Computers, Materials & Continua CMC, vol.65, no.3, pp.2153-2167, 2020

Abnormal Behavior Detection and Recognition Method Based on Improved ResNet Model

Huifang Qian[1], Xuan Zhou[1, *] and Mengmeng Zheng[1]

Abstract: The core technology in an intelligent video surveillance system is that detecting and recognizing abnormal behaviors timely and accurately. The key breakthrough point in recognizing abnormal behaviors is how to obtain the effective features of the picture, so as to solve the problem of recognizing them. In response to this difficulty, this paper introduces an adjustable jump link coefficients model based on the residual network. The effective coefficients for each layer of the network can be set after using this model to further improving the recognition accuracy of abnormal behavior. A convolution kernel of 1×1 size is added to reduce the number of parameters for the purpose of improving the speed of the model in this paper. In order to reduce the noise of the data edge, and at the same time, improve the accuracy of the data and speed up the training, a BN (Batch Normalization) layer is added before the activation function in this network. This paper trains this network model on the public ImageNet dataset, and then uses the transfer learning method to recognize these abnormal behaviors of human in the UTI behavior dataset processed by the YOLO_v3 target detection network. Under the same experimental conditions, compared with the original ResNet-50 model, the improved model in this paper has a 2.8% higher accuracy in recognition of abnormal behaviors on the public UTI dataset.

Keywords: ResNet, abnormal behavior recognition, YOLO_v3, adjustable jump link coefficients model, standard normal distribution.

1 Introduction

At present, domestic and foreign scholars have conducted a lot of research in the detection and recognition of human abnormal behavior, which can be roughly divided into two aspects [Rao, Gubbi, Marusic et al. (2016)]: On the one hand, it is based on the particle flow method to detect human behavior. This type of method assumes that the person is moving under long-term external forces, calculates the interaction force, and sets a threshold to detect human abnormalities. On the other hand is a method based on low-level visual feature extraction. This type of method [Francesco, Simone and Rita (2016)] first utilizes computer vision technology to extract low-level features of the

[1] School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China.

[*] Corresponding Author: Xuan Zhou. Email: zhou_xuan668@163.com.

Received: 01 June 2020; Accepted: 25 June 2020.

human body image, and then uses a classifier to detect and recognize the human body. Among them, the recognition of abnormal behavior of human based on vision [Zhu, Zhu and Xu (2018); Li, Wang and Wang (2014); Feng, Arshad, Zhou et al. (2019)] is more flexible and adaptable. Abnormal behavior recognition methods can be divided into traditional methods and deep learning methods, among which the deep learning method [Herath, Harandi and Porikli (2017)] has stronger robustness and higher recognition accuracy. In particular, CNN (Convolutional Neural Network) performed well, so many classic models were born based on this model, such as AlexNet, VGG (Visual Geometry Group) [Simonyan and Zisserman (2015)], GoogLeNet (Google Inception Network) [Szegedy, Liu, Jia et al. (2015); Ioffe and Szegedy (2015); Szegedy, Vanhoucke, Ioffe et al. (2016)], etc. However, there are still some difficulties in the recognition method of human behavior based on deep learning [Zhu, Zhao, Wang et al. (2016)]: In deep neural networks, the main problem encountered is the vanishing gradient. This problem is an obstacle to building a deep network. If each layer of the deep neural network can be optimized, the error will not increase when the network depth is increased. Based on this assumption, ResNet (Residual Network) [He, Zhang, Ren et al. (2016)] was proposed by identity shortcut connections, which can directly map low-level features to high-level features, and thus can skip defective training layers. ResNet solves the vanishing gradient problem of deep networks to a certain extent [Wang, Jiang, Luo et al. (2019)].

In order to efficiently detect and recognize abnormal behavior in video, it is necessary to capture some appearance and dynamics of the human body in order to detect any abnormal behavior existing in the scene and determine its spatial position. At present, there are many methods for detecting abnormal behaviors, such as SSD (Single Shot MultiBox Detector) [Liu, Anguelov, Erhan et al. (2016)], Fast R-CNN (Fast Regions with CNN features) [Girshick (2015)], Faster R-CNN (Faster Regions with CNN features) [Ren, He, Girshick et al. (2017)], YOLO_v3 and other network structures. Compared with the abnormal behavior detection algorithms based on the deep learning [Bengio, Lecun and Hinton (2015)] model (such as SSD, Fast R-CNN, Faster R-CNN), the detection speed of the YOLO (You Only Look Once) network is the fastest. Because this paper needs to carry out real-time detection of abnormal behavior in the video surveillance system, that is, there are requirements for the detection speed, so this experiment utilizes YOLO_v3 to complete the target detection task. ResNet network is applied in recognizing abnormal behavior. According to the gradient formula of the network, the variation rule of each layer of ResNet can be obtained, which provides a theoretical basis for improving ResNet. Finally, a convex strategy with adjustable jump link coefficients model is proposed, which provides appropriate gradient increment for each layer parameter to improve the performance of ResNet.

2 Target detection algorithm

There three target detection methods: R-CNN, Fast RCNN, and Faster RCNN. They mainly generate a large number of potential bounding boxes that may contain the object to be detected in the candidate area, and then use the classifier to determine whether the potential bounding box contains objects, as well as the probability or confidence of the category to which the object belongs. The characteristic of YOLO is that it treats object

detection as a regression problem. Using the neural network to take the entire picture as input, the position and category of the regression bounding box are directly output in the output layer. As one of the best target detection networks, YOLO can process images up to 45 frames per second in real-time.

