Two Stream Convolution Fusion Network based on Attention Mechanism

Shengbo Chen¹,*, Yuzhi Wang¹ and Zhou Lei¹
¹ School of Computer Engineering and Science Shanghai University, Shanghai University, Shanghai, Shanghai, 200444, China
*Corresponding author’s e-mail: schen@shu.edu.cn

Abstract. Due to the intuitiveness and low cost of image data, action recognition based on image data has always been the focus of action recognition research. The most of studies show that spatio-temporal two-stream network has very superior performance in the field of action recognition. However, traditional two-stream convolutional model needs multiple training, and the method of fusion two-stream models is too simple. In this research, a new two stream convolution fusion model is proposed, which can train our model end-to-end. At the same time, the basic model of spatio-temporal flow model is replaced by CBAM_Resnet. The model is improved on the basis of resnet_101, which gives weight to channel and spatial features, so as to improve the performance of the model. In this paper, the RGB image and the optical flow in X and Y directions are taken as the input features of the fusion network, and the final classification result of the model is obtained. Our method is validated on HMDB51 datasets, and the experimental results show that the accuracy of our model is improved by 4%.

1. Introduction

Human action recognition plays an important role in computer vision and it can be used in video monitoring, health detection, video content analysis and other fields. so action recognition based on deep-learning is still a research hotspot. Based on the summary of a large number of studies, some researchers [1], [2] discussed in detail the current challenges in the field of action recognition, including object occlusion, the differences caused by different individuals with the same action, the different execution speed of the same action, and variation in viewpoint. In view of the existing research, image data are mainly used as basic data in this field.

RGB, optical flow, infrared image and other image data can well describe the background, interactive objects, motion and other features in the images, so image-based action recognition has been the concern of a large number of researchers. According to different research methods, this paper divides the field of image-based action recognition into two categories:

- Action recognition based on convolution neural network(CNN) and recurrent neural network(RNN).
- Action recognition based on three-dimension convolution neural network(3D-CNN).

1.1. Action recognition based on CNN and RNN

Convolution neural network(CNN) and Recurrent neural network(RNN) are suitable for processing image, image sequence and other data, they are usually used to handle this type of data. Simonyan et al.[3] introduced the model of two-stream convolutional neural network(2s-CNN), which is one of the most critical researches in the field of action recognition. This research takes motion features into
account in image-based action recognition for the first time. The spatial convolution network and temporal convolution network are trained by RGB and optical flow respectively. For the action of a certain frame in the video, the model extracts RGB and optical flow respectively, and then inputs RGB into spatial convolution network, optical flow into temporal convolution network. Finally, the final classification result is obtained by average fusion or SVM fusion. Wang et al. [4] proposed TSN to further improve the performance of 2s-CNN. TSN is improved on the basis of traditional 2s-CNN, it divides the video averagely to solve the problem that the traditional 2s-CNN is difficult to deal with long video. The basic CNN model CNN-M2048 is replaced by BN-Inception and the input feature of module are replaced by RGB-difference and warped optical flow which are obtained by image enhancement. Gammulle et al. [5] combined CNN and LSTM, and they used the output of CNN as the input of LSTM. Extensive experiments on UCF11, UCFSports and jHMDB, demonstrate that the model is superior. And Dai et al. [6] consider that RNN can better deal with the features of temporal, which combines the long short-term memory(LSTM) with two stream network, and uses spatial attention and temporal attention to make the model pay attention to the effective features in the image. Inspired by the two stream network, Ardianto et al. [7] studies the influence of more branch features on the model.

1.2. Action recognition based on 3D-CNN

Although using multiple two-dimensional convolutions(2D-CNN) to extract temporal and spatial features respectively, and then fusing can meet the needs of action recognition, it is difficult to extract spatio-temporal correlation features by this method. 3D-CNN structure naturally satisfies the feature extraction of spatial-temporal, so it is suitable for application in the field of video classification. Yang et al. [8] proposed an asymmetric 3D-CNN structure based on 3D-CNN. They designed an asymmetric one-directional 3D convolutions to approximate the traditional 3D convolution, so as to reduce the cost of computation to some extent. To improve the efficiency of the model, they design a set of local 3D convolutional networks to incorporate multi-scale 3D convolution branches. In addition, Huang et al. [9] also improved the computational costs of 3DCNN. In this research, a 2D expansion and 3D convolution parallel operation are designed to convert the pre-trained 2D convolution parameters to 3D convolution, so as to avoid the pre-training of video data on 3D convolution, so as to reduce the training cost. The 3D2CNN introduced by Liu et al. [10] can extract features from the depth image sequence and bone data in the video. Finally, the two parts of features are fused through Support Vector Machines(SVM) to get the final score. Li et al. [11] input single frame GCN and optical flow images into different 2D-CNN, whereas 3D-CNNs are used to handle the clip and video stream. In each branch, LSTM are used to further model the temporal information in the frame, motion and clip streams. The action recognition result of the input video is obtained by linearly fusing the scores from all the branch of input stream. In order to improve the ability of long video processing, Yu et al. [12] combined 3DCNN with LSTM to get D3D-LSTM. At the same time, the model optimizes the original attention module by assigning different weights to each frame in real time.

