Long-term recurrent convolutional network violent Behaviour recognition with attention mechanism

Qiming Liang[1], *Yong* Li[2,*], *Kaikai* Yang[1], *Xipeng* Wang[1], and *Zhi* Li[1]

[1]Graduate team, Engineering University of PAP, 710086 Xi'an, China
[2]College of Information Engineering, Engineering University of PAP, 710086 Xi'an, China

Abstract. Violent behavior recognition is an important direction of behavior recognition research. For traditional violent behavior recognition algorithms, there is too much background information when processing video information, which will cause greater interference in feature extraction, so the recognition accuracy is not high. Improved on the basis of effective recurrent convolutional network, a long-term recurrent convolutional network with attention mechanism is proposed. In the video preprocessing stage, a variety of attention mechanisms are introduced. In the feature extraction stage, the lightweight end-to-side neural network architecture GhostNet and convLSTM are selected to build a long-term recurrent convolutional network. The global average pooling and fully connected layer are used in the classification process. The combined approach realizes the classification of behaviours. The final results show that in the Hockey dataset, the algorithm in this paper has increased by 0.4% compared to LRCN, in the RWF-2000 dataset with more samples, it has increased by 10.5% compared to LRCN, and has increased by 1.75% compared to I3D, indicating that the algorithm in this paper can effectively suppress the background information. Interference, improve the performance of the algorithm.

1 Introduction

Violent behavior recognition has broad application prospects in the field of intelligent security and is a key field of behavior recognition research. Behavior recognition mainly includes three steps: video preprocessing, behavior expression and behavior classification [1]. According to different ways of expressing behavior, behavior recognition can be divided into traditional learning framework and deep learning framework.

The feature extraction of behavior in the traditional learning framework mainly includes global feature extraction and local feature extraction. For example, Bobick et al[2] established a motion energy map on the basis of background subtraction to achieve behavior classification. Yang [3] determined the coordinates of joint points establish a three-dimensional contour of the human body for feature extraction. Willems[4] et al. proposed a

[*] Corresponding author: liyong@nudt.edu.cn

Harris3D-based spatiotemporal interest point detection method, and Wang[5,6] et al. proposed dense trajectories extraction related algorithms DT (Dense Trajectories) and IDT (Improved Dense Trajectories).

The framework of deep learning currently commonly used feature extraction methods mainly include Two-stream CNN model, spatio-temporal model and time series model. The Two-stream CNN model[7] uses two parallel CNNs to extract the spatial information and timing information in the video respectively, and classify them through channel fusion; The spatio-temporal model uses 3D convolution[8] to make the 3D convolution kernel perform convolution operations with several adjacent frames of video images at the same time to extract the spatial and timing information in the video; The time series model mainly uses the long-term recurrent convolutional network (LRCN)[9] method to cascade the CNN and the recurrent neural network, and extract the spatial information of the video through the CNN , And then extract the timing information through a recurrent neural network.

The long-term recurrent CNN of time series model is mixed with a large amount of background information when extracting spatiotemporal information, which leads to low accuracy of behavior recognition. In order to effectively extract the key information of the human body in the video and reduce the impact of complex background information on the model, this paper uses a variety of attention mechanisms for data preprocessing on the basis of the time series model long-term recurrent convolutional neural network, and uses lightweight convolution Network GhostNet[10] extracts the spatial information of the video, uses ConvLSTM to extract the timing information, and finally uses global average pooling (GAP) and fully connected layer for classification. The results show that the algorithm in this paper can reduce the interference of background information and improve the accuracy of the network.

2 Related Works

2.1 Attention mechanism

When humans observe things, they will focus their attention on a certain part, ignoring the secondary information at the edge of the sight line. Similarly, the computer can also use a similar attention mechanism to reduce the interference of background information when processing image information. The attention mechanism is widely used in various fields of deep learning. Adding the attention mechanism to the network structure can improve the feature extraction ability of the network and reduce the amount of calculation parameters. At present, the attention mechanism in the visual field mainly includes the spatial domain attention mechanism, the channel domain attention mechanism and the fusion hybrid attention mechanism [11]

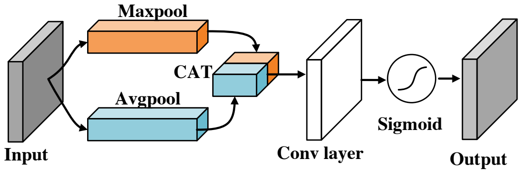

Fig. 1. Spatial attention mechanism.