2.1 YOLO_v3 network

YOLO_v3 divides the input image into S×S grids [Chen, Sun, Zhang et al. (2019)], and each grid is responsible for predicting boundary boxes with quantity B. Each bounding box contains its own location information (x, y, h, w) and confidence. The confidence represents the confidence that the prediction box contains the target and the accuracy of the prediction of this box. The calculation is as follows:

$$confidence = \Pr(Object) \times IOU_{pred}^{\text{truth}} \tag{1}$$

$$IOU_{\text{pred}}^{truth} = \frac{Detection \cap GroundTruth}{Detection \cup GroundTruth} \tag{2}$$

If the grid contains targets, $\Pr(Object)=1$, otherwise, $\Pr(Object)=0$. $IOU_{pred}^{\text{truth}}$ represents the intersection ratio of the actual bounding box and the predicted bounding box, which is used to measure the accuracy of the predicted bounding box [Qian, Zhou and Zheng (2019)]. $Detection$ is the bounding box predicted by the model. $GroundTruth$ is the bounding box marked in the sample dataset.

In the prediction stage, in order to obtain the confidence score of the category to which the predicted bounding box belongs, the probability of the category to which the target belongs is multiplied by the confidence of the prediction of the bounding box. As shown in Eq. (3):

$$\begin{aligned} Class_i \_ score &= confidence \times \Pr(Class_i \mid Object) \\ &= \Pr(Object) \times IOU_{pred}^{\text{truth}} \times \Pr(Class_i \mid Object) \\ &= \Pr(Class_i) \times IOU_{pred}^{\text{truth}} \end{aligned} \tag{3}$$

After obtaining the confidence score for each category, set a threshold to filter out the category boxes with lower scores to obtain the final detection result. In YOLO_v3, the network is increased to 53 layers, which significantly improves the detection accuracy. Moreover, YOLO_v3 adopts the method of multi-scale fusion to detect the target, which has a good adaptability to the scale change of the target, so the detection effect has been greatly improved. In this paper, the YOLO_v3 network is only used to detect humans in video.

2.2 Human abnormal behavior detection framework based on YOLO_v3

When using target detection technology to detect abnormal behaviors of human in video, its network detection effect is particularly important. Using YOLO_v3 target detection algorithm can not only realize real-time detection, but also ensure the accuracy of multi-scale target detection. With the continuous deepening of the network, the gradient disappears easily during the training process, and the introduction of the residual network can solve this problem well. Combining the features before entering the residual block with the features output by the residual block can extract deeper abnormal behavior feature information. The feature extraction network of YOLO_v3 utilizes darknet-53, and

a residual network is used in this network. The framework of human abnormal behavior detection based on YOLO_v3 is shown in Fig. 1.

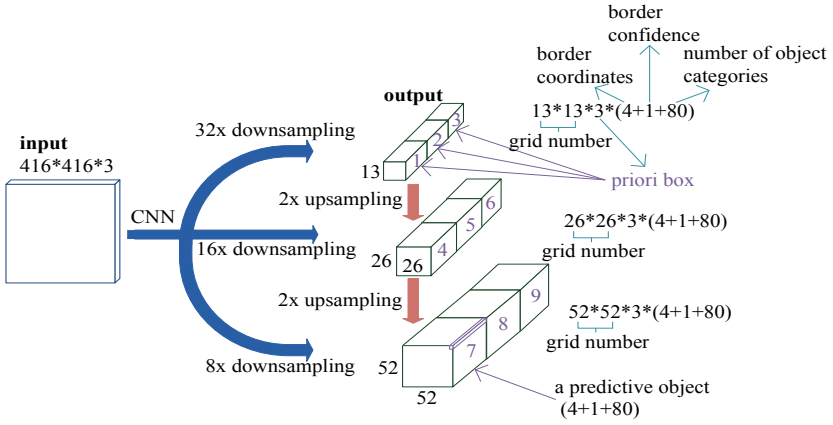


Figure 1: YOLO_v3 abnormal behavior detection framework

Without considering the structural details of the neural network, for the input image, YOLO_v3 maps it to the output tensor of three scales, which indicates the probability of the existence of various objects at each position of the image. For the input image of 416×416, three prior boxes are set in each grid of the feature map of each scale, totaling 10647 predictions. Each prediction is an 85 dimensional vector. The 85-dimensional vector contains border coordinates (4 values), border confidence (1 value), and object class probability (for COCO dataset, there are 80 objects). Compared with YOLO_v2, YOLO_v2 utilizes 845 predictions, and the number of predicted borders of YOLO_v3 increases by more than 10 times, and it is executed at different resolutions, so the mAP and the detection effect on small objects have been improved to a certain extent.

YOLO_v3 network can quickly and accurately detect 80 kinds of target objects, such as cats, dogs, cars, human, etc. In this paper, the detection target of the YOLO_v3 network is changed so that it only detects and extracts the human behavior target, which provides input for the next abnormal behavior recognition network.

