier or using a weighted average method. Figure 5 shows the structure of DSCNN.

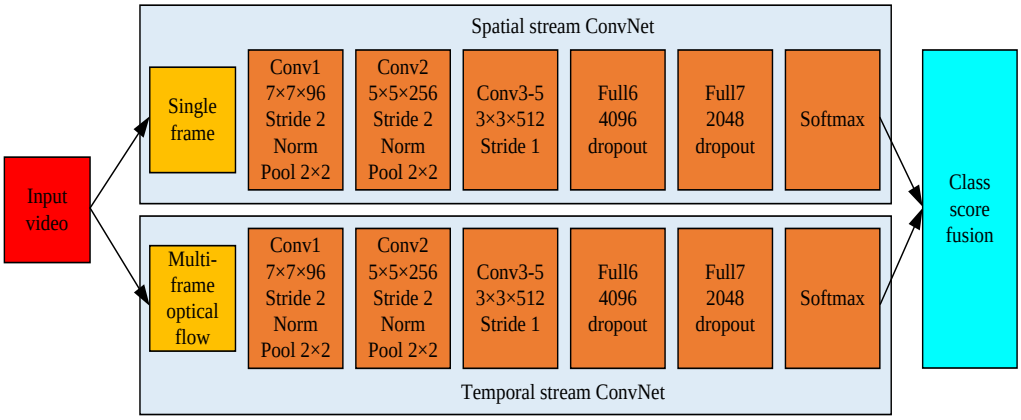


Figure 5: DSCNN model structure frame diagram

In the research of student classroom behavior recognition, the spatial networks use standard CNNs such as VGG-Net to process single-frame RGB images, with image sizes set to 224×224×3. The spatial network is responsible for analyzing these static images. The temporal network captures motion information by analyzing the changes between consecutive frames. Time networks extract motion patterns by comparing consecutive frames for dynamic classroom teaching behaviors of students. Deep learning technology simplifies the calculation of optical flow maps. First, deep CNN is used to extract features, then feature matching is performed, and finally optical flow information is obtained. This CNN-based method involves multiple steps, such as feature extraction and post-processing of optical flow features. Furthermore, this layer enhances its classification ability through the Dense Net structure. Multi-layer networks are used for optical flow refinement processing. The parameter learning process of the entire model is represented by equation (10) [28-30].

$$y = \sum_{\beta} \square_{\beta} \sum_{\Omega} | \omega_{\Phi}^{\beta}(\Omega) - \omega_{GT}^{\beta}(\Omega) |_2 \quad (10)$$

In equation (10), y represents model parameters, β is the weight factor at the pyramid level. In the optical flow feature extraction network, the size and direction of the flow vector are first determined. These vectors contain two components in equation (11), U_x and U_y, representing the displacement in the horizontal and vertical directions, respectively.

$$\gamma = \frac{\sqrt{U_x^2 + U_y^2}}{Max(\gamma)}, \psi = \arctan(\frac{U_y}{U_x}) \quad (11)$$

In equation (11), $Max(\gamma)$ standardizes the intensity of exercise. The time extracted from the features is input into the pulse neural network. The integration trigger or leakage integration trigger neurons are used for numerical generation. Pulse neural networks include hidden layers and Softmax-based classification layers, trained using various optimization techniques. A pulse neural network consists of an input layer, a hidden layer, and an output layer, which generates an output at time point p, represented by equation (12).

$$\begin{cases} X^p = [x_1^p, x_2^p, ..., x_\lambda^p] \\ V^p = V_{rest} + \sum_{i=1}^{\lambda} (W_I^T X_i^p + b_i) \\ V_p = V_{p-1} - \gamma (V_{p-1} - V_{rest}) \\ T^p = T^{p-1} - \begin{cases} T_{decay} * (T^{p-1} + T_{plus}), neuron \quad spikes \\ T_{decay} * T^{p-1}, elsewhere \end{cases} \end{cases} \quad (12)$$

In equation (12), V_{rest} is the baseline potential of the neuron. V_p means that the membrane potential of neurons decays proportionally by γ. The peak neuron T^p in the hidden layer resets to the baseline potential after activity. Furthermore, DSCNN incorporates attention mechanism and constructs a DSCNN based on it in Figure 6. This model processes student behavior videos and divides them into a single RGB image and a stacked optical flow map of 10 consecutive frames, with a unified size of 224 × 224. The core network structure includes pre-trained VGG-Net and attention modules.

