doi:10.3772/j.issn.1006-6748.2021.02.007

Research on behavior recognition algorithm based on SE-I3D-GRU network①

Wu Jin (吴 进)②, Yang Xue, Xi Meng, Wan Xianghong
(School of Electronic and Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, P. R. China)

Abstract

In order to effectively solve the problems of low accuracy and large amount of calculation of current human behavior recognition, a behavior recognition algorithm based on squeeze-and-excitation network (SENet) combined with 3D Inception network (I3D) and gated recurrent unit (GRU) network is proposed. The algorithm first expands the Inception module to three-dimensional, and builds a network based on the three-dimensional module, and expands SENet to three-dimensional, making it an attention mechanism that can pay attention to the three-dimensional channel. Then SENet is introduced into the I3D network, named SE-I3D, and SENet is introduced into the GRU network, named SE-GRU. And, SE-I3D and SE-GRU are merged, named SE-I3D-GRU. Finally, the network uses Softmax to classify the results in the UCF-101 dataset. The experimental results show that the SE-I3D-GRU network achieves a recognition rate of 93.2% on the UCF-101 dataset.

Key words: behavior recognition, squeeze-and-excitation network (SENet), Incepton network, gated recurrent unit (GRU)

0 Introduction

Due to the increasing status of video human behavior recognition[1-2] in the field of artificial intelligence, people's demand for behavior recognition intelligent systems is increasing. Therefore, video-based behavior recognition has a wide range of applications in human-computer interaction[3-4], social public safety[5], intelligent security and other fields. As early as the middle of the 18th century, the physiologist Marey had begun a series of studies on animal behavior. In 1973, the psychologist Johansson proposed the moving light display (MLD) model. After that, the discriminative models based on human behavior are improved based on the MLD model. Based on the 3D convolutional neural network (3D-CNN)[6-8], 3D convolutional neural network can extract feature information in both spatial and temporal dimensions simultaneously. For the first time, they applied 3D-CNN to human behavior recognition, and the recognition rate on the UCF-101[9] dataset reached 85.2%. The dual stream[10] structure network was proposed in Ref. [11]. It is composed of two dimensions of time and space. First, in the time domain[12] network section, the optical flow field between consecutive multiple frames is used as multiple input channels, and finally the fully connected layer is used for classification. The spatial domain network part uses RGB[13] images for training. The remaining steps are the same as the optical flow operation. Finally, feature fusion is performed in the classification prediction layer to obtain the final result. Ref. [14] expanded the optical flow[15] and RGB flow into three dimensions, and improved the fusion processing of related features, and achieved the extreme fusion of dual flow features. The dynamic features in the time domain and the features in the space domain of feature extraction are fused, and finally good results are obtained. SENet[16] is a brand-new network structure proposed by the autonomous driving company Momenta in 2017. The fusion model of the network obtained an error rate of 2.251% on the test set, which is nearly 25% more accurate than the first place last year. Initially, the methods of behavior recognition can be divided into two methods based on deep learning and traditional classification based on manual extraction of feature information. Based on the manual classification method, first of all, it is necessary to preprocess the data of the incoming video, and then manually extract the feature information of the video. The steps of human behavior recognition based on deep learning are: (1) feature extraction of video information; (2) use a classifier to

① Supported by the Shaanxi Province Key Research and Development Project (No. 2021GY-280), the Natural Science Foundation of Shaanxi Province (No. 2021JM-459) and the National Natural Science Foundation of China (No. 61772417, 61634004, 61602377).

② To whom correspondence should be addressed. E-mail: wujin1026@126.com

Received on June 10, 2020

classify the extracted features.

A SE-I3D-GRU network based on deep learning is proposed. First, the Inception module is expanded to three-dimensional, and then this module is used to build the network, which can solve the problem that the low-level spatiotemporal features cannot be learned well in the two-dimensional convolutional network. Then, it combines the recurrent neural network GRU[17] to solve the problem that the front-end network is not good at modeling high-level timing features. Later, in order to improve the performance of the network, the attention mechanism SENet is introduced into the network. Finally, the performance of the network model is tested on the UCF-101 data set, and the UCF-101 data set achieves a recognition rate of 93.2%.

1 SE-I3D-GRU network structure design

1.1 Use the Inception module to build an I3D network

The information contained in the video is divided into two parts: time and space. Spatial information exists in the form of independent frames, which describes the characteristic information of objects or scenes in the video frames. Time information describes the movement of the camera and target in the form of motion in a video[18]. In order to make better use of this information, the Inception module of the GoogLeNet[19] network is used as the basic network to build a network model, and the 2D network module is expanded to the 3D network module.

1.1.1 The principle of the two-dimensional Inception module

In general, the most direct way to improve network performance is to increase the depth and width of the network. The depth of the network refers to the number of layers of the network, and the width refers to the number of channels per layer. However, this method will bring two shortcomings: (1) prone to overfitting, when the depth and width continue to increase, the parameters that need to be learned continue to increase, huge parameters are prone to overfitting; (2) uniform increasing the size of the network will increase the amount of calculation. Therefore, GoogLeNet introduced the Inception[20] module. GoogLeNet is composed of 22 layers, including 9 Inception modules. The basic structure of the Inception module is shown in Fig. 1.

In the Inception structure, a large number of 1 × 1 matrices are used, mainly for two purposes: (1) dimensionality reduction of the data; (2) the introduction of more nonlinearity, improve the generalization ability, after the convolution is activated by ReLU function.