3 Abnormal behavior recognition algorithm

There are some difficulties in the method of human abnormal behavior recognition based on deep learning: in deep neural networks, the main problem encountered is the disappearance of gradients. ResNet solves this problem well through identity shortcut connections. Based on the original ResNet-50 network, this paper optimizes the convolution kernel of its network, and introduces the BN layer and adjustable jump link coefficients model to improve the accuracy of recognizing abnormal behavior.

3.1 Improved ResNet

This paper improves its network structure on the basis of the original ResNet-50 network. The entire network structure includes four parts: the improved ResNet module, the global average pooling layer, and two fully connected layers, as shown in Fig. 2. The entire

network has fifty convolutional layers, one global average pooling layer and two fully connected layers.

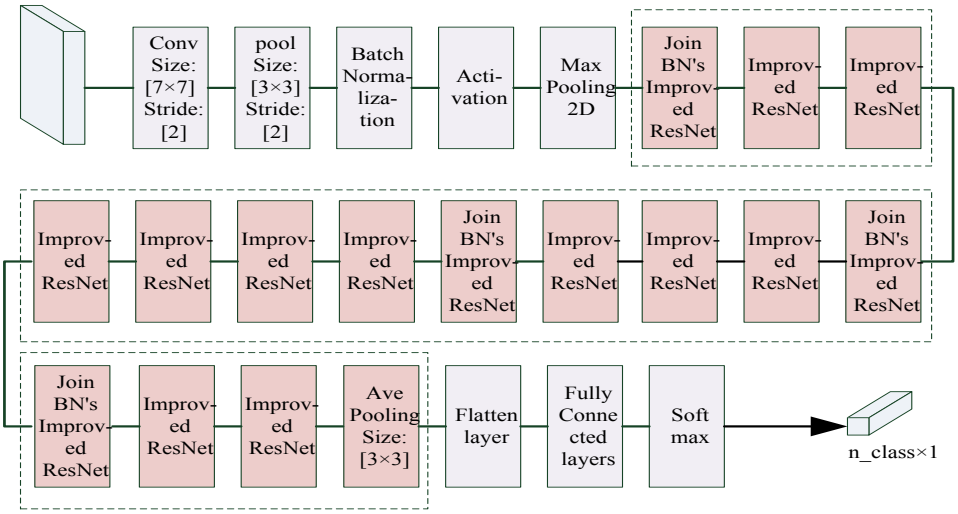


Figure 2: The improved ResNet overall network structure diagram

3.1.1 The improved ResNet block

This paper optimizes the size of the convolution kernel in the residual block based on the original residual network. In addition, the convolution kernel of 1×1 size is added to achieve the purpose of reducing the number of parameters, furthermore, improving the speed of operation.

For x of the input network, assuming that $H(x)$ (the output of the residual block) is the optimal solution mapping, $F(x)=H(x)-kx$ is constructed, so that $H(x)$ is expressed as $F(x)+kx$, the specific expression is:

$$\mathrm{F(x)=W_3\sigma(W_2\sigma(W_1x))} \tag{4}$$

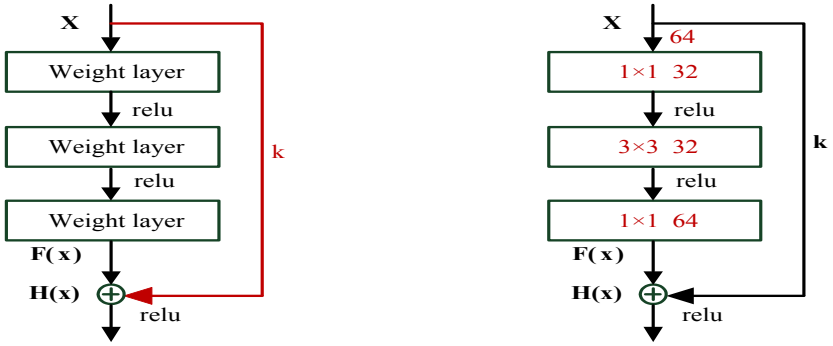
The final output of the residual block is:

$$\mathrm{H(x)=W_3\sigma(W_2\sigma(W_1x))}+kx \tag{5}$$

where, $F(x)$ represents the output of the residual block before the activation function of the third layer, and W_1, W_2, and W_3 represent the weights of the first, second, and third layers, and σ represents the linear activation function *Relu*. The schematic diagram of the improved residual block is shown in Fig. 3(a).

Eq. (5) demonstrates that even if the weight $W_i \approx 0$, the final output $H(x)$ will not be 0. Since the number of main and bypass channels is different, the parameter k that does the linear operation is adjusting the dimension of x by the convolution operation. GoogLeNet first proposed a 1×1 size convolution kernel. This kernel can realize the reduction and addition of the dimensionality of convolution kernel channels, thereby reducing the number of parameters and accelerating the training speed. As shown in Fig.

3(b), this paper first reduced the input of 64-dimensional to 32-dimensional by 1×1 32 convolution, and finally restored to 64-dimensional by 1×1 64 convolution.



(a) (b)

Figure 3: Basic structure of improved ResNet: (a) Residual block of improved ResNet; (b) The number of network channels for the improved ResNet.

3.1.2 Add batch normalization layer

In order to reduce the noise of the data edge, improve the accuracy of the data and speed up the training, normalization processing (BN layer) is added before the activation function *relu* in this paper, as shown in Fig. 4. After the BN layer is added, the *dropout* layer can also be without used to enhance the generalization ability of the network.