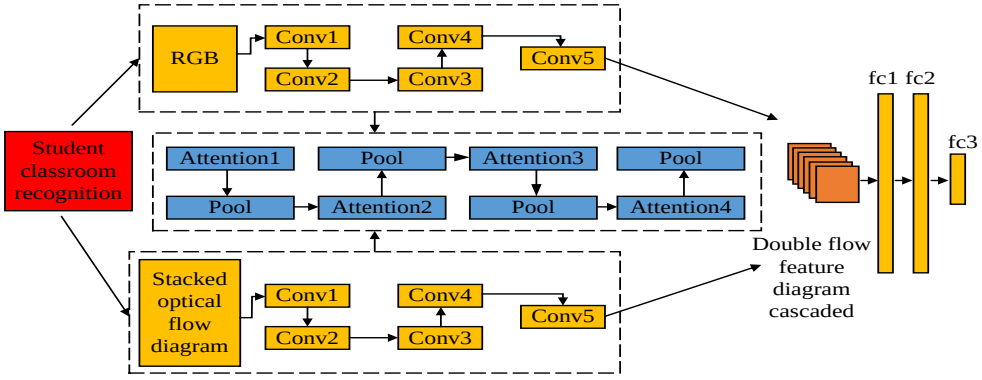


Figure 6: DSCNN combined with attention mechanism

In Figure 6, due to the lack of interaction and focus area filtering between two independent networks, attention mechanisms are introduced in this study. Attention mechanism achieves effective filtering of key regions by embedding attention matrices between convolutional layers. It consists of four matrices and three pooling layers. Each convolutional layer is followed by an attention matrix, which is followed by a pooling layer to reduce parameters and perform filtering. When implementing, the weight of the first attention matrix is first calculated. Then the remaining matrices are obtained through three pooling operations. The spatial and temporal networks process images of size 224 × 224 × 3 and output feature maps. Then a cascaded feature map of size 112 × 112 × 128 is obtained by combining the first layer outputs of the two networks at the channel level. Finally, attention weights are obtained through fully connected layers, represented by equation (13).

$$\begin{cases} G_{IJ} = \tanh(W_1^t H_{i,j} + B_1) \\ G_{IJ} = \dfrac{e^{g^{i,j}}}{\sum_{i=1}^{112}\sum_{j=}^{112} e^{g^{i,j}}} \end{cases} \tag{13}$$

In equation (13), B_1 is the bias term. g_{ij} represents the original attention weight. G_{IJ} means normalization through the Softmax function. Figure 7 shows the implementation of the attention mechanism. After determining the first attention weight, corresponding attention weights are applied to the output feature map of the second convolutional layer. A weighted feature map is formed as input for the next convolutional layer. Similarly, the first attention weight matrix is pooled to generate the second, third, and fourth attention weight matrices. The size of each attention weight matrix is halved in sequence, with values of 56 × 56, 28 × 28, and 14 × 14, respectively. These weighted feature maps are ultimately fed into the fully connected layer to identify behavioral actions. The four attention mechanism modules introduced optimize the learning process and eliminate interfering information and strengthen the correlation between time and space networks. This improves the accuracy of identifying student classroom teaching behaviors.

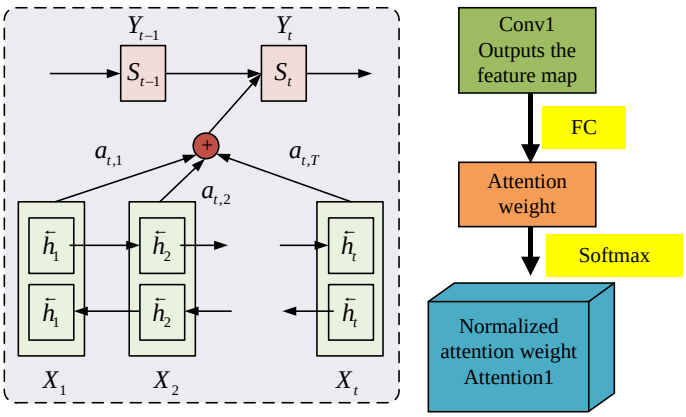


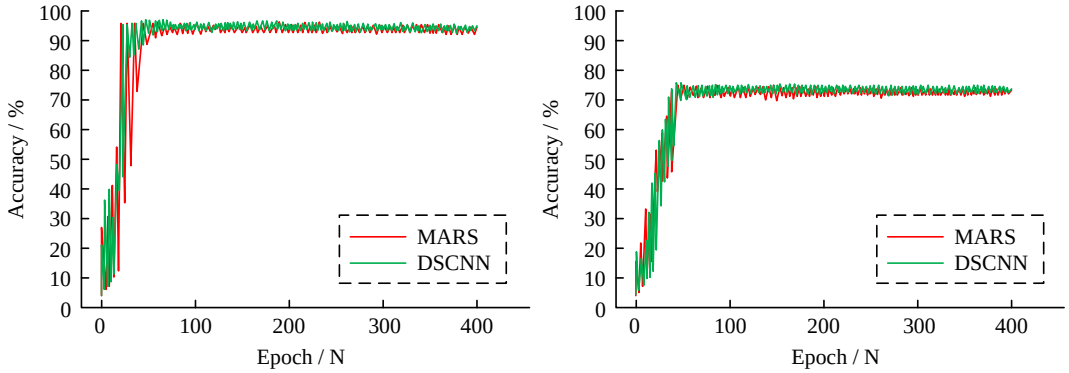
Figure 7: The realization process of attention mechanism

4 Analysis of student classroom teaching behavior recognition results based on DSCNN Model

First, the study introduced three different data sets and elaborated on the network configuration used in this experiment. Next, performance evaluation of DSCNN and motion enhanced RGB streaming network architecture was conducted. Visualization analysis of intermediate layer features was conducted. Subsequently, the study identified and analyzed the designated categories of student classroom teaching behavior on the UCF-101 and STUDENT data sets. Finally, a comparative analysis was conducted on the performance of the attention mechanism combined with DSCNN and other existing models to highlight the advantages and characteristics.