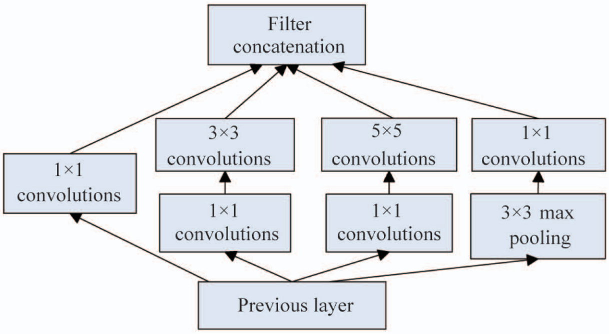


Fig. 1 Inception module structure diagram

1.1.2 Expansion of the Inception module

The expansion of the Inception module is to expand the two-dimensional convolutional neural network into a three-dimensional convolutional neural network, that is, to increase the time dimension information on the convolutional layer and the pooling layer. After the expansion, take the time dimension as N, all the kernels of the pooling layer and the convolutional layer will expand from two-dimensional $N \times N$ to $N \times N \times N$ cubes. The process of Inception module expansion is shown in Fig. 2.

In the Inception module, some adjustments are made to the steps of some convolution kernels, the steps of the largest pooling layer, and the global average pooling layer. The specific method is that in the first two largest pooling layers, the stride in the time dimension is set to 1, and the step size and spatial consistency of the last two largest pooling layers are both taken to be 2, the largest average pooling layer used is 2 × 7 × 7. In order to more intuitively show the expansion process of the GoogLeNet network, the network before and after the transformation is shown in the following two figures. The specific design of the network before transformation is shown in Fig. 3, and the network architecture after transformation is shown in Fig. 4.

In the entire I3D network model, except for the Softmax classification layer of the last layer, a ReLU activation function is added after each convolutional layer. The Softmax classification layer and ReLU activation function are not drawn in the figure.

1.2 Inception network and SENet network

1.2.1 SENet module

After the construction of the I3D network, the network architecture is initially completed. There are 9 Inception modules from the top layer to the bottom layer.