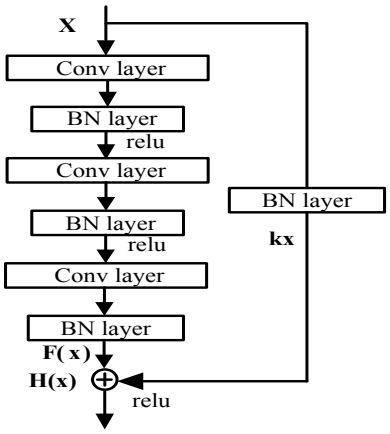


Figure 4: Improved residual block with BN layer added

3.1.3 Adjustable jump link coefficients model

In order to solve the following problems: how to find which layers in the network can't get the ideal training, the reasons that cause the training of some layers to be unsatisfactory, how to deal with the undesirable network layers, etc. Li et al. [Li and He (2018)] have found that the parameters of the middle layer were not easy to train by analyzing the

forward and backward propagation of ResNet, and proposed to provide different gradient increments for different layer parameters with different change rules. This paper sets an effective coefficient k for the adjustable jump links in the improved ResNet block to efficiently recognize abnormal behaviors. By considering the changing rules of ResNet parameters in different regions, this paper proposes to take 11 data of its x value in the form of arithmetic progression based on the standard normal distribution. The strategy in this paper is more suitable for improving the recognition performance of ResNet.

3.1.4 Global average pooling layer and fully connected layer

In Fig. 2, there is a global average pooling layer, which has the following characteristics: (1) The convolution structure is simpler by enhancing the consistency of the feature map and the category. (2) No parameter optimization is required, so this layer can be avoided overfitting. (3) It sums the spatial information, so it is more stable than the input spatial transformation. Unlike the traditional fully connected layer, we perform global average pooling on the entire image of each feature map, so that each feature map can be output.

There are two fully connected layers in Fig. 2. The first fully connected layer is the output of 1024 1×1 nodes, and is input to the classifier with six nodes in the second layer.

3.2 The overall framework for detecting and recognizing abnormal behaviors

The purpose of this study is to build a system that first detects the human body in the video and then recognizes the abnormal behavior of human. The YOLO_v3 detection method proposed in the second part of the target detection algorithm is used to detect and extract the target object of each frame from the continuous video clip, and then use the improved ResNet proposed above to recognize the human behavior target extracted in the previous step. The two neural network models are connected to form an overall framework for detecting and recognizing abnormal behaviors, as shown in Fig. 5.

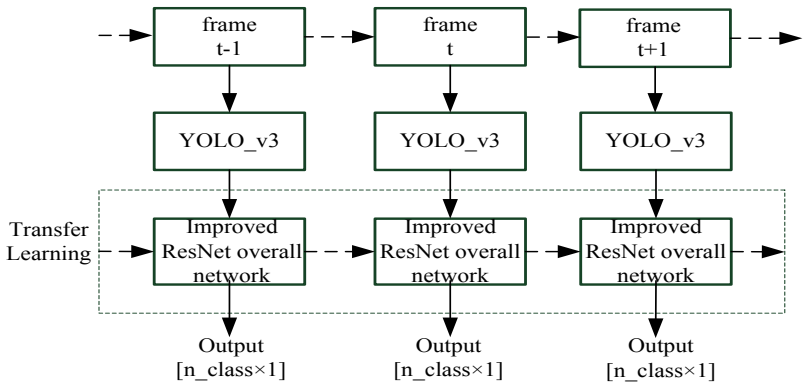


Figure 5: The overall framework for detecting and recognizing abnormal behaviors

Among them, frame t-1, frame t, and frame t+1 represent the images of frame t-1, frame t, and frame t+1 in the input video, respectively. The YOLO_v3 network is only used to detect human form, while the improved ResNet overall network is used to classify the input images into categories of abnormal behavior defined during training. The improved

ResNet overall network receives the image inside the bounding box output by the YOLO_v3 network, which includes information about each human body and behavior status. Among them, the YOLO_v3 network will frame the human body and cut out the human target according to the coordinate values of the bounding box, and then use the improved ResNet overall network to recognize abnormal behaviors.

The UTI behavior dataset is utilized to evaluate the performance of the improved residual network model proposed in this paper, in which the YOLO_v3 network will frame the human body and cut out the human target according to the coordinate values of the bounding box, and then utilize the improved ResNet network to recognize abnormal behavior. The experimental process and results are shown in Fig. 6.

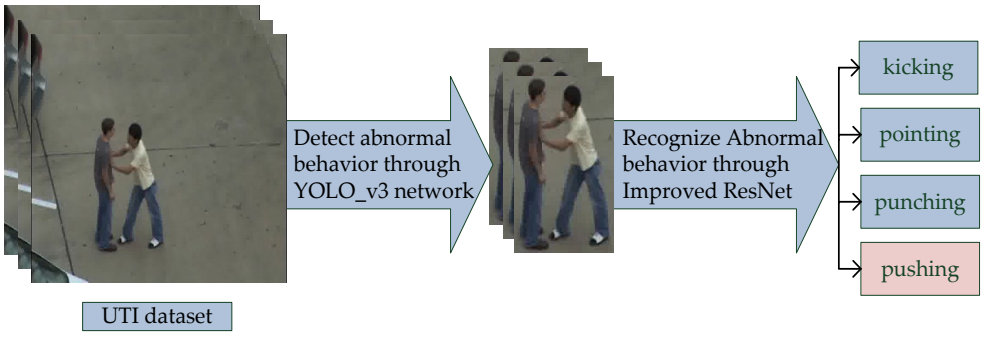


Figure 6: Model process and results of detecting and recognizing abnormal behaviors