4.1 Analysis of motion enhanced RGB stream results based on DSCNN

This research experiment used three data sets, namely UCF-101, HMDB-51, and a self-built data set. UCF-101 is a publicly available data set widely used in behavior recognition research, which includes 101 different action categories and involves approximately 13000 videos from real-life scenarios. These videos showcase diverse scenes from daily life to specific activities, covering changes in the scale, appearance, posture, and lighting conditions of various objects. HMDB-51 is a relatively small publicly available data set, containing 6849 video clips distributed across 51 action categories. These videos are mainly annotated manually, involving various categories, such as facial behavior, body movements, and interaction with objects. In response to the shortcomings of these two data sets in covering student classroom teaching behavior, the STUDENT data set is created. STUDENT focuses on five classroom behaviors such as using a mobile phone, writing, raising hands, talking, and sleeping. These videos are shot in a real classroom environment, reflecting different perspectives and background settings. In terms of training DSCNN, both spatial and temporal networks were set with a Batch size of 35, an Epoch range of 100 to 500, an initial learning rate of 0.001. The learning rate was adjusted to one tenth of the original after every 50 Epochs. To ensure the performance of the proposed network architecture, these two models were trained on two publicly available data sets, UCF-101 and HMDB-51. Figure 8 shows the results after training with 400 Epochs.



(a) UCT-101 (b) HMDB-51

Figure 8: Validity result of Dual-actionstream structure

In the UCT-101 data set in Figure 8 (a), the accuracy of the MARS network reached 0.946, while the accuracy of the DSCNN was 0.952. In the HMDB-51 data set in Figure 8 (b), the accuracy of the MARS network was 0.723, while the DSCNN achieved an accuracy of 73.7. These results not only demonstrated the effectiveness of the proposed network architecture in complex scene behavior recognition, but also demonstrated the advantages of DSCNN in recognition accuracy. This indicated that the proposed model had practical application value in learning the basic features and their interrelationships of network layers. A visualization analysis of intermediate layer features was conducted to explore this point in depth. Ten behavior categories were selected. The feature representations of baseline network and DSCNN were compared, and they were mapped to the feature space in Figure 9.

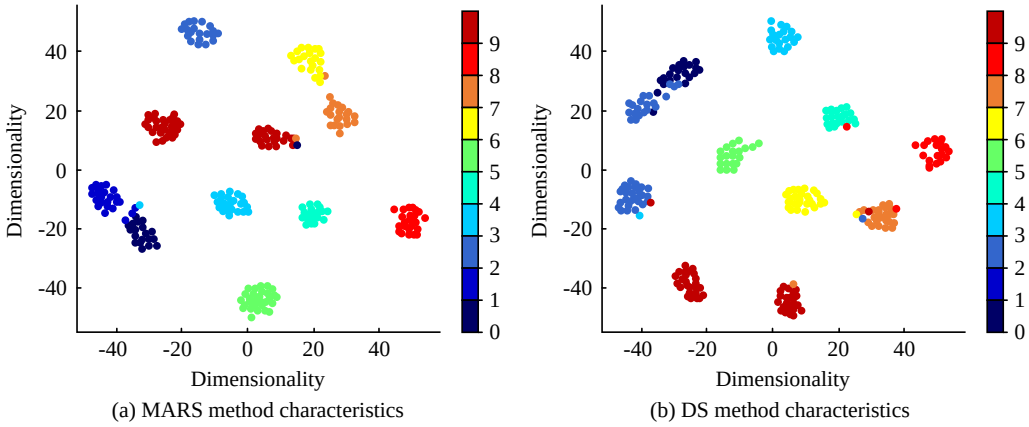


Figure 9: MARS and DS feature space results

Figures 9 (a) and 9 (b) show the feature maps of MARS network and DS network in the final layer, respectively. Each point represents the feature representation of a behavioral video clip sequence after network processing. The coordinate axis represents the two most significant feature dimensions after dimensionality reduction processing. Points of the same color belong to the same behavioral category. The aggregation and separation of points from other categories reflect the distinguishing ability of features. The results indicated that the proposed model was slightly better than the baseline model in feature learning.

4.2 Analysis of student classroom teaching recognition results using attention mechanism combined with DSCNN

This study evaluated the combination of DSCNN and attention mechanism on three different data sets. They were UCF-101, HMDB-51, and STUDENT data sets specifically designed for student classroom teaching behavior. The experiment included long-term testing of 1200 Epochs. Firstly, a series of preliminary experiments were conducted to confirm the performance of the model in behavioral feature extraction and classification. This experiment was mainly conducted on the UCF-101 and STUDENT data sets to identify the selected categories of student classroom teaching behavior. This experiment aimed to compare the performance of DSCNN with attention mechanism in terms of accuracy. Figure 10 shows the comparison of average recognition rates between two models.