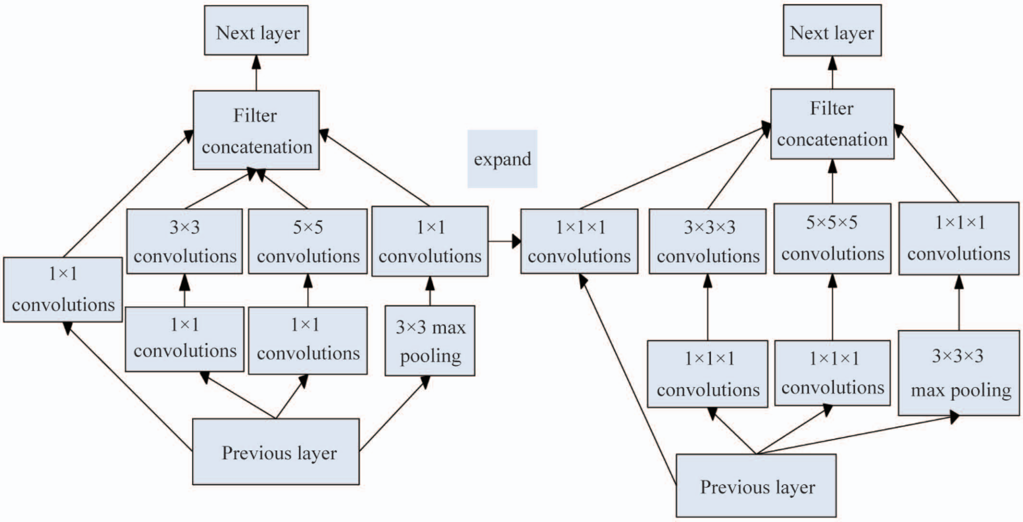


Fig. 2 Inception module expansion process

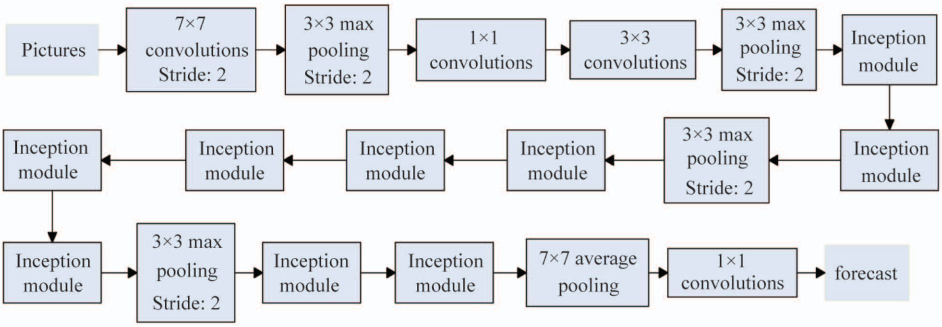


Fig. 3 2D network model before expansion

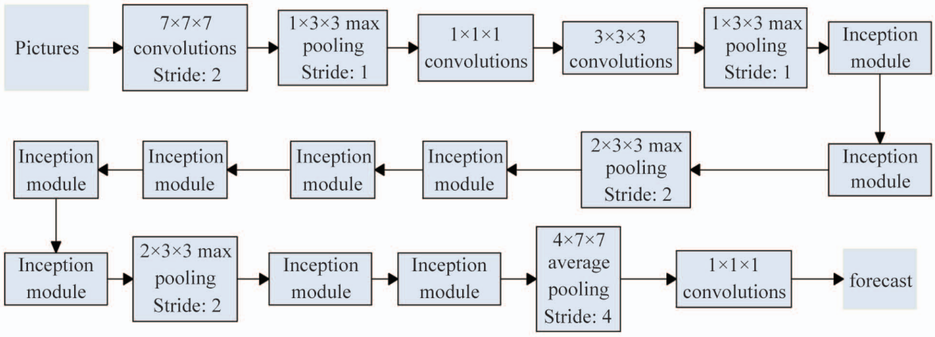


Fig. 4 3D network model after expansion

The three-dimensional Inception module is represented by Inc. They are Inc1, Inc2, Inc3, Inc4, Inc5, Inc6, Inc7, Inc8 and Inc9. SENet takes into account the relationship between feature channels and uses the 'feature recalibration' strategy to show the interdependence between features and feature channels. Specifically, the model is used to automatically obtain the importance of each feature channel through the learning method, and then according to the importance, to enhance the useful channel features and weaken the useless channel features, and can also achieve the feature channel granularity attention extraction. Here, the Inc9 module of the I3D network (the last Inception module) is recalibrated, the SENet module is fused, and the SE-I3D network is constructed for low-level space-time feature extraction.

Hu et al.[16] proposed SENet from the feature channel. The two main operations of SENet are Squeeze and Excitation. The SENet module is shown in Fig. 5.

Given an arbitrary F_{tr} mapping, first of all, input an x to the feature map $U(U^{C\times W\times H})$, its feature channel number is C', width and height are W', H'. After a series of convolution transformations, a feature with a channel number of C can be obtained. Then Squeeze and Excitation operations are performed, and finally, these weights are applied to the feature map U to obtain the output of the SE block, and the output result can be directly output to the subsequent network layer for use.

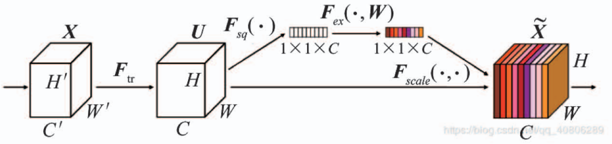


Fig. 5 SENet network model

The two main operations of the SENet module are Squeeze and Excitation operations. The specific operation process is as follows. First, the features are subjected to Squeeze operation, and F_{tr} is a convolution operation. After learning, the kernel of the learned filter is represented by $V = [v_1, v_2, v_3, \cdots, v_c]$, v_c is the parameter of the C-th filter kernel, as mentioned above, when inputting x to the feature map $U(U^{C\times W\times H})$, the output is $U = [u_1, u_2, u_3, \cdots, u_c]$, where u_c is shown in Eq. (1).

$$u_c = v_c * X = \sum_{s=1}^{c'} v_c^s * x^s \tag{1}$$

where, $*$ is the convolution, $v_c = [v_c^1, v_c^2, \cdots, v_c^{C'}]$, $X = [x^1, x^2, \cdots, x^{C'}]$, $u_c = R^{W\times H}$, and the two-dimensional convolution kernel is expressed by v_c^s. By compressing the features along the spatial dimension of each feature map, the two-dimensional feature channel is converted into a real number, so that the number of input feature channels is the same as the output dimension, and in a sense this real number has a global receptive field. Allow information from the global acceptance domain of the network to be used by all its layers.

In order to solve the problem of channel dependence, a global average pooling layer is introduced. Then the second important operation is Excitation operation. This operation is used to obtain the independence of the channel dimensions. Its role is also similar to the mechanism of gates in recurrent neural network. The weight of each feature channel is generated by w, and the parameter w is learned to explicitly model the correlation between feature channels. The gate mechanism uses the Sigmoid activation function, as shown in Eq. (2).

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{2}$$

where, δ is the ReLU activation function, σ is the Sigmoid activation function, $W_1 = R^{\frac{C}{r}\times C}$, $W_2 = R^{C\times\frac{C}{r}}$. The most important are two fully connected layers. The first fully connected layer is followed by ReLU operation, followed by a fully connected layer. The final output block is obtained by rescaling U by the Sigmoid activation function, as shown in Eq. (3).

$$\tilde{x}_c = F_{scale}(u_c, s_c) = u_c \cdot s_c \tag{3}$$

where, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_c)$, $F_{scale}(u_c, s_c)$ is the product of scalar and feature map $u_c \in R^{W\times H}$.

Finally, there is a Reweight operation. The weight of the output of Excitation is regarded as the importance of each feature channel after feature selection, and then multiply channel-by-channel weighting to the previous features to complete the recalibration of the original features in the channel dimension.

1.2.2 Integration of inception network and SENet network

The structure diagram after the fusion of two-dimensional Inception module and SENet is shown in Fig. 6. The SE module is very flexible and can be directly applied to the standard convolutional layer, which means that it is very simple to apply the SE module to the Inception network. By modifying each module in the architecture, the SE-Inception network is constructed. As can be seen from Fig. 5, after the introduction of the two-dimensional Inception module, only the input of the SE module has been changed, and the input characteristics of the SE module are the output characteristics of the Inception network. First, the features $X \in R^{C\times W\times H}$ is input to the two-dimensional global average pooling layer for operation.