As shown in Figure 1, this paper uses the spatial attention module SAM[12] for data processing in the data preprocessing stage. For the input data, the maximum pooling and average pooling are performed respectively, and then the feature maps are concatenated (CAT), form an efficient feature indicator, and then perform convolution operations through the convolution layer to extract the key information area of the feature map, suppress irrelevant information, and finally activate it through the sigmoid function. The specific expression is:

$$F_{CAT} = CAT\left[maxpool\left(F_{in}\right),\ avgpool\left(F_{in}\right)\right] \tag{1}$$

$$F_{out} = \sigma\left[f^{3\times 3}\left(F_{CAT}\right)\right] \tag{2}$$

where F_{in} represents the input frame difference map, *maxpool* and *avgpool* represent maximum pooling and global pooling respectively, *CAT* represents the results of connection maximum pooling and global pooling, $f^{3\times 3}$ represents convolution operation with a convolution kernel of size 3. σ represents feature activation through the sigmoid function, and finally output .

As shown in Fig 2, this paper also uses another concurrent spatial and channel squeeze & excitation (scSE)[13] module, which uses the spatial attention module cSE (Spatial Squeeze and Channel Excitation Block) concurrently with the channel domain module sSE (Channel Squeeze and Spatial Excitation Block), it strengthens important feature data and reduces the impact of non-critical data. This paper changes all the convolution operations in the original scSE module to the Ghost module to reduce the size of the module, denoted as iscSE, the specific expression is:

$$F_1 = \sigma_1\left\{F_{Ghost}\left[avgpool\left(F_{in}\right)\right]\right\} \tag{3}$$

$$F_{cSE} = F_{in}\sigma_2\left[F_{Ghost}\left(F_1\right)\right] \tag{4}$$

$$F_{sSE} = F_{in}\sigma_2\left[F_{Ghost}\left(F_{in}\right)\right] \tag{5}$$

$$F_{scSE} = F_{cSE} + F_{sSE} \tag{6}$$

where F_{in} represents the input frame difference map, *avgpool* represents the average pooling, F_{Ghost} represents the use of the Ghost module instead of the convolution operation, σ_1 represents the feature activation through the relu function, σ_2 represents the feature activation through the sigmoid function, and F_{cSE} represents the output feature map of the cSE module , F_{sSE} represents the output characteristic diagram of the sSE module, F_{scSE} represents the final output result of the scSE module.

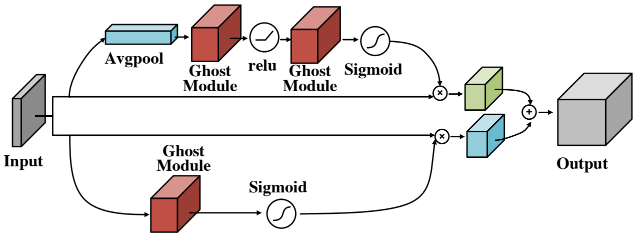

Fig. 2. Improved scSE module.

2.1 Convolutional Neural Network

At present, the mainstream CNN has a lot of redundancy, and it consumes a lot of computing power in the calculation process. Han Kai and Wang Yunhe of Huawei's Noah's Ark Laboratory proposed a new end-to-side neural network architecture GhostNet, which can reduce the time complexity of the network and reduce the computational cost during the experiment. The traditional convolutional layer can be expressed as:

$$Y = X * f + b \tag{7}$$

where X represents the input data, and $X \in R^{c\times h\times w}$ (c, h, w represent the channel and length and width of the input data respectively);* represents the convolution operation;f represents the convolution kernel of the convolution operation, and $f \in R^{c\times k\times k\times n}$ (c, k, n respectively represent the number of input data channels, the size of the convolution kernel and the number of feature maps of the convolution kernel); b represents the bias term; Y represents the output result of the convolution, and $Y \in R^{h'\times w'\times n}$ (h' 、 w' 、 n represent the length, width and number of feature maps of the output data respectively).