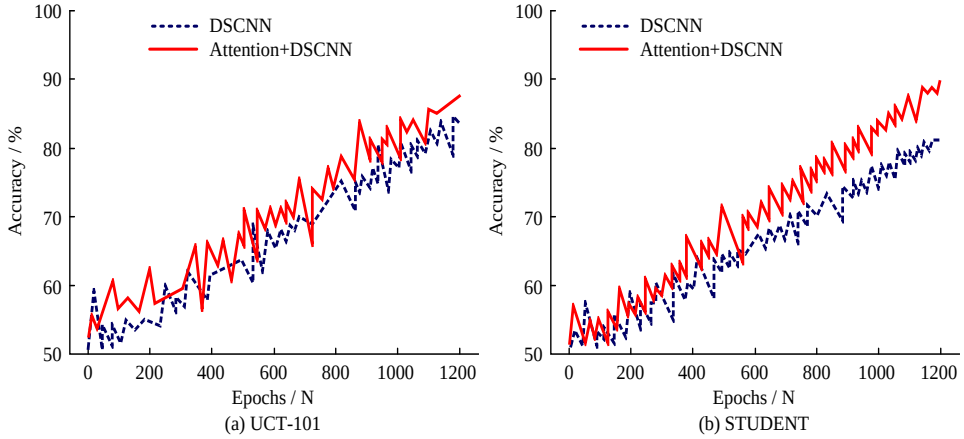


Figure 10: Comparison of the average recognition rate of two models

In the UCF-101 data set of Figure 10 (a), the recognition accuracy of DSCNN reached 84.0%, while the recognition accuracy of DSCNN with attention mechanism increased to 88.1%. In the STUDENT data set of Figure 10 (b), the recognition accuracy of DSCNN was 80.9%. The attention mechanism DSCNN achieved an accuracy of 89.4%, an increase of 8.5 percentage points. The results confirmed that DSCNN with added attention mechanism exhibited higher accuracy in student classroom teaching behavior recognition tasks,

confirming its superiority. Next, in the STUDENT data set, common student behaviors in the classroom, such as using mobile phones, writing, raising hands, talking, and sleeping, were represented as categories 1 to 5, respectively. Figure 11 shows the recognition performance of attention mechanism combined with DSCNN.

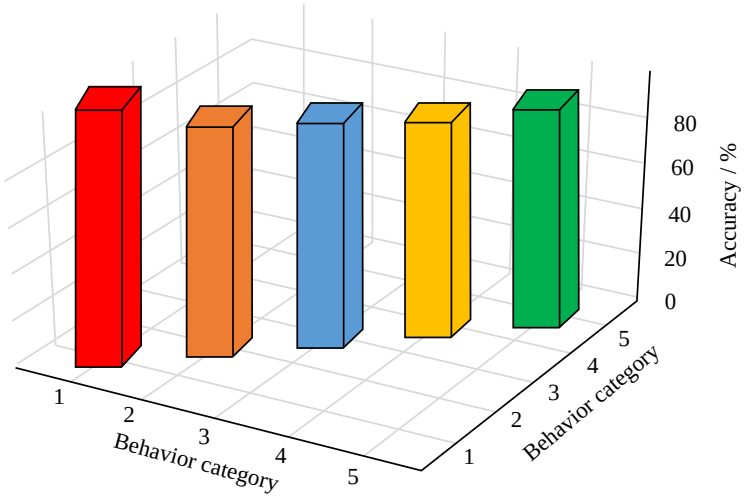


Figure 11: Identification of students' classroom teaching behavior

From Figure 11, the accuracy rate for recognizing the behavior of using a mobile phone was 97.0%, while the accuracy rate for recognizing writing actions was 87.9%. In addition, an accuracy rate of 90.7%was achieved in detecting conversational behavior, 89.2% in recognizing hand movements, and 96.1% in recognizing sleeping behavior. These data indicated that regardless of the type of classroom teaching behavior of students, the attention mechanism combined with DSCNN demonstrated excellent recognition ability.

4.3 Comparison and analysis of attention combined with DSCNN performance

To further evaluate the performance of the model, the study compared the recognition accuracy of the proposed model with VGG-Net+DSCNN, TSN+DSCNN, and R (2+1) D+DSCNN on three data sets: UCF-101, HMDB-51, and STUDENT. Figure 12 shows the results of 20 different tests.

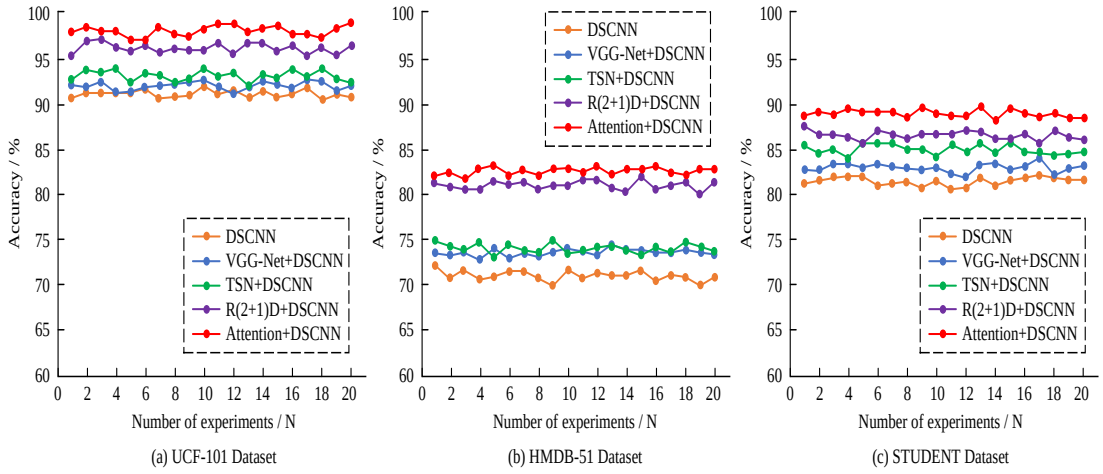


Figure12: Comparison of accuracy of different algorithms

The DSCNN combined with attention mechanism significantly outperformed most comparison models in terms of performance from the UCT-101 data set in Figure 12 (a), HMDB-51 data set in Figure 12 (b), and