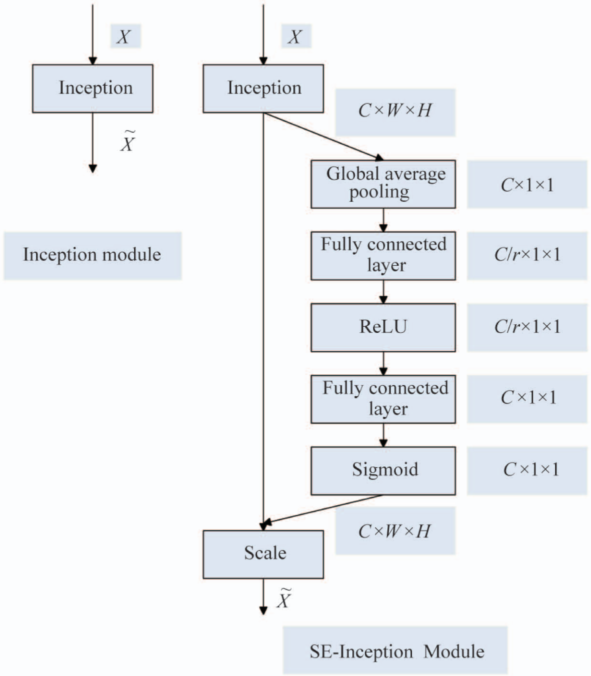


Fig. 6 Inception module and SE-Inception module

The output features are $C \times 1 \times 1$, and then the size of C is further reduced through the fully connected layer, the dimension reduction operation is performed first, and the features after the dimension reduction is $\frac{c}{r} \times 1 \times 1$, r is the ratio of dimensionality reduction, set r to 16. Then the ReLU operation is performed. After that, the output features are input to another fully connected layer for dimensionality upgrading operation. The output features are $C \times 1 \times 1$, then go through Sigmoid, and finally, the final output is obtained through the Scale operation.

1.3 I3D network and 3D-SENet network

In order to make SENet an attention mechanism that can pay attention to the three-dimensional channel, the Squeeze and Excitation operations in SENet are redefined and expanded to three-dimensional. For any given transformation $F_{tr}: X \rightarrow U$, $X \in R^{D'W'H'}$, $U \in R^{D'W'H'}$, where F_{tr} is the 3D convolution operator, $V = [v_1, v_2, \cdots, v_c]$ represents the learned filter kernel set, v_c is the parameter of the c-th filter kernel, and $U = [u_1, u_2, \cdots, u_c]$ is the output of F_{tr}, where u_c is shown in

$$u_c = v_c * X = \sum_{s=1}^{c'} v_c^s * x^s \qquad (4)$$

where, $*$ is a convolution operation, $v_c = [v_c^1, v_c^2, \cdots, v_c^{C'}]$, $X = [x^1, x^2, \cdots, x^{C'}]$, v_c^s is 3D convolution kernels.

3D-Squeeze Feature compression along the spatial dimension, while rotating three-dimensional feature channels into real numbers, the real numbers have a global receptive field to some extent, and the output dimensions match the number of input feature channels, and it also represents the global distribution of the response on the feature channel. The information of the global acceptance domain can be obtained near the input layer.

$$z_c = F_{sq}(u_c^{Inc9}) = \frac{1}{D \times W \times H} \sum_{i=1}^{D} \sum_{j=1}^{W} \sum_{k=1}^{H} u_c^{Inc9}(i, j, k) \qquad (5)$$

where, the statistic $z \in R^c$ is compressed by U through the spatial dimension $D \times W \times H$, z_c represents the c-th element of z, c is the number of channels of the Inc9 module layer in the I3D network, and u_c^{Inc9} represents the c-th of the ninth Inception module feature maps, where c is 512 and $D = W = H = 7$. Inc9 is the last Inception module.

3D-Excitation Excitation operations are similar to two-dimensional operations, and their functions are similar to the gate mechanism in recurrent neural network. A parameter is used to generate weights for each feature channel. This parameter is learned to explicitly model the correlation between feature channels. Then, the Sigmoid activation function is used as a simple gating mechanism. The mechanism of action of Sigmoid is the same as that of two-dimensional, and its 3D-Excitation operation is

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \qquad (6)$$

where, δ is the ReLU activation function, σ is the Sigmoid activation function, $W_1 \in R^{\frac{C}{r}} \times C$ acts on the first fully connected layer for dimensionality reduction operation, $W_2 \in R^{C \times \frac{C}{r}}$ acts on the second fully connected layer for dimensionality increase operation ($r = 16$). The final output is a reweighting operation, which is obtained by activating the scaling of the conversion output U. The new output of the Inc9 module is shown in Eq. (7). $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_c)$, $F_{scale}(u_c, s_c)$ is the product of scalar s_c and feature map $u_c \in R^{W \times H}$.

$$\tilde{x}_c = F_{scale}(u_c^{Inc9}, s_c) = s_c \cdot u_c^{Inc9} \qquad (7)$$

where $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_c)$, $F_{scale}(u_c^{Inc9}, s_c)$ is the channel-by-channel product of scalar S_c and feature map $u_c^{Inc9} \in R^{D \times W \times H}$. The 3D-SE structure is an attention mechanism that can pay attention to the three-dimensional channel, c represents the channel, r represents the ratio, after performing global average pooling operation on $\tilde{x}_c$, the obtained video spatial feature is represented as $\widetilde{U} \in R^{T \times C \times 1}$, where T is the total number of frames of the video , the video with $C \times 1$ as the feature vector is globally tied, and then the formed three-dimensional feature matrix is sent to the SE-GRU network for time series modeling. Fig. 7 is a schematic diagram of the SE-I3D structure. It is a schematic diagram of the fusion structure of 3D-Inception and 3D-SE in the I3D module.

1.4 Integration of SENet network and GRU network

GRU is an RNN network. In order to consider the interdependence between video frames and the importance of different frames, the SE module is very flexible, the SE module is also embedded in the GRU network. The GRU network is shown in Fig. 8.

In order not to introduce new feature sizes, the ‘Functional Recalibration’ strategy is still used to implement SE-GRU. The model can automatically understand the importance of each frame through learning, and improve the useful functions of the current frame and suppress the functions of useless frames. The SE-GRU network framework operates as follows for Squeeze and Excitation.