Although two-stream method is generally accepted at present, it is too redundant to train spatial convolution network and temporal convolution network separately. In addition, it should be noted that for different action categories, the prediction of spatial convolution network and temporal convolution network have different effects on the final classification results. For example, the action of reading a book, because the background is almost static, the optical flow features extracted with RGB images are few, but scene features in the background often play a decisive role in the correct recognition of the action. Therefore, in this action, the output features of the spatial convolution network should be more important. However, the input feature of temporal convolution network will be more decisive for the more obvious action such as running. This is not taken into account by using average fusion or SVM fusion.

In this paper, we propose a new two-stream convolutional network, two-stream fusion net(2s-FusionNet). The network uses the end-to-end training method, and the final model can be obtained through one training. At the same time, the network can automatically fuse the output features of spatial
convolution and temporal convolution, and then train the fused features through deep neural network, which can assign weights to the spatial convolution and temporal convolution networks.

Our contributions can be summarized as follows:

- In order to improve the reliability of the model, channel attention and spatial attention are embedded in Resnet. In this way, the model can pay more attention to the features that are useful to the classification results.
- An end-to-end two-stream convolution network structure, 2s-FusionNet, is proposed in this paper. The model fuses the output features of spatial convolution network and temporal convolution network, and a better performance model can be obtained by one training.
- After verification, the recognition effect of our model on HMDB51 datasets is obviously better than that of similar models.

2. Materials and Methods

In this section, the framework of 2s-FusionNet will be introduced in detail. The following will describe the research content from CBAM-Resnet and 2s-FusionNet.

2.1. CBAM-Resnet

Channel attention and spatial attention(CBAM) [13] is a lightweight framework that combines channel attention and spatial attention for end-to-end training.

CBAM improves the traditional channel attention and spatial attention network, and integrates them. The specific structure is shown in Figure 1.

![Figure 1. This figure shows the structure of the CBAM module](image)

For the channel attention module, the input features are processed by the maximum pooling layer and the average pooling layer respectively, and the spatial dimension of the feature map is reduced to 1. Then the two feature maps are processed by the Multi-Layer Perceptron(MLP) respectively, and the final channel attention vector is obtained by adding the two feature maps and sigmoid processing. The specific formula is shown in equation (1).

\[
X' = \sigma(MLP(AvgPool(X)) + MLP(MaxPool(X)))
\]  

(1)

Where \(X\) is the input feature, \(X'\) is the output feature, \(AvgPool(\cdot)\) and \(MaxPool(\cdot)\) are the average pooling and maximum pooling operations respectively, \(\sigma(\cdot)\) is sigmoid activation function, and \(MLP(\cdot)\) is the MLP operation.

The input of spatial attention model in CBAM is the feature map with channel weight. In this module, the input features are pooled averagely and maximally on the channel dimension, and then the two pooled results are stacked on the channel dimension. Then a convolution layer with kernel size of 7 is used to reduce the channel dimension to 1. Finally, the feature graph is processed by sigmoid to get the final spatial attention weight. It can be shown by equation (2).

\[
X'' = \sigma(f^{7 \times 7}(AvgPool(X'); MaxPool(X')))
\]  

(2)

As a network structure widely used in the field of deep learning, Resnet has good performance. In order to further improve its performance, this paper applies CBAM structure to Resnet. For each bottleneck unit in Resnet, the CBAM module is embedded after it. The specific structure is shown in figure 2.
2.2. 2s-FusionNet

In order to fuse spatial convolution and temporal convolution neural network structure effectively, we proposed a new model of 2s-FusionNet, which is used CBAM-Resnet101 as the basic structure of spatial convolution network and temporal convolution network. Due to the Resnet101 structure is not...
suitable for the structure of the fusion network, this research will modify the Resnet101 model, delete the adaptive pooling layer and full connection layer in the end of model, and take its output features as the input feature of the fusion network. The 2s-FusionNet structure is shown in figure 3. It should be noted that RGB is directly used as the input feature in the spatial convolution of the model, so the number of channels for the input feature is 3. For the temporal convolution, the input feature is obtained by stacking the optical flow in X and Y on the channel dimension, so the input channel is 2. The output features of the