the self-built STUDENT data set in Figure 12 (c). In Figure 12 (a), the proposed combination of attention mechanism and DSCNN improved recognition accuracy by 7.0% compared to the original DSCNN, 6.0% compared to VGG-Net+DSCNN, and 4.9% compared to TSN+DSCNN. In Figure 12 (b), the model improved by 11.6% compared to DSCNN. Furthermore, in Figure 12 (c) of the STUDENT data set, the model improved by 7.4% compared to DSCNN, 6.0% by VGG-Net+DSCNN, and 2.0% by R (2+1)D+DSCNN. This indicated that DSCNN with enhanced attention mechanism had higher accuracy in the student classroom teaching behavior recognition. Next, 100000 iterative training experiments were conducted on the STUDENT data set. Figure 13 shows the behavior recognition results of student classroom teaching in the STUDENT data set.

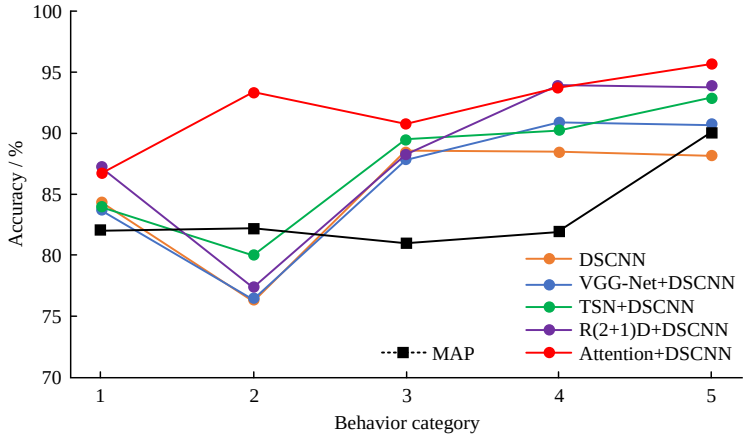


Figure 13: Comparison of the accuracy of classification of students' classroom teaching behavior

In Figure 13, the Mean Average Precision (MAP) of the five classroom teaching behaviors of students significantly improved. Especially when dealing with writing actions, the attention mechanism combined with DSCNN performed well, with a recognition accuracy of up to 88.3%. In terms of recognizing conversation actions, the model also performed excellently, with an accuracy rate of 92.0%. This value was 17.1 percentage points higher than DSCNN. The model combining attention mechanism achieved an accuracy of 90.2%in the recognition of hand raising movements. The recognition performance of sleeping and playing with mobile phones was even better, with accuracy rates of 95.5% and 97.3%, respectively. The results confirmed that the proposed attention mechanism combined with DSCNN exhibited superior performance over other models in all five categories. The overall MAP accuracy of this method also improved by about 8.0%, confirming its effectiveness and superiority in the student behavior recognition. Figure 14 shows the effectiveness of attention mechanism combined with DSCNN in student classroom teaching. This model demonstrated high accuracy and comprehensiveness in analyzing student behaviors such as writing, raising hands, talking, and sleeping. Especially when dealing with complex classroom environments, the proposed attention mechanism combined with DSCNN could effectively improve the accuracy and adaptability of behavior recognition. This innovative method provided strong technical support for the progress of intelligent campus education and promoted the intelligent transformation of the education industry. In the construction of intelligent campus education, the introduction of this model indicates a broad prospect for the application of educational technology.

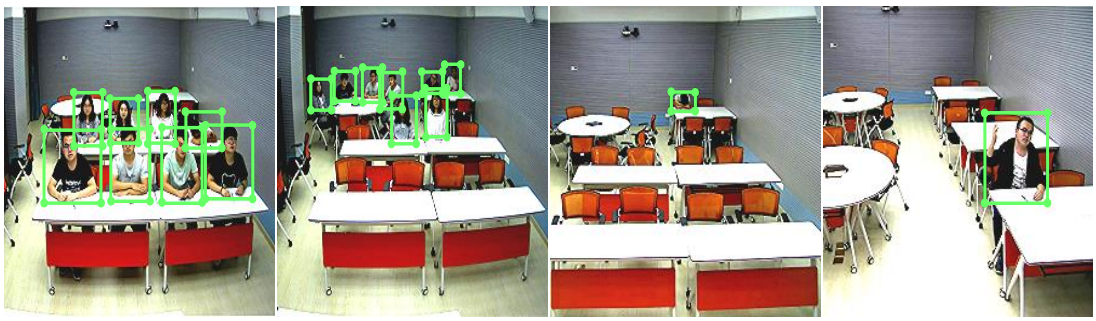


Figure 14: The effect of attention mechanism combined with DSCNN model on students' classroom teaching behavior recognition