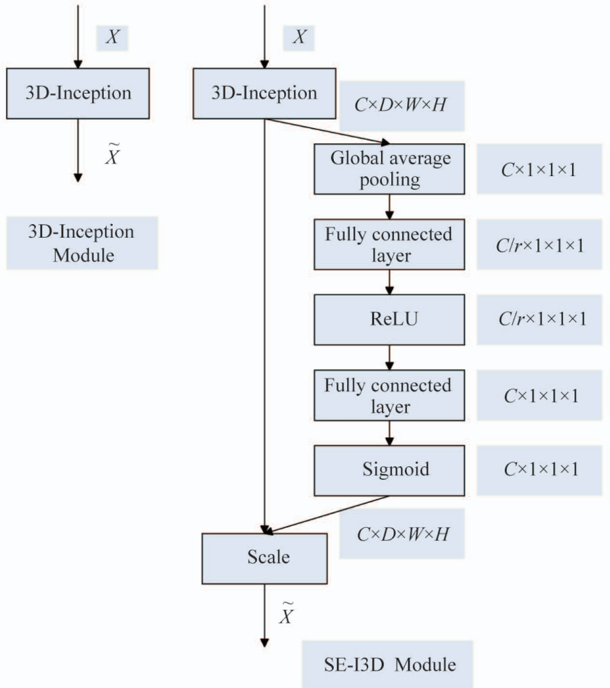


Fig. 7 3D-Inception module and SE-I3D module

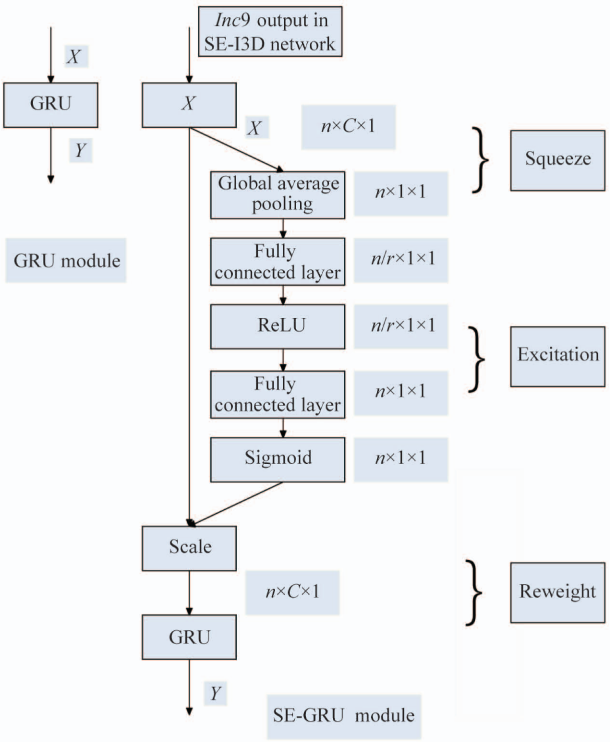


Fig. 8 SE-GRU structure diagram

There are two Squeeze operations that can enhance the reliability and importance of granular feature channels, as shown in Eq. (8) and Eq. (9).

$$z_t = F_{sq}(\tilde{U}_t) = \frac{1}{C}\sum_{c=1}^{C}\tilde{U}_{t,c} \tag{8}$$

$$\tilde{z}_C = F_{sq}(\tilde{U}_C) = \frac{1}{T}\sum_{t=1}^{T}\tilde{U}_{t,c} \tag{9}$$

where, z_t is the t-th element of $z \in R^{T\times 1}$, C is the channel number of Inc9 module layer in the SE-I3D network, and T is the total number of video frames. Only Eq. (8) is used to enhance the frame granularity, dependence and importance of the channel, Eq. (9) is also applicable to the SE-GRU network.

Excitation operation is similar to Eq. (7), and the calculation formula of weight operation is shown in

$$h_n = f(Ul_n + Wh_{n-1} + b) \tag{10}$$

where, the offset value is represented by b, and the weight values are represented by w and u.

1.5 Overall architecture of SE-I3D-GRU

In the previous introduction, the I3D network, SE-I3D network module, and SE-GRU network module are designed. On this basis, the network structure SE-I3D-GRU designed in this paper is obtained by fusing SE-I3D and SE-GRU, as shown in Fig. 9.

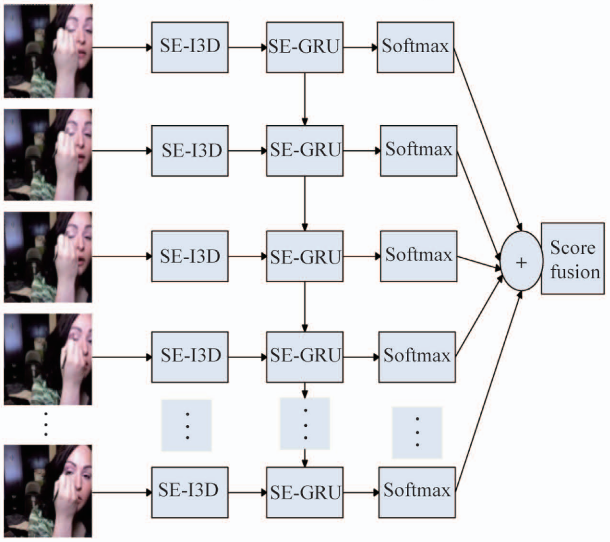


Fig. 9 SE-I3D-GRU structure diagram

First, SE-I3D is used to extract the features of the video. Then, the activation tensor output is selected from the Inc9 module layer as the feature value and it is input into the SE-GRU network for time series modeling. Finally, the output result is predicted using Softmax.

2 Experimental results and comparative analysis

2.1 Experimental environment

The hardware and software environment is shown in Table 1.

