Convolutional Neural Networks for Identifying Human Behavior

游泽平 36220201154082, 程晨 36220201154077
陈姚伶 2302020115373, 赵佳驰 23020201153826

Abstract

Human behavior recognition technology is becoming more and more popular. The focus of many deep learning network models is just on improving accuracy. Lightweight models are therefore of particular importance. A lightweight convolutional neural network (LWCNN) was designed specifically for human behavioural recognition tasks. This paper then proposes a combined training strategy approach to train the network, including pre-training, fine-tuning training and migration training, to optimize the LWCNN parameters. On the basis of this,this method can effectively reduce computational complexity compared to existing methods while maintaining identification performance.The accuracy rate of the network model in our data set reached 72.34%, and the experimental results proved that the compressed model can accurately and objectively identify human behavioral images.

Introduction

Nowadays people use different types of cameras to capture a large number of videos and pictures every day. In some interactive scenes, it is necessary to recognize different human behaviors to identify whether the human body is maintaining the correct posture. For example, we can recognize the fall of the old or children by putting a camera at home.

Video behavior recognition refers to the process that video is handed over to computer to judge what operation people or interested objects are doing. Cameras are widely used in video data recording of all walks of life, such as traffic management, recording daily life etc. Video data has become more and more important in most industries. Therefore, the research of video recognition based on deep learning emerges . Especially in recent years, with the vigorous development of technology and the substantial improvement of computer computing power, the human behavior recognition technology based on deep learning has attracts more and more attention.

Howerver,recognizing human behavior is a challenging task because of problems such as background clutter, partial occlusion, changes, etc. Developing a fully automated human activity recognition system is a difficult task.

Related Work

Most researches on human action recognition focus on human detection and motion tracking. The first step is to extract the person from the video.The Gaussian mixture model is the most effective way to cut out the background in a noisy background[10]. CheroGN took pictures as input, extracted different features in the hidden layer, and realized the recognition of static pictures.This method is not suitable for video[2].To solve this problem, Siminyan proposed two-StreamCNN model, which input the original im-

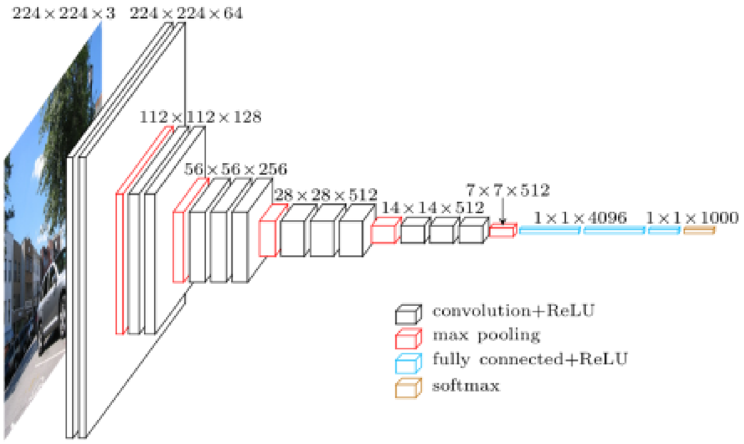

Figure 1: VGG16

age and optical flow image into the network, and extract the apparent features of the human body from the original image, and extract the dynamic information of the behavior from the optical flow image[9].Large-scale data sets make the structure and parameters of CNN network too complex. IjinaEP proposes to use genetic algorithm to initialize CNN network parameters to improve the accuracy[7].To solve the problem that continuous movement cannot be recognized, Kiu C et al. combined CNN and conditional random field to build a human action recognition model based on convolutional neural random field[12]. After years of exploration and research by researchers, the technology of human behavior recognition has made great progress. With the rapid development of computers and the popularization of vision-related sensors, the massive data sets generated from the network and in real life also provide a large amount of cheap preliminary materials for video behavior recognition, which has also laid a solid foundation for the development of this technology Foundation.

VGG16[8] is a convolutional neural network model proposed by A. zisserman of Oxford University in the paper "deep neural network in large scale image recognition" . The model achieves 92.7% top-5 test accuracy in Imagenet.Its layers include convolution layer, maximum pooling layer, activation layer and full connection layer,which is shown in figure1. It is worth noting that there are three full connection layers, which are also the areas to be improved in this paper.

Proposed Solution

We want to use the collected human behavior data and train a model to identify dynamic or static images of different behavior categories. The human behavior recognition algorithm based on deep learning that has emerged over the years solves the problem of low accuracy. It uses deep neural networks to extract characteristics of human behavior types and uses BP[1] neural networks to perform loss regression, thereby greatly improving Accuracy. But this also has some shortcomings, that is, the amount of calculation caused by deep learning, which leads to a certain degree of decline in recognition speed.