Traditional convolution will produce a lot of redundancy during the operation. In order to reduce the complexity of the network, as shown in Fig 3, the author established the Ghost module. The basic idea of the Ghost module is to use a smaller convolution kernel to perform convolution operations, and then cascade with linear operations to achieve a similar effect to traditional convolution operations. The Ghost module has plug-and-play features, which can improve the performance of CNNs, and can form Ghost bottlenecks through superposition, thereby building a lightweight high-performance network GhostNet. The Ghost module first uses a smaller convolution kernel to perform convolution operations, which can be expressed as:

$$Y' = X * f' \tag{8}$$

where f' represents a convolution kernel with a dimension slightly smaller than f, and $f' \in R^{c\times k\times k\times m}$ ($m < n$), Y' is the newly obtained output feature after the operation. For the obtained new feature map, linear operation is performed again, specifically:

$$y_{ij} = \varphi_{ij}\left(y_i'\right) \tag{9}$$

y_i' is the i-th inherent feature map in Y' ,and φ_{ij} is the j-th (except the last) linear operation, used to generate the j-th feature map y_{ij} .

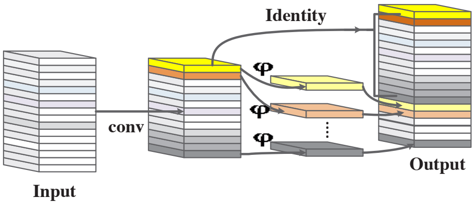

Fig. 3. Ghost module structure.

Ghost module uses linear operation to achieve part of the function of convolution, and the complexity is less than convolution operation. As shown in Figure 4, similar to the residual block of ResNet, Ghost bottleneck (G-bneck) is built through the superposition of Ghost modules. G-bneck integrates multiple convolutional layers and a direct connection (shortcut) structure. For G-bneck with a step size of 1, the first Ghost module is used as an expansion layer to increase the number of channels, and the second Ghost module reduces the number of channels is matched with the direct connection structure; For the G-bneck with a step size of 2, a deep convolution module (Depthwise Convolution, DWConv) with a step size of 2 is also included between the two Ghost modules. Direct connection connects the input and output of the convolutional layer. ReLU is not used after the second Ghost module, and the other layers apply batch normalization (BN) and ReLU nonlinear activation after each layer.

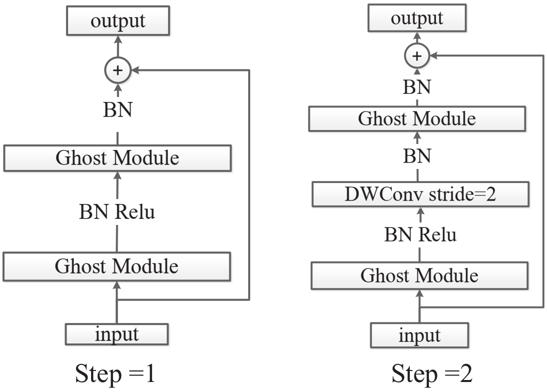

Fig. 4. Ghost bottleneck structure.

On the basis of the MobileNetV3[14] network structure, G-bneck is used as the basic unit to assemble, and SE (Squeeze-and-Excitation) modules are introduced in some layers, and finally a lightweight network GhostNet is built.

2.3 Long-Short Term Memory network

Long-Short Term Memory (LSTM) [15], a variant of RNN, combines short-term memory and long-term memory through gate control. Compared with RNN, it can retain long-term memory, which helps to obtain timing information, text generation and speech recognition are widely used in fields. LSTM departments rely on similar feedforward neural networks to calculate, so it is also called fully connected LSTM (FC-LSTM). FC-LSTM can easily process time series data, but for spatial data, due to its strong local features, FC-LSTM cannot describe the local features, which will bring serious redundancy. Sudhakaran[16] et al introduced convLSTM to replace the traditional LSTM, and realized the fusion of spatiotemporal information.

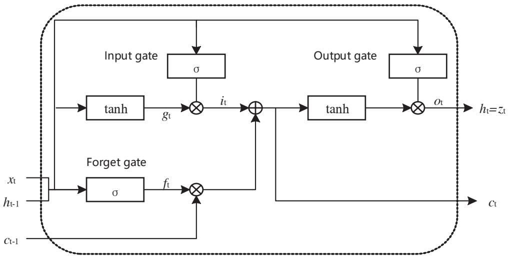

Fig. 5. Convlstm unit structure.