4 Experimental results and analysis

4.1 Experimental environment and dataset

Experimental configuration: CPU is Inter(R) Core(TM) i9-9900x, the main frequency is 3.5GHz, the memory is 16GB, the mechanical hard disk is 4TB, the solid state hard disk is Inter SSD D3-S4510 480GB, the graphics card is NVIDIA RTX 2080TI 11G turbo card. On this basis, the software environment is built: Ubuntu16.04, Python3.6, Opencv, Cuda, etc. The framework uses Tensorflow framework and Keras framework.

In this experiment, the public ImageNet dataset is first used as the input of the improved ResNet network model proposed above, and the network is trained. Then, the trained network is used to conduct behavior recognition on the videos in the UTI [Ryoo and Aggarwal (2010)] behavior dataset. Among them, the ImageNet dataset is used by LSVRC-2016 competition. This dataset contains 2 million color images, the training set contains 1.65 million images, and the test set contains 350,000 images. This dataset is divided into 1000 categories, the major categories include: animal, appliance, bird, device, fish, flower, food, fruit, furniture, geometric formation, musical instrument, plant, sport, tool, tree, vehicle, person, etc. The UTI dataset consists of 20 video clips, the train dataset contains 16 videos, and the test dataset contains 4 videos. Each video clip contains multiple moving human bodies, including the following six different behaviors: “hand shaking”, “hugging”, “kicking”, “pointing”, “punching” and “pushing”. Each video contains at least one of these six interactions, resulting in an average of eight human activities per video clip. The resolution image frame corresponds to HD (720×480), and the frame rate corresponds to

30fps (frames per second). In this paper, the four behaviors of "kicking", "pointing", "punching" and "pushing" are defined as abnormal behaviors, and the others are normal behaviors. However, due to the small scale of this dataset makes it difficult to train a robust network to detect and recognize abnormal behaviors. For this reason, this paper adopts the method of transfer learning [Shao, Zhu and Li (2015)] to train the improved ResNet. Yosinski et al. [Yosinski, Clune, Bengio et al. (2014)] demonstrated the effectiveness of feature transfer in transfer learning. In the experiment, the UTI dataset is used as the target domain, and the ImageNet dataset is used as the source domain, and the fine-tuning method is used to transfer the model trained by the improved ResNet. The UTI behavior dataset is used to evaluate the performance of the overall network model proposed in this paper for detecting and recognizing abnormal behaviors.

4.2 Setting of experimental parameters

In the experiment, the learning rate is set as 0.01, the SGD (Stochastic Gradient Descent) random gradient descent method is adopted by the optimizer, where the momentum is set as 0.9, and L2 regularization is used to realize weight attenuation to avoid the network falling into overfitting.

4.3 Experimental contents and results analysis

4.3.1 Add batch normalization layer

In the ResNet-50 network model trained on the ImageNet dataset, the Epoch is set to 500 and the Batch size is set to 128. When using the transfer learning method to recognize abnormal behaviors in the UTI behavior dataset, the Epoch is set to 20 and the Batch size is set to 32. The loss values and the accuracy of the train dataset and the test dataset are shown in Fig. 7.

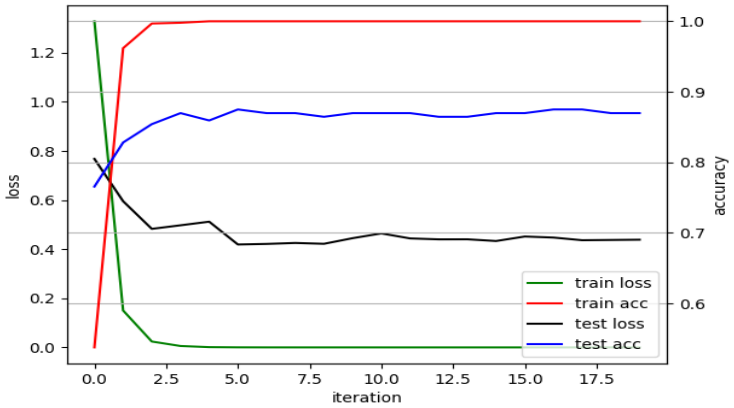


Figure 7: The recognition rate of the ResNet-50 network model for abnormal behavior

4.3.2 Experimental results of the improved ResNet model

Based on the original ResNet-50 network, this paper optimizes the convolution kernel in the residual block, and the BN layer and adjustable jump link coefficients model are introduced to construct an improved ResNet for the ImageNet and UTI datasets. From the

improved ResNet overall network structure diagram (Fig. 2), it can be seen that 11 adjustable jump link coefficients need to be set. Li et al demonstrated the effectiveness of the adjustable jump link coefficients changing with convex trend.

On this basis, this paper sets convex values of different variation trends for k in different regions of the network in order to better recognize abnormal behaviors.