Model architecture

In recent years, increasingly complex CNN architectures have been proposed to improve the performance of large-scale image recognition tasks. However, the improvement of recognition performance comes at the cost of a large increase in computing and storage resources. In order to achieve a good trade-off between performance and computational complexity, we want to design a lightweight CNN(LWCNN) network architecture for human behavior recognition.It consists of 13 convolution layers, 1 global pooling layer and 2 full connection layers, with a total of 16 layers. In order to reduce the number of parameters of the fully connected layer, we specially designed a global average pooling layer between the last convolutional layer and the first fully connected layer.In this way,each feature map can only output an average value, which is equivalent to a dimension reduction operation.This can greatly reduce the number of network parameters and speed up training, avoiding the risk of overfitting.

Training Methods

Random initialization is a common way to set CNN model parameters. The experimental results[11] show that the final classification effect is better than the method of parameter random initialization when the unsupervised learning method is used for pre training and the supervised learning method is used for fine tuning.The combination of pre training and fine tuning is the most commonly used strategy to optimize CNN model parameters. However, the use of large data sets for pre training usually consumes a lot of time and computing resources. In order to solve this problem, this paper proposes a transfer training[6] method for human behavior recognition. The strategy of combination training including pre training, fine-tuning training and transfer training is used in this paper. In this way, the parameters of the network are initialized based on a more complex model.Due to the different network structures of LWCNN and VGG, the number of parameters in each layer does not match. A parameter selection method is proposed to overcome this problem.

- **Pre training:** VGG16 net is trained by supervised learning on Pascal VOC dataset, and the initial weight coefficient of the network is obtained.
- **Fine-tuning training:** Fine tuning is performed on the collected data sets. And only the last fully connected layer is adjusted in the fine-tuning process.
- **Transfer training:** Parameter extraction is mainly used to extract the parameter weight w and offset B of the corresponding nodes in each layer from the initial model 2 to obtain the parameter matrix W (h, W, C, n) and offset B (n), where w × h is the size of a single convolution kernel, C is the number of input channels, and N is the number of convolution cores in the current layer. After parameter extraction, 16 corresponding parameter matrices and offset vectors can be obtained.

S_i is defined as the average correlation coefficient and its formula is as follows:

$$S_i = \sum_{j=1}^{n} |R_{ij}|, \qquad i, j = 1, 2, ..., n$$

Where n is the number of convolution kernels in the current layer.

R_{ij} is the correlation coefficient between the i_{th} kernel parameter and the j_{th} kernel parameter. The calculation process can be expressed as:

$$R_{ij} = \frac{Cov(w_i, w_j))}{\sqrt{Cov(w_i, w_i)Cov(w_j, w_j)}}, i, j = 1, 2, ..., n$$

Where $w_i, w_j \in$ W, W is the parameter matrix of convolution layer and N is the number of convolution kernels

Then, all convolution kernels of each layer are sorted by S_i value. The first m kernel parameters are selected for LWCNN initialization (M is the number of convolution kernels in the corresponding layer of LWCNN) weight

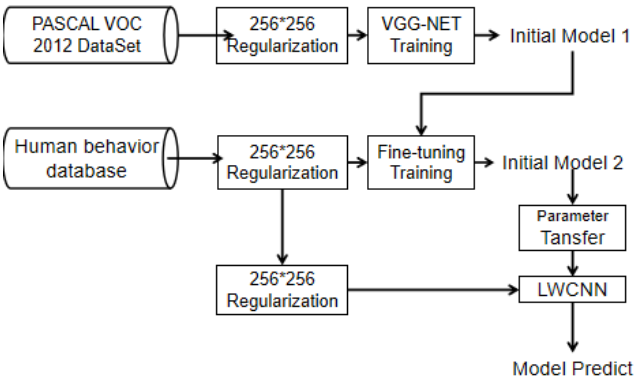

figure 2: Algorithm Process

coefficient. After the parameter selection is completed, the final classification model is obtained by network training based on human behavior data set using the principle similar to fine tuning. This method can ensure that the output of convolution neurons has approximate distribution, thus improving the convergence. Training speed and recognition performance of network model.

Algorithm Process

The human behavior recognition method proposed in this chapter includes two parts. (1) In order to reduce the number of network parameters and reduce the demand for computing and storage resources, a lightweight convolutional neural network for human behavior recognition is designed. (2) A joint training strategy, including pre training, fine tuning training and migration training, is proposed to optimize the network parameters to improve the recognition performance of deep CNN model.

In the pre training stage, the image size of Pascal VOC dataset is regularized to 256 × 256, and then VGG16 net is used for training to obtain the initial model 1 (including network structure and weight). In the fine-tuning stage, the image size of human behavior dataset is regularized to 256 × 256, and then the regularized image is used to fine tune the training of initial model 1. The weight of initial model 1 is updated to generate initial model 2. In the migration training stage, through the analysis of the initial model 2, the appropriate initialization parameters of LWCNN are obtained. The normalized human behavior images are trained to get the final classification model. In the recognition stage, the input image is fed back to the classification model for recognition.