Table 1 Hardware and software environment

Experimental hardware and software environment and description
1. Operating system, Ubuntu 18.04, Linux version
2. CUDA10.1, the underlying software platform for GPU acceleration
3. Keras version 2.1.1, the backend uses the TensorFlow framework
4. TensorFlow version 1.2.0
5. Graphics card, GTX1080Ti, 11 GB video memory
6. Memory, Kingston HyperX Savage DDR4, main frequency: 2400 MHz, capacity: 8 GB
7. Hard disk, Western Digital, 1000 GB memory, SATA interface, maximum transfer rate 300 MB/s.

2.2 Data set

UCF-101 data set is currently the larger human behavior data set, as shown in Fig. 10. Fig. 10 shows all the categories of the UCF-101 dataset.

The UCF-101 data set contains 101 action categories. The 101 categories of video are divided into 25 groups, each group contains 4 – 7 video clips, and in these groups, each group of video may have some similar characteristics, such as similar background or action. For the video actions inside, it can be divided into five types: musical instrument performance, sports, interaction between people, interaction between people and objects, and only body movement.

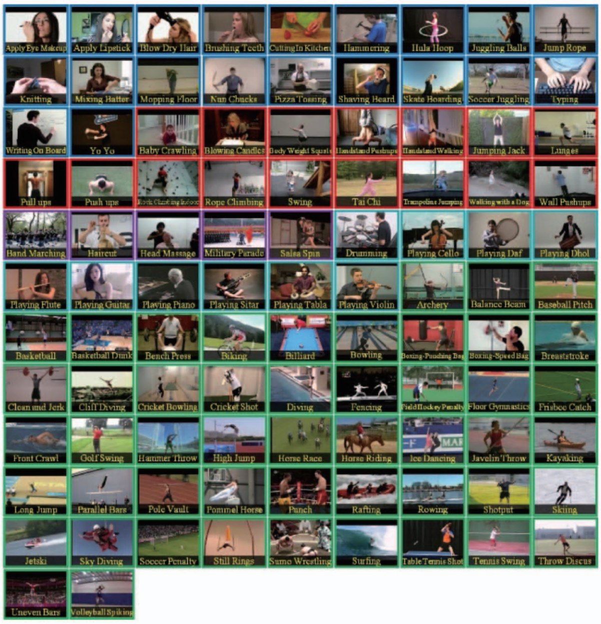


Fig. 10 All categories of UCF101 dataset

The UCF-101 data set contains 13 320 short videos, with an average time of 10 – 15 s. Therefore, a higher frame rate can be used for video frame extraction (for example, 6 fps). The specific statistical information is shown in Fig. 11.

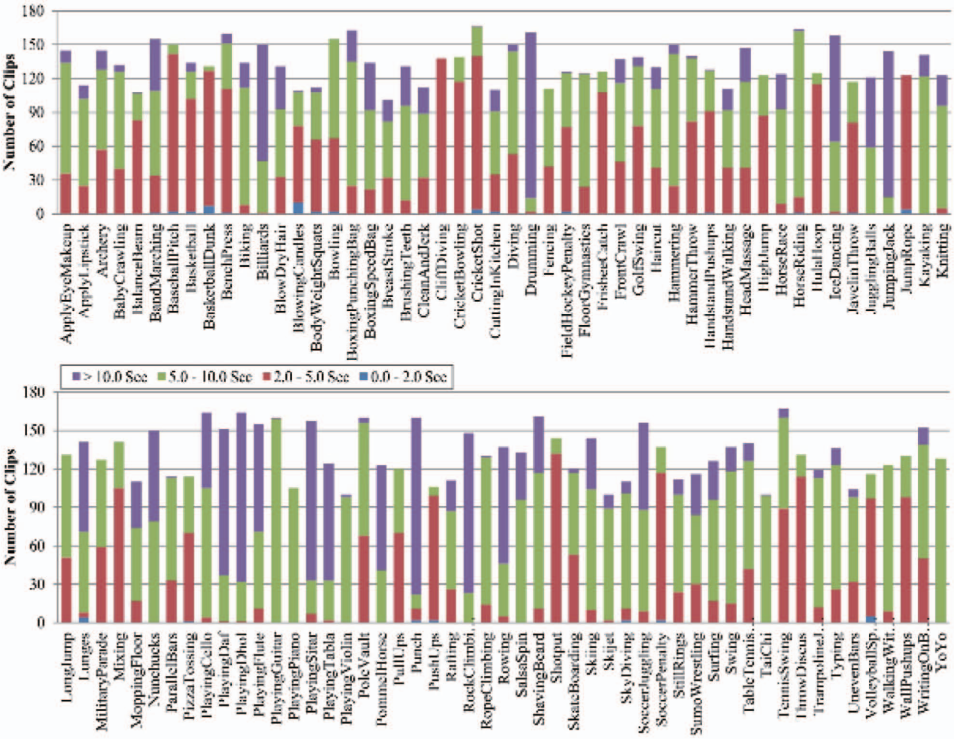


Fig. 11 UCF101 data set video length statistics

The abscissa is the category of behavior recognition, the ordinate is the number of videos of each length included in each video type, and different colors represent different lengths. It can be clearly seen from Fig. 11 that the number of each video is similar, and each video is basically more than 2 s. And each video has a complete description of each action, which is a good quality video. The resolution is 320 × 240.

The time statistics of each behavior category of the UCF-101 data set are shown in Fig. 12. The abscissa represents each behavior category, the ordinate represents the length of time of the video in seconds.