This study evaluated the efficiency of combining attention mechanism with DSCNN by conducting experiments on processing running speed on two data sets, UCF-101 and STUDENT in Figure 15. By comparing different models, the processing speed of this model was more than twice that of DSCNN on the UCF-101 data set in Figure 15 (a). On the STUDENT data set in Figure 15 (b), the processing efficiency of this method was 1.5 times that of DSCNN. This significant speed improvement is attributed to the embedded attention mechanism, which optimizes resource allocation during the recognition process, and effectively improves the operational efficiency of the network model.

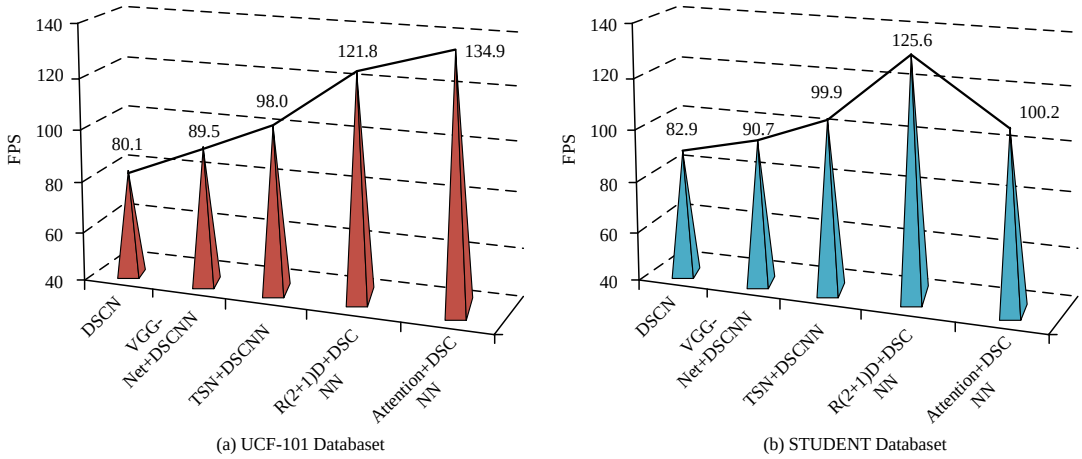


Figure 15: The running speed of different models is compared

In this study, the complexity of the models was analyzed and the computational complexity testing experiments were conducted on each model. The results are shown in Table 2. The Attention + DSCNN designed in this study performed better than the comparison model in terms of computational time and memory consumption indicators. These indicators represent the time and space complexity. For example, the computation time and memory consumption of the Attention + DSCNN model on the entire test set were 62s and 36MB, respectively. This is because the attention mechanism in the model reduces the number of neurons and parameters required for training the overall model, allowing the model to calculate quickly.

Table 2: Comparison of computational complexity of various models

Index	Test set sample usage ratio/%	VGG-Net+ DSCNN	TSN+DSCNN	R（2+1）D+DSCNN	Attention + DSCNN
Calculation time/s	20%	37	22	16	13
	50%	76	47	34	28
	100%	173	119	94	62
Memory consumption/MB	20%	9	72	86	29
	50%	11	83	96	34
	100%	14	109	102	36

A domestic high school was invited to improve the performance of the model in more situations. The experiment involved 600 middle school students, who were divided into an experimental group and a control group. The experimental group needed to run a student behavior monitoring system made of the designed model during class. The control group did not receive any intervention. The experiment lasted for 30 days, and the average scores of the experimental group and control group students before the experiment were 72.6 and 72.8 points, respectively. The average scores of the two groups of students after the experiment were 85.9 and 72.7, respectively. As a result, the model designed in this study also played a certain role in improving the learning effectiveness of students in practical applications.

5 Discussion

In curriculum education, students may lose focus or engage in various activities unrelated to class due to physical fitness, illness, interests, and other reasons. Therefore, it is necessary to design a system for monitoring classroom student behavior. This study used DSCNN and attention modules to design a model for identifying abnormal behavior in student courses.

For the test result accuracy, the accuracy of the designed model on the UCF-101 data set was 88.1%. This model had a higher accuracy of 89.4% on the customized student data set, which was higher than the traditional DSCNN and other comparative algorithms. This indicated that the data extraction performance of the DSCNN algorithm could be improved by incorporating attention modules into this algorithm. The Attention + DSCNN model had a computing time and memory consumption of 62s and 36MB on the entire test set, respectively, which was much lower than the comparison model. This is mainly because the attention mechanism is added to concentrate the algorithm's data processing flow and reduce the resource consumption of the algorithm on irrelevant data and features.

Finally, the designed model in this study had good scalability. Because the basic algorithm DSCNN used in this model is a universal algorithm for processing non-time series data. Meanwhile, this study does not separately optimize the structure based on the research object, and the overall scalability of the model is preserved. In addition, the model designed in this study does not involve ethical and privacy issues. Because students receiving education in school is not a private behavior, but a behavior with an obligation nature. Therefore, students' performance in the classroom can be monitored by various intelligent systems.