As shown in Figure 5, convLSTM uses input gates, output gates and forget gates to determine the input and output of a unit module. The forward recurrent formula is:

$$i_t = Sigmoid(w_x^i * I_t + w_h^i * h_{t-1} + b^i) \tag{10}$$

$$f_t = Sigmoid(w_x^f * I_t + w_h^f * h_{t-1} + b^f) \tag{11}$$

$$\tilde{c}_t = Tanh(\tilde{w}_x^c * I_t + \tilde{w}_h^c * h_{t-1} + \tilde{b}^c) \tag{12}$$

$$c_t = \tilde{c}_t \, i_t + c_{t-1} f_t \tag{13}$$

$$o_t = Sigmoid(w_x^o * I_t + w_h^o * h_{t-1} + b^o) \tag{14}$$

$$h_t = o_t Tanh(c_t) \tag{15}$$

3 Proposed method

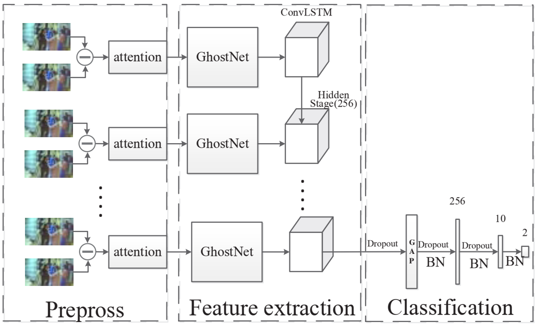

Fig. 6. Long-term recurrent convolutional neural network with attention mechanism.

The network structure established in this paper includes a video preprocessing module, a feature extraction module, and a classification module. The network structure is shown in Figure 6. In the video preprocessing module, the frame difference operation is performed on two adjacent frames, and the key information in the data is extracted through the attention mechanism module to reduce the parameters. In the feature extraction module, the lightweight CNN GhostNet is used to extract high-dimensional visual features, and the corresponding time-series high-dimensional visual features are modeled through the convLSTM network to obtain the spatiotemporal feature representation of the video. Finally, in the classification module, the classification is realized through the global average pooling and the fully connected layer to determine whether it is a violent behavior

In order to assemble GhostNet and convLSTM modules, this paper uses a convolutional layer to replace the fully connected layer of GhostNet when constructing the network. The output of the convolutional layer contains a 256-dimensional feature map of spatial features, which participates in the extraction of spatio-temporal features together with the 256-dimensional data generated by the ConvLSTM module to initialize the hidden layer. Finally, the last ConvLSTM network with long-term video information is retained, and the spatio-temporal features extracted before aggregation are used as the overall representation of the video. Add batch normalization and dropout to the global average pooling layer and the fully connected layer to reduce the risk of network overfitting, improve the generalization ability of the model, and realize the classification of behaviors. In this paper, the scSE module, the improved scSE module and the SAM module are respectively selected in the attention mechanism module, forming scSE-GhostNet-convLSTM, iscSE-GhostNet-convLSTM and SAM-GhostNet-convLSTM.

4 Implementation details

4.1 Datasets

The Hockey dataset contains 1,000 violent and non-violent videos collected from ice hockey matches. The training set includes 800 video clips, and the validation set includes 200 video clips. The main content of the video is the violent actions in the ice hockey game. Each video is 2 seconds long and contains 41 frames. Due to the small number of videos in the Hockey dataset, single scenes, high risk of overfitting, and limited application value, it is difficult to meet the needs of deep neural network learning. Therefore this paper introduces the latest RWF-2000[17] dataset. The dataset contains 2000 surveillance video clips collected from YouTube. The training set includes 1,600 video clips, and the verification set includes 400 video clips. Each video clip is 5 seconds in length and contains 150 frames. It mainly includes violent behaviors such as two persons, multiple persons and crowds, with rich scenes and difficult identification. And the video clips are all obtained by security cameras, without multimedia technology modification, fit the actual scene, and have high research value.

Fig. 7. Hockey dataset.

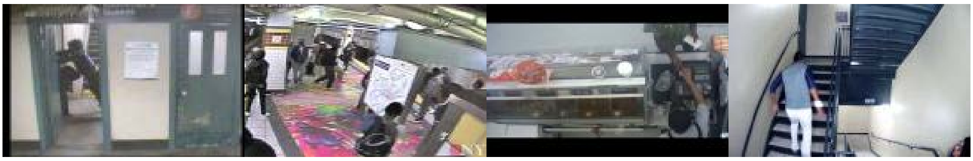

Fig. 8. RWF-2000 dataset.