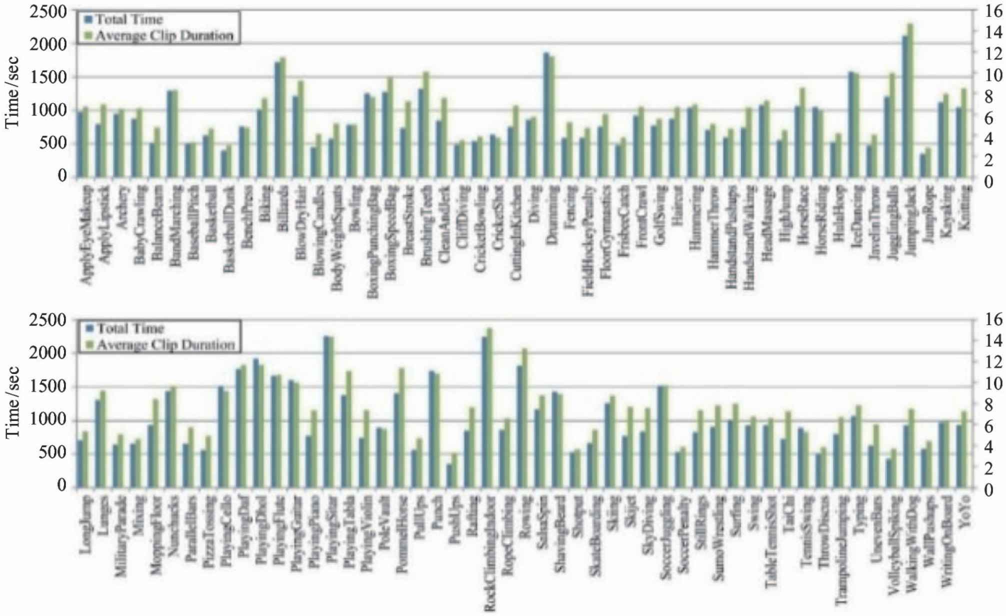


Fig. 12 Statistics of the time of each behavior category in the UCF101 data set

2.3 RGB video data preprocessing

The behavior recognition datasets all exist in video format, such as the UCF-101 dataset used in this article. Since the method of directly inputting the video into the network is not desirable, the video must be converted into a sequence of picture frames. Not only that, but it can also speed up the training speed of the network. Therefore, this article uses the open source software FFmpeg as a conversion tool before doing preprocessing. After FFmpeg decoding, all video sequences are parsed into a sequence of pictures, and named and arranged in a certain order. The parsed pictures are stored with JPG image encoding. Some of the parsed pictures are shown in Fig. 13.

Fig. 13 UCF-101 partial resolution picture

After decompressing the data set of the video format UCF-101, the size is 6.8 GB, and after conversion, the size of the picture format is 56 GB.

In the experiment, the following operations were performed on the data preprocessing.

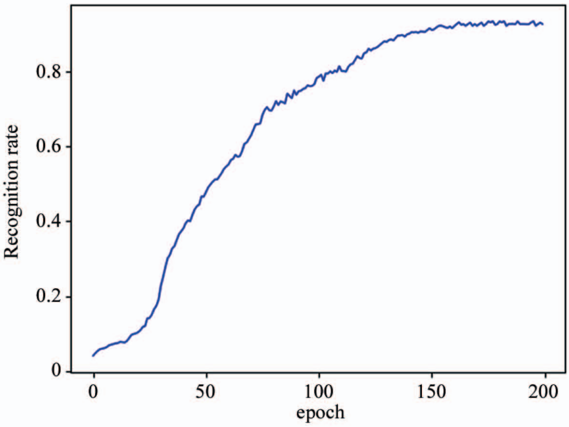
(1) Adjust the size of the input image. The size of the last image through the fully connected layer is (112, 112).

(2) Data enhancement. In the experiment, the image was first flipped, and then the data set was processed by means of noise disturbance.

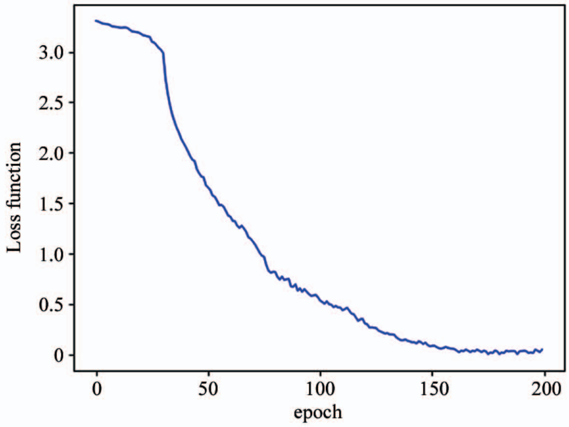
2.4 Analysis of experimental results

In the training process of the UCF-101 data set, since stochastic gradient descent (SGD) can quickly converge on large data sets where the noise is not particularly large, SGD is used as an optimizer for SE-I3D-GRU network. The mean square error function selected by the loss function is also added to the SGD optimizer. After several trainings in this paper, the learning rate is 0.01 when the effect is the best. The learning rate decay is 1e-9, and the learning rate will decrease by 1e-9 every iteration.

The maximum number of training steps of the network is 40 000 steps, which means that during the training process, a total of 40 000 trainings are used. When recording training data, the History module of the callback function Callbacks in Keras is used. Recorded once at the end of each epoch, a total of 200 data were recorded throughout the training process, including 200 iterations of each epoch. The batch size is set to 16, which means 16 random videos for each training. Fig. 14 shows the SE-I3D-GRU network training process on the UCF-101 dataset. Fig. 14(a) is the recognition rate curve during training, and Fig. 14(b) is the loss function curve during training.