The model designed in this study is beneficial in reducing the amount of energy teachers spend on supervising students in the classroom, thereby improving their learning outcomes.

6 Conclusion

In traditional student classroom teaching, relying on manual observation and recording is not only time-consuming and laborious, but also difficult to accurately capture and analyze student behavior. Therefore, the study proposed DSCNN and optimized the model efficiency by incorporating knowledge distillation techniques. The study also integrated attention mechanisms to develop DSCNN that combined attention mechanisms to further enhance performance. The results confirmed that on the UCF-101 data set, the recognition accuracy of the attention mechanism combined with DSCNN reached 88.1%, which was 7.0% higher than traditional DSCNN. On the STUDENT data set, the recognition accuracy of this model was 89.4%, which was 8.5% higher than traditional models. The accuracy of attention mechanism combined with DSCNN reached 97.0%, 87.9%, 90.7%, 89.2%, and 96.1%, respectively, in the recognition of the five behaviors. In addition, the proposed combination of attention mechanism and DSCNN improved recognition accuracy by 7.0% compared to the original DSCNN, by 6.0% compared to VGG-Net + DSCNN, and by 4.9% compared to TSN+DSCNN. In addition, the running speed of this model on the UCF-101 and STUDENT data sets was twice and 1.5 times that of traditional DSCNN, respectively. This significant speed improvement is attributed to the embedded attention mechanism, which effectively improves the operational efficiency of the network model. These data indicate that the attention mechanism combined with DSCNN performs well in recognizing different behaviors. The limitation of this study is that the student course behaviors in the data set are not yet abundant. Future research directions will focus on collecting data sets with richer information to further train and test the designed models.

7 Funding

The research is supported by the Doctoral Foundation Project of Minzu Normal University of Xingyi (No.20XYBS01).

References

[1] S. Cheng, Q. Yang, and H. Luo, “Design of neural network-based online teaching interactive system in the context of multimedia-assisted teaching,” Informatica, vol. 48, no. 7, pp. 53-62, 2024. https://doi.org/10.31449/inf.v48i7.5205

[2] H. Ou, and J. Sun, “Multi-scale spatial temporal information deep fusion network with temporal pyramid mechanism for video action recognition,” Journal of Intelligent & Fuzzy Systems, vol. 41, no. 3, pp. 4533-4545, 2021. https://doi.org/10.3233/JIFS-189714

[3] J. Tang, L. Li, and M. Tang, “A novel micro-expression recognition algorithm using dual-stream combining optical flow and dynamic image convolutional neural networks,” Signal, Image and Video Processing, vol. 17, no. 3, pp. 769-776, 2022. https://doi.org/10.1007/s11760-022-02286-0

[4] Y. Yoshimoto, and H. Tamukoh, “FPGA implementation of a binarized dual stream convolutional neural network for service robots,” Journal of Robotics and Mechatronics, vol. 33, no. 2, pp. 386-399, 2021. https://doi.org/10.20965/jrm.2021.p0386

[5] N. S. Russel, and A. Selvaraj, “Fusion of spatial and dynamic CNN streams for action recognition,” Multimedia Systems, vol. 27, no. 5, pp. 969-984, 2021. https://doi.org/10.1007/s00530-021-00773-x

[6] C. Chen, “A novel motion recognition method based on improved two-stream convolutional neural

network and sparse feature fusion," Computer Science and Information Systems, vol. 19, no. 3, pp. 1329-1348, 2022. https://doi.org/10.2298/CSIS220105043C

[7] V. T. R. Pavan Kumar, M. Arulselvi, and K. B. S. Sastry, "Comparative assessment of colon cancer classification using diverse deep learning approaches," Journal of Intelligent Information Systems, vol. 1, no. 2, pp. 128-135, 2023. https://doi.org/10.47852/bonviewJDSIS32021193.

[8] Y. Xu, and T. T. Qiu, "Human activity recognition and embedded application based on convolutional neural network," Journal of Artificial Intelligence and Technology, vol. 1, no. 1, pp. 51-61, 2021. https://doi.org/10.37965/jait.2020.0051.

[9] W. Huang, L. Zhang, W. Gao, F. Min, and J. He, "Shallow convolutional neural networks for human activity recognition using wearable sensors," IEEE Transactions on Instrumentation and Measurement, vol. 70, no. 1, pp. 1-11, 2021. https://doi.org/10.1109/TIM.2021.3091990

[10] J. Zhai, X. Yao, G. Dong, Q. Jiang, and Y. Zhang, "3D dual-stream convolutional neural networks with simple recurrent unit network: A new framework for action recognition," 2022 4th International Conference on Communications, Information System and Computer Engineering (CISCE), vol. 1, no. 1, pp. 27-29, 2022. https://doi.org/10.1109/CISCE55963.2022.9851166

[11] R. Singh, R. Khurana, A. K. S. Kushwaha, and R. Srivastava, "A dual stream model for activity recognition: exploiting residual-CNN with transfer learning," Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, vol. 9, no. 1, pp. 28-38, 2022. https://doi.org/10.1080/21681163.2020.1805798