Selecting a k value which is far too small will lead to inaccurate extraction of features in the network layer, otherwise, it will lead to excessive redundancy of extracted features. In view of the experience, the k value selected in this paper range from 0.5 to 1.5, and five different strategies are used in the improved ResNet. The first strategy is called convex 1. This strategy adopts the idea of arithmetic progression. The expression formula of arithmetic progression is:

$$k_n = \mathrm{k}_1 + (n-1)d \tag{6}$$

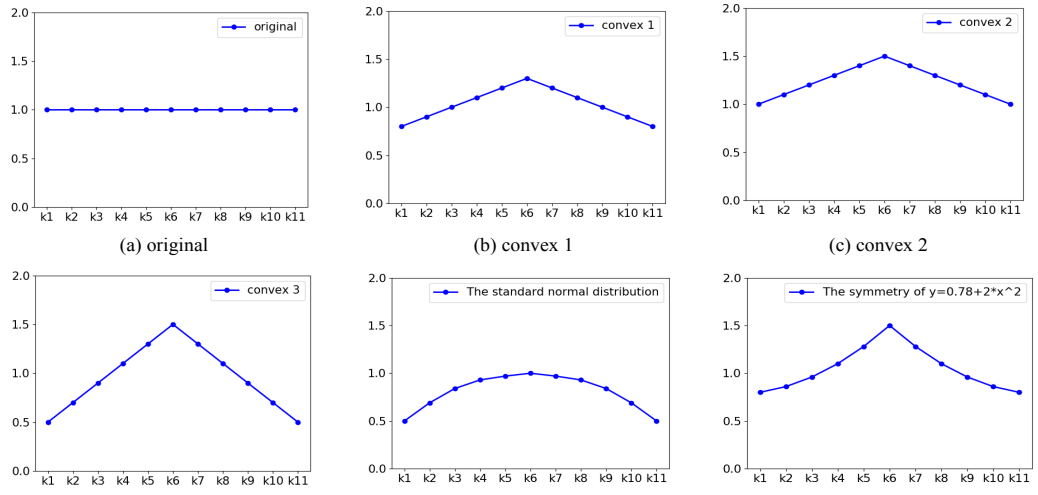
In convex 1, the first term k_1 is 0.8, and the common difference d is 0.1. After taking 6 values in turn, the common difference d becomes -0.1 and then 5 data, that is, the k values are 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.2 1.1, 1.0, 0.9, 0.8. The second strategy, called convex 2, takes 6 values in turn in the form of an arithmetic progression $\mathrm{k}_n = 1.0 + 0.1 \times (n-1)$ and then changes the common difference d to -0.1 and then takes 5 data. The third strategy, called convex 3, takes 6 values in turn in the form of arithmetic progression $\mathrm{k}_n = 0.5 + 0.2 \times (n-1)$ and then changes the common difference d to -0.2 and then takes 5 data. The fourth strategy is the standard normal distribution, and its expression formula is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{7}$$

Take 6 values of its x value in the form of an arithmetic progression $\mathrm{x}_n = 0 + 0.5 \times (n-1)$ and then change the common difference d to -0.5 and then take 5 data. The corresponding function values of its x value are 0.5, 0.69, 0.84, 0.93, 0.97, 1.0, 0.97, 0.93, 0.84, 0.69, 0.5, respectively. The fifth strategy is obtained on the basis of quadratic function, and its expression formula is as follows:

$$k = 0.78 + 2 \times x^2 \tag{8}$$

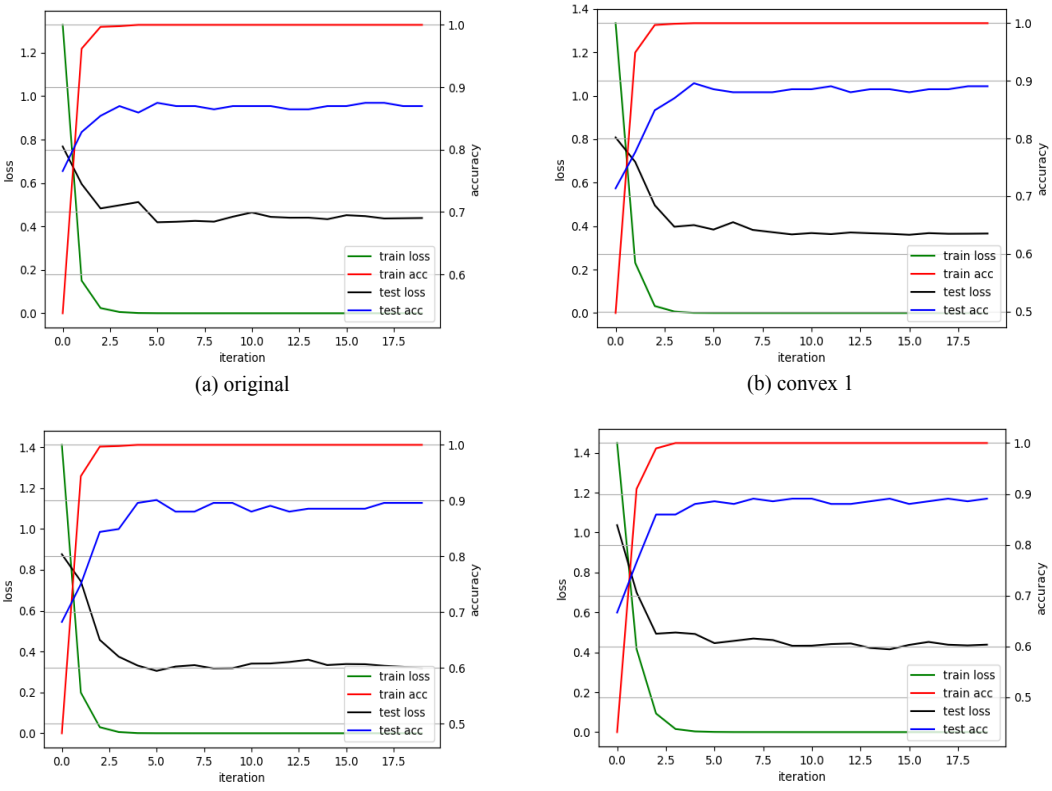
Take 6 values of its x value in the form of an arithmetic progression $\mathrm{x}_n = 0.1 + 0.1 \times (n-1)$ and then change the common difference d to -0.1 and then take 5 data. The corresponding function values of its x value are 0.8, 0.86, 0.96, 1.1, 1.28, 1.5, 1.28, 1.1, 0.96, 0.86, 0.8, respectively. Fig. 8 shows six different k strategies, the first of which is the k value of the original ResNet-50 network model. In order to better analyze these six strategies, the characteristics of each of them applied in the improved ResNet are regarded as reference. Similarly, all models utilize the ImageNet dataset for 500 iterations, and then use the transfer learning method to perform 20 iterations on the UTI behavior dataset. Finally, the classification accuracy of the UTI abnormal behavior dataset is used to evaluate the performance of six different k strategies. The loss values and the accuracy of the train dataset and test dataset are shown in Fig. 9.