4.2 Parameter configuration and preprocessing

This paper uses the Pytorch deep learning framework and CUDA 10.2 throughout the training and testing process, the GPU is NVIDIA GeForce GTX 1080, and the CPU is Intel I9-10920x. When processing RWF-2000, in order to reduce the data and the computational burden, at the same time, in order to keep the timing information in the data as much as possible, one frame is extracted every two frames from each video segment. The initial learning rate of this experiment is set to 0.0001, the learning rate drops to 50% of the original every 100 epochs, and the batch size is 6. The ratios of the video clips in the RWF-2000 dataset are quite different. In order to facilitate model processing, the screen size is uniformly adjusted to 256×256 for training. Use RMSprop to optimize the algorithm, and use the GhostNet model trained on ImageNet to reduce the risk of overfitting and reduce the amount of calculation for network training.

5 Results

Randomly select 200 from the Hockey dataset as the verification set, and extract each video as a continuous 41 frames of images for experimentation. This paper uses GhostNet-convLSTM as the basic skeleton. Before GhostNet, the scSE, iscSE, and SAM modules are added for experiments. After 600 iterations of training, the optimal accuracy rate has been improved. The experimental results are shown in Table 1.

Table 1. The optimal accuracy of different algorithms in the Hockey dataset.

Algorithm	**Accuracy**
HOF+BoW[18]	88.6
HOG+Bow[18]	91.7
3D-CNN[8]	91
3D-CNN(RGB+FLOW) [9]	94.4
LRCN[9]	97.1
GhostNet-ConvLSTM(ours)	94.5
scSE-GhostNet-ConvLSTM(ours)	95.5
iscSE-GhostNet-ConvLSTM(ours)	96.4
SAM-GhostNet-ConvLSTM(ours)	97.5

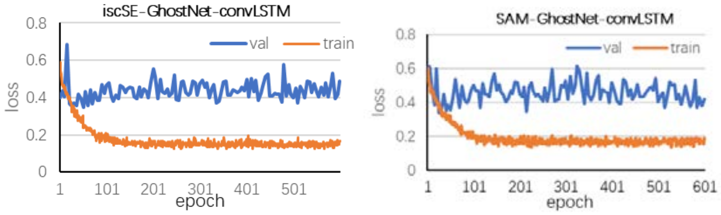

Fig. 9. The loss curve of iscSE-GhostNet-convLSTM and SAM-GhostNet-convLSTM in the RWF-2000 dataset.

It can be seen from Table 1 that after 600 epochs of training, the long-term recurrent convolutional network with GhostNet-convLSTM as the basic structure has a certain improvement in accuracy compared to some traditional algorithms. The spatial attention mechanism (SAM) is introduced. The long-term recurrent convolutional network has a 0.4% improvement over the traditional LRCN. Since the Hockey dataset has fewer training samples, the included scenes are too single, the ability of the attention mechanism to deal with complex backgrounds is not uncovered, and the algorithm is prone to overfitting, so the improvement of the algorithm is limited.

In order to further verify the performance of the algorithm, this paper introduces RWF-2000, a violent behavior dataset with more samples and more complex scenes. The RWF-2000 dataset contains 2000 video clips, and each video contains 150 frames. One frame is intercepted every two frames to form a continuous image of 75 frames for training. Figure 9 and Figure 10 are the loss value curves of iscSE-GhostNet-convLSTM and SAM-GhostNe-convLSTM on the RWF-2000 dataset, respectively.

As can be seen from the figure above, after 600 trainings, the long-term recurrent convolutional network with the attention mechanism is trained in the more complex RWF-2000 dataset with small amplitude and stable data, indicating that this algorithm is generalized Strong ability and performance. The optimal accuracy of different algorithms on the RWF-2000 dataset is shown in Table 2:

Table 2. The accuracy of different algorithms on the RWF-2000 dataset.