(a) SE-I3D-GRU training process recognition rate curve



(b) SE-I3D-GRU training process loss function curve

Fig. 14 SE-I3D-GRU training process

It can be seen from Fig. 14(a) that the recognition rate rises slowly at the beginning of the network training. It may be because the initial learning rate is high and the current local optimal solution has not been found. As the learning rate continues to decrease, the parameters becomes small, and slowly approaching the best advantage makes the recognition rate that starts to rise. When training to 140 epochs, the recognition rate and loss function tend to ease, and the network has begun to converge. When training is 160 epochs, the loss function and recognition rate is no longer on the rise, and the network has converged.

In the process of network training, in order to improve the performance of the network, the introduction of the attention mechanism SENet and the fusion of the GRU network have been carried out. Table 2 is a comparison of the recognition rate of the algorithm in different frameworks on the UCF-101 dataset. The tested algorithm frameworks are the following five types: I3D network, I3D-GRU network, SE-I3D + GRU network, I3D + SE-GRU network, and SE-I3D + SE-GRU network.

Table 2 The recognition rate of the model on the UCF-101 dataset under different frameworks

Algorithm	UCF-101 accuracy
I3D	90.6%
I3D-GRU	91.4%
SE-I3D + GRU	92.1%
I3D + SE-GRU	91.8%
SE-I3D + SE-GRU	93.2%

In Table 2, the I3D network expanded using the two-dimensional GoogLeNet was tested on the UCF-101 dataset, and the recognition rate was only 90.6%. When integrated into the GRU network, the recognition rate was 91.4%, which was 0.8% higher than before. Recognition rate, which shows that the introduction of GRU network in the I3D module can make the feature extraction ability stronger, also shows that GRU is suitable for modeling high-level timing features in 3D networks. With the introduction of SENet, the recognition rate of I3D-GRU is 92.1%, and the accuracy rate is improved by 0.7%. The introduction of SENet only in the GRU network increases the recognition rate by 0.4%. At the same time, SENet is introduced, and the SE-I3D-GRU recognition rate is 93.2%, which is 2.1% higher than I3D-GRU and 2.9% higher than I3D. It can be seen that SENet's role in improving the network is still very objective, and it can effectively achieve the attention extraction of pixel and frame granularity to increase the recognition rate. And the recognition rate of SE-I3D + GRU on the UCF-101 dataset is higher than that of I3D + SE-GRU, which shows that the role of SENet attention mechanism in convolutional layer is more obvious than in recurrent neural network GRU. The feasibility of the proposed algorithm on the recognition rate is verified.

3 Conclusion

Aiming at the advantages and disadvantages of current video behavior recognition, a network model SE-I3D-GRU for behavior recognition is proposed. First, considering the superiority of 3D-CNN for extracting spatio-temporal features and the modeling of GRU networks at high-level time-series features, the I3D network is designed based on the Inception framework and the inflation network is introduced. It is SE-I3D. The important features of each feature channel are automatically obtained through learning, according to the importance of the features, enhance useful features and limit unimportant features. In order to automatically acquire the importance of each framework through learning, the attention mechanism SENet is introduced into the GRU network, named SE-GRU. In order to further improve the performance of the model, SE-I3D and SE-GRU are merged to design the SE-I3D-GRU network. It is used to enhance useful channel features, weaken useless channel features, and extract the attention of the frame granularity of the feature channels. Finally, the network performs final debugging on the UCF-101 data set. The efficiency of designing the network is verified.

References

[1] Dai X, Liu X, Lai J, et al. Human behavior deep recognition architecture for smart city applications in the 5G environment[J]. *IEEE Network* , 2019, 33(5): 206-211

[2] Sigurdsson G A, Russakovsky O, Gupta A. What actions are needed for understanding human actions in videos?[C]//2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 2156-2165

[3] Herath S, Harandi M, Porikli F. Going deeper into action recognition: a survey[J]. *Image and Vision Computing*, 2017, 60(1): 4-21

[4] You Q, Jiang H. Action4D: real-time action recognition in the crowd and clutter[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 11849-11858

[5] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6479-6488

[6] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4489-4497

[7] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(1): 221-231

[8] Li J, Liu X, Xiao J, et al. Dynamic spatio-temporal feature learning via graph convolution in 3D convolutional networks[C]//2019 International Conference on Data Mining Workshops, Beijing, China, 2019:646-652

[9] Khurram S, Amir R, Mubarak S. UCF101: a dataset of 101 human action classes from videos in the wild[J]. *arXiv*: 1212.00402v1, 2012

[10] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1933-1941

[11] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. *Computational Linguistics*, 2014, 1(4):568-576

[12] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks for action recognition in videos[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(11):2740-2755

[13] Wang P, Li W, Ogunbona P, et al. RGB-D-based human motion recognition with deep learning: a survey[J]. *Computer Vision and Image Understanding*, 2018, 1(1): 1-22

[14] Liu Y, Zhang K, Wang C X. Human behavior recognition method based on dual-stream convolutional neural network [J]. *Application of Computer Systems*, 2019, 28(7): 234-239 (In Chinese)

[15] Sun S, Kuang Z, Sheng L, et al. Optical flow guided feature: a fast and robust motion representation for video action recognition[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1390-1399

[16] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 2018: 7132-7141

[17] Cho K, Van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014: 623-631

[18] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778

[19] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015:1-9

[20] Szegedy C, Vanhoucke V, Ioffe S, et al, Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016:2818-2826

Wu Jin, born in 1975. She received her B. S. degree from Xi'an Jiaotong University in 1998, and she also received her M. S. degree from Xi'an Jiaotong University in 2001. Her research focuses on key techningues for signal and information processing.