(a) original (b) convex 1 (c) convex 2

(d) convex 3 (e) The standard normal distribution (f) The symmetry of y = 0.78+2*x^2

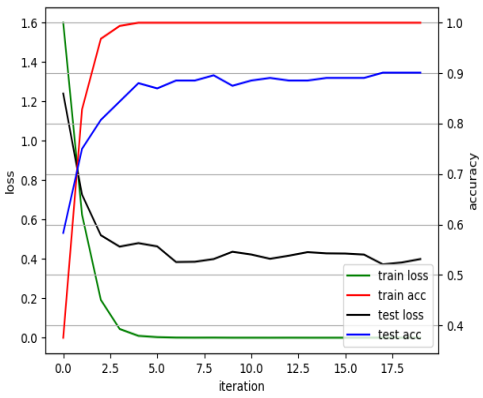
Figure 8: Six different *k* strategies



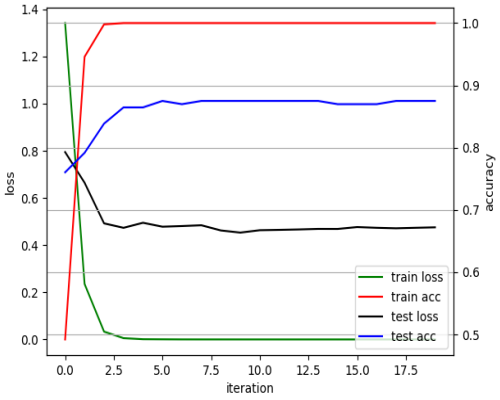
(a) original (b) convex 1

(c) convex 2 (d) convex 3

Figure 9: The accuracy of six different *k* strategies in recognizing abnormal behaviors



(e) The standard normal distribution



(f) The symmetry of y = 0.78+2*x^2

Figure 9: The accuracy of six different *k* strategies in recognizing abnormal behaviors (continued)

By comparing Figs. 8 and 9, it can be seen that after the value of x is sequentially taken as 6 values in the form of the arithmetic progression $x_n = 0 + 0.5 \times (n-1)$, then the common difference d is changed to -0.5 and then 5 data are taken. The adjustable jump link coefficients model that changes in the form of the standard normal distribution corresponding to the x value has the test accuracy rate is the highest and the fluctuations is small during the iteration process, which shows the effectiveness and stability of this model. The experimental results are shown in Tab. 1.

Table 1: Comparison of experimental results

Strategy	**k_i**		**Testing accuracy (UTI)**
	$[k_1,k_2,k_3,k_4,k_5]$	**$[k_6,k_7,k_8,k_9,k_{10},k_{11}]$**	
original	[1.0,1.0,1.0,1.0,1.0]	[1.0,1.0,1.0,1.0,1.0,1.0]	87.2%
convex 1	[0.8,0.9,1.0,1.1,1.2]	[1.3,1.2,1.1,1.0,0.9,0.8]	89.2%
convex 2	[1.0,1.1,1.2,1.3,1.4]	[1.5,1.4,1.3,1.2,1.1,1.0]	89.1%
convex 3	[0.5,0.7,0.9,1.1,1.3]	[1.5,1.3,1.1,0.9,0.7,0.5]	89.1%
The standard normal distribution	[0.5,0.69,0.84,0.93,0.97]	[1.0,0.97,0.93,0.84,0.69,0.5]	90%
The symmetry of $y = 0.78+2 \cdot x^2$	[0.8,0.86,0.96,1.1,1.28]	[1.5,1.28,1.1,0.96,0.86,0.8]	87.5%

Tab. 1 demonstrates that the highest accuracy of the original ResNet-50 to recognize abnormal behaviors is 87.2%. While the adjustable jump link coefficients model in the improved ResNet utilizes the first convex strategy to recognize abnormal behaviors, the highest accuracy rate is 89.2%. The highest accuracy of the second convex strategy and the third convex strategy is 89.1%, while the highest accuracy of the fourth convex strategy (The standard normal distribution) is 90%, and the highest accuracy of the fifth

convex strategy (The symmetry of $k=0.78+2\cdot x^2$) is 87.5%. The experimental data shows that the adjustable jump link coefficients model in the improved ResNet uses the fourth convex strategy (The standard normal distribution) proposed in this paper can effectively improve the accuracy of abnormal behavior recognition. And on the basis of the original ResNet-50, the recognition accuracy of this model has increased by 2.8%. Therefore, the improved ResNet framework proposed in this paper is very valuable.