Algorithms	Accuracy
LRCN[9]	77
3D-CNN[8]	82.75
I3D[19]	85.75
GhostNet-ConvLSTM(ours)	86.1%
scSE-GhostNet-ConvLSTM(ours)	86.4%
iscSE-GhostNet-ConvLSTM(ours)	87%
SAM-GhostNet-ConvLSTM(ours)	87.5%

As shown in Table 2, after 600 trainings, The accuracy of iscSE-GhostNet-convLSTM on the RWF-2000 dataset has reached 87%, which is a 10% improvement over LRCN, an improvement of 1.25% over I3D, and an improvement of 0.9% over the GhostNet-ConvLSTM proposed in this paper. The accuracy of SAM-GhostNe-convLSTM on the RWF-2000 dataset reached 87.5%, which is 1.4% higher than GhostNet-ConvLSTM, indicating that the algorithm can effectively suppress the interference of background information when dealing with violent behavior in complex backgrounds. Thereby improving the accuracy of violent behavior recognition.

6 Summary

In order to suppress the interference of complex background information on violent behavior recognition and improve the accuracy of violent behavior recognition, this paper redesigned the network architecture on the basis of the long-term recurrent convolutional network, and chose the lightweight CNN GhostNet and convLSTM to form the basic structure. In the preprocessing part, a variety of attention mechanisms are introduced, and the global average pooling and fully connected layer are combined in the classification part. After the experimental verification of Hockey dataset and RWF-2000 dataset, the long-term recurrent convolutional network with attention mechanism proposed in this paper can effectively suppress the interference of background information in dealing with violent behavior under complex background conditions. Compared with the traditional behavior recognition algorithm, the accuracy rate is higher, the generalization ability is stronger, and the violent behavior recognition can be effectively performed.

References

1. Cheng Shilei. Research on Feature Extraction and Recognition Method of Human Behavior in Video Sequence: University of Electronic Science and Technology, (2020).
2. BOBICK A F, DAVIS J W. The Recognition of Human Movement using Temporal Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence. 38(1):142-158. (2016)

3. YANG X D,TIAN Y L. Effective 3D Action Recognition using EigenJoints. Journal of Visual Communication and Image Representation, 25(1):2-11,(2014).
4. WILLEMS G, TUYTELAARS T, GOOL L. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. //European conference on computer vision. Springer, Berlin, Heidelberg: 650-663, (2008).
5. Wang H , Klser A , Schmid C , et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. International journal of computer vision, 103(1):60–79 ,(2013).
6. Wang H , Schmid C. Action Recognition with Improved Trajectories. Proceedings of the 2013 IEEE International Conference on Computer Vision. IEEE, (2013).
7. Simonyan K , Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos. Advances in neural information processing systems, (2014).
8. Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 35(1):221–231,(2012).
9. Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description.//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 2625-2634,(2015).
10. Han K, Wang Y, Tian Q, et al. GhostNet: More features from cheap operations.//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 1580-1589,(2020).
11. Zhang Lianchao, Qiao Ruiping, Dang Qiwei, et al. Spatial attention mechanism with global characteristics. Journal of Xi 'an Jiaotong University,54(11):129-138,(2020).
12. WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module.//Proceedings of the 15th European Conference on Computer Vision. Cham:Springer International Publishing: 3-19,(2018).
13. Roy A G, Navab N, Wachinger C. Concurrent spatial and channel ‘squeeze & excitation’in fully convolutional networks.//International conference on medical image computing and computer-assisted intervention. Springer, Cham: 421-429, (2018).
14. Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3.//Proceedings of the IEEE International Conference on Computer Vision: 1314-1324, (2019).
15. Hochreiter S , Schmidhuber J. Long Short-Term Memory. Neural Computation, 9(8):1735-1780,(1997).
16. SUDHAKARAN S L O. Learning to detect violent videos using convolu-tional long short-term memory.//2017 14th IEEE International Conference on Ad-vanced Video and Signal Based Surveillance (AVSS), IEEE: 1-6, (2017).
17. Ming Cheng, Kunjing Cai, Ming Li. "RWF-2000: An Open Large Scale Video Database for Violence Detection.". arXiv preprint arXiv:1911.05913 (2019).
18. Bermejo N E, Deniz S O, Bueno G, et al. Violence Detection in Video Using Computer Vision Tech-niques.// International conference on Computer analysis of images and patterns, August 29-31, 2011, Seville, Spain. Heidelberg: Springer Berlin Heidelberg: 332-339, (2011).
19. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset.//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 6299-6308. (2017).