Layer	VGG-16	LWFCNN	LWCNN
Input	224×224×3	224×224×3	224×224×3
Layer 1	Conv,64	Conv,32	Conv,32
Layer 2	Conv,64	Conv,32	Conv,32
Layer 3	Conv,128	Conv,48	Conv,48
Layer 4	Conv,128	Conv,48	Conv,48
Layer 5	Conv,256	Conv,64	Conv,64
Layer 6	Conv,256	Conv,64	Conv,64
Layer 7	Conv,256	Conv,64	Conv,64
Layer 8	Conv,512	Conv,96	Conv,96
Layer 9	Conv,512	Conv,96	Conv,96
Layer 10	Conv,512	Conv,96	Conv,96
Layer 11	Conv,512	Conv,128	Conv,128
Layer 12	Conv,512	Conv,128	Conv,128
Layer 13	Conv,512	Conv,128	Conv,128
Layer 14	Fc,4096	Fc,1024	GAP
Layer 15	Fc,4096	Fc,1024	Fc,1024
Layer 16	Fc,281	Fc,281	Fc,281
Memory size	513M	170.4M	27M
Parameter quantity	1.7×10^8	4.7×10^7	1.3×10^6
Accuracy	80.05%	70.07%	72.08%

Table 1: Performance comparison results

Experiments

Influence of GAP

In order to illustrate the effect of global mean pooling layer replacing the full connection layer of vgg16-net on human behavior recognition performance, we compared it with vgg16-net initial model 1 (VGG-16). Through pre training and fine tuning methods, we got the initial model 2 (LWFCNN) of VGG16-net classification model, and compared it with VGG16-net classification model

Method	Model size	Accuracy
ResNet[3]	37M	59.47%
MobileNet[4]	30M	63.34%
DesNet[5]	30.8M	67.03%
LWCNN	27M	72.08%

Table 2: Experimental results of different methods

(initial model 1) Compared with the classification model of VGG16-net, the performance of human behavior recognition is reduced to 10%, and the accuracy is reduced by 10%. Then, a comparative experiment was carried out. The configuration of these experiments is as follows. Firstly, all the convolution layers of LWFCNN are reserved and initialized with the same parameters; then, the 14_{th} layer of LWFCNN is changed to the global mean pooling layer. Finally, the combination of the two parts forms the final network, called LWCNN. The comparison results of different models are shown in Table 1.Compared with VGG16-net model, the recognition accuracy of LWCNN proposed in this paper only decreases about 8%. However, the number of network parameters and the memory occupied by the model are less than 10%, which greatly reduces the requirements of computing and storage resources in the process of network training and testing. Compared with the original LWCNN model, the proposed method can achieve a recognition accuracy of more than 2%, and the memory occupancy rate of both network parameters and model is less than 30%. It can be concluded that the designed lightweight CNN network adopts the global mean pool layer, which can effectively realize the convolution feature fusion and greatly reduce the complexity of the network.

Comparison of Different Methods

In order to verify the effectiveness of the proposed method, we compared it with the current popular lightweight recognition methods. The experimental results of different recognition methods are shown in Table 2

Conclusion

Based on the idea of global average pooling, this paper makes some improvements on the network structure of VGG. On the premise of ensuring the accuracy, it achieves the purpose of compressing the size of the model and improving the speed of calculation. The idea of how to change the parameters of the network efficiently is proposed, that is, how to maximize the compression capacity of the deep neural network as far as possible under the condition of ensuring the accuracy of the network, so as to meet our needs. We should not only keep the depth of the network constant, but also ensure the minimum weight of the network. Finally, under the proposed training strategy, the network compression is successfully completed, which has certain reference significance for the follow-up practitioners.In this paper, a human behavior recognition method based on lightweight convolutional neural network and combined learning strategy is proposed. Compared with other methods, the proposed method can effectively reduce the complexity and maintain the network performance. In the future work, a new parameter selection method can be designed to further improve the recognition performance.

References

[1] Timur Ash. "Dynamic node creation in backpropagation networks". In: *Connection science* 1.4 (1989), pp. 365–375.

[2] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. "P-CNN: Pose-based CNN Features for Action Recognition". In: *IEEE International Conference on Computer Vision*. 2015.

[3] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[4] Andrew G Howard et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[5] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[6] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[7] Earnest Paul Ijjina and C. Krishna Mohan. "Human action recognition using genetic algorithms and convolutional neural networks". In: *Pattern Recognition* (2016), pp. 199–212.

[8] Hussam Qassim, Abhishek Verma, and David Feinzimer. "Compressed residual-VGG16 CNN model for big data places image recognition". In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. 2018.

[9] Karen Simonyan and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos". In: *Advances in Neural Information Processing Systems* 1 (2014).

[10] C. Stauffer. "Adaptive Background Mixture Model for Real-Time Tracking". In: *Proc Computer Vision Pattern Recognition* 2 (1998), p. 2246.

[11] Nima Tajbakhsh et al. "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1299–1312.

[12] Zhai et al. "Convolutional neural random fields for action recognition". In: *Pattern Recognition the Journal of the Pattern Recognition Society* 59 (2016), pp. 213–224.