5 Conclusions

In this paper, an improved network model structure based on ResNet is constructed, from the perspective of obtaining effective features of pictures and unsatisfactory recognition effect of abnormal behavior. The YOLO_v3 network in this model utilizes feature fusion to perform multi-scale target detection, which can quickly and accurately extract human targets, and cut out human targets according to the coordinate values of the bounding box to remove redundant information in the background. This paper proposes to embed an adjustable jump link coefficient model that changes in the form of a standard normal distribution into an improved ResNet, and then constructs a new network model. The experimental data shows that the new network model proposed in this paper can effectively avoid the gradient vanishing problem caused by the deep network, and effectively improves the accuracy of the network in recognizing abnormal behavior. The disadvantage of this paper is that the problem of recognition errors has not been solved. The next step is to optimize the algorithm to solve such problems, thereby further improving the accuracy of recognizing abnormal behaviors. Compared with the original ResNet-50 model, the new network model proposed in this paper is 2.8% more accurate in recognizing abnormal behavior on the public UTI dataset.

Acknowledgement: We especially thank the experimental equipment provided by the 212 laboratory of the School of Electronics and Information, Xi'an Polytechnic University.

Funding Statement: This research was funded by the Science and Technology Department of Shaanxi Province, China, Grant Number 2019GY-036.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Bengio, Y.; Lecun, Y.; Hinton, G. (2015): Deep learning. *Nature*, vol. 52, no. 1, pp. 436-444.

Chen, Y. Q.; Sun, L. Q.; Zhang, Y. Z.; Fu, Q. M.; Lu, Y. et al. (2019): Research on bird detection technology of transmission line based on YOLO v3. *Computer Engineering*. vol. 46, no. 4, pp. 294-300.

http://kns.cnki.net/kcms/detail/31.1289.TP.20190809.1140.013.html.

Feng, C.; Arshad, S.; Zhou, S.; Cao, D.; Liu, Y. (2019): Wi-Multi: a three-phase system for multiple human activity recognition with commercial WiFi devices. *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 7293-7304.

Francesco, S.; Simone, C.; Rita, C. (2016): Socially constrained structural learning for groups detection in crowd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 5, pp. 995-1008.

Girshick, R. (2015): Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440-1448.

He, K.; Zhang, X.; Ren, S. Q.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Herath, S.; Harandi, M.; Porikli, F. (2017): Going deeper into action recognition: a survey. *Image and Vision Computing*, vol. 60, pp. 4-21.

Ioffe, S.; Szegedy, C. (2015): Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448-456.

Li, B.; He, Y. (2018): An improved ResNet based on the adjustable shortcut connections. *IEEE Access*, vol. 6, pp. 18967-18974.

Li, R. F.; Wang, L. L.; Wang, K. (2014): A survey of human body action recognition**.** *PR&AI*, vol. 27, no. 1, pp. 35-48.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. et al. (2016): SSD: Single shot multibox detector. *Proceedings of European Conference on Computer Vision*, pp. 21-37.

Qian, H. F.; Zhou, X.; Zheng, M. M. (2019): Detection and recognition of abnormal behavior based on multi-level residual network. *IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference*, pp. 2572-2579.

Rao, A. S.; Gubbi, J.; Marusic, S.; Palaniswami, M. (2016): Crowd event detection on optical flow manifolds**.** *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1524-1537.

Ren, S. Q.; He, K. M.; Girshick, R.; Sun, J. (2017): Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 1, pp. 1137-1149.

Ryoo, M. S.; Aggarwal, J. (2010): UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). *IEEE International Conference on Pattern Recognition Workshops*, vol. 2, pp. 4.

Shao, L.; Zhu, F.; Li, X. (2015): Transfer learning for visual categorization: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019-1034.

Simonyan, K.; Zisserman, A. (2015): Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations*, pp. 1-14.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. et al. (2015): Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. (2016): Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826.

Wang, W.; Jiang, Y. B.; Luo, Y. H.; Li, J.; Wang, X. et al. (2019): An advanced deep residual dense network (DRDN) approach for image super-resolution. *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1592-1601.

Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. (2014): How transferable are features in deep neural networks? *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 3320-3328.

Zhu, H. L.; Zhu, C. S.; Xu, Z. G. (2018): Research advances on human activity recognition datasets. *Acta Automatica Sinica*, vol. 44, no. 6, pp. 978-1004.

Zhu, Y.; Zhao, J. K.; Wang, Y. N.; Zheng, B. B. (2016): A review of human action recognition based on deep learning. *Acta Automatica Sinica*, vol. 42, no. 6, pp. 848-857.