[12] L. Yang, X. Shan, C. Lv, J. Brighton, and Y. Zhao, "Learning spatiotemporal representations with a dual-stream 3-D residual network for non-driving activity recognition," IEEE Transactions on Industrial Electronics, vol. 69, no. 7, pp. 7405-7414, 2021. https://doi.org/10.1109/TIE.2021.3099254

[13] C. Xu, R. Liu, T. Zhang, Z. Cui, J. Yang, and C. Hu, "Dual-stream structured graph convolution network for skeleton-based action recognition," ACM Transactions on Multimedia Computing Communications and Applications, vol. 17, no. 4, pp. 1-22, 2021. https://doi.org/10.1145/3450410

[14] Z. Du, and H. Mukaidani, "Linear dynamical systems approach for human action recognition with dual-stream deep features," Applied Intelligence, vol. 52, no. 1, pp. 452-470, 2022. https://doi.org/10.1007/s10489-021-02367-6

[15] C. Zhang, Y. Xu, Z. Xu, J. Huang, and J. Lu, "Hybrid handcrafted and learned feature framework for human action recognition," Applied Intelligence, vol. 52, no. 11, pp. 12771-12787, 2022. https://doi.org/10.1007/s10489-021-03068-w

[16] Y. Tang, L. Zhang, H. Wu, J. He, and A. Song, "Dual-branch interactive networks on multichannel time series for human activity recognition," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 10, pp. 5223-5234, 2022. https://doi.org/10.1109/JBHI.2022.3193148

[17] W. Yang, J. Zhang, J. Cai, and Z. Xu, "HybridNet: Integrating GCN and CNN for skeleton-based action recognition," Applied Intelligence, vol. 53, no. 1, pp. 574-585, 2022. https://doi.org/10.1007/s10489-022-03436-0

[18] A. Jisi, and S. Yin, "A new feature fusion network for student behavior recognition in education," Journal of Applied Science and Engineering, vol. 24, no. 2, pp. 133-140, 2021. https://doi.org/10.6180/jase.202104_24(2).0002

[19] H. Zan, and G. Zhao, "Human action recognition research based on fusion TS-CNN and LSTM networks," Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 2331-2345, 2022. https://doi.org/10.1007/s13369-022-07236-z.

[20] Y. Sun, Y. Weng, B. Luo, G. Li, and D. Chen, "Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images," IET Image Processing, vol. 17, no. 4, pp. 1280-1290, 2022. https://doi.org/10.1049/ipr2.12712

[21] Y. Abdulazeem, H. M. Balaha, W. M. Bahgat, and M. Badawy, "Human action recognition based on transfer learning approach," IEEE Access, vol. 9, no. 1, pp. 82058-82069, 2021. https://doi.org/10.1109/ACCESS.2021.3086668

[22] B. Cao, and Z. Liu, "A video abnormal behavior recognition algorithm based on deep learning," IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference, vol. 4, no. 1, pp. 755-759, 2021, https://doi.org/10.1109/IMCEC51613.2021.9482114

[23] X. Li, and X. Cao, "Human motion recognition information processing system based on LSTM recurrent neural network algorithm," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 7, pp. 8509-8521, July, 2023, https://doi.org/10.1007/s12652-021-03614-x.

[24] W. Yu, P. Zhou, L. Shu, C. Jiang, and H. Zheng, "Research progress on the human behavior recognition based on machine learning methods," Recent Patents on Mechanical Engineering, vol. 16, no. 1, pp. 32-44, 2023. https://doi.org/10.2174/2212797615666220827164210

[25] Z. Deng, Q. Gao, Z. Ju, and X. Yu, "Skeleton-based multi-features and multistream network for real-time action recognition," IEEE Sensors Journal, vol. 23, no. 7, pp. 7397-7409, 2023. https://doi.org/10.1109/JSEN.2023.3246133

[26] L. Deng, R. Fu, Q. Sun, M. Jiang, Z. Li, H. Chen, and X. Bu, "Abnormal behavior recognition based on feature fusion C3D network," Journal of

Electronic Imaging, vol. 32, no. 2, pp. 21605-21606, 2023. https://doi.org/10.1117/1.JEI.32.2.021605

[27] Z. Chen, and N. Chen, "Children's football action recognition based on LSTM and a V-DBN," IEIE Transactions on Smart Processing & Computing, vol. 12, no. 4, pp. 312-322, 2023. https://doi.org/10.5573/IEIESPC.2023.12.4.312

[28] S. Zhang, "Cascade attention-based spatial-temporal convolutional neural network for motion image posture recognition," Journal of Computers, vol. 33, no. 1, pp. 21-30, 2022. https://doi.org/10.53106/199115992022023301003

[29] D. Li, S. Wang, J. Li, Y. Yang, and X. S. Tang, "Dual-stream shadow detection network: biologically inspired shadow detection for remote sensing images," Neural Computing and Applications, vol. 34, no. 12, pp. 10039-10049, 2022. https://doi.org/10.1007/s00521-022-06989-w

[30] X. Yao, J. Zhang, R. Chen, D. Zhang, and Y. Zhang, "Weakly supervised graph learning for action recognition in untrimmed video," The Visual Computer, vol. 39, no. 11, pp. 5469-5483, 2022. https://doi.org/10.1007/s00371-022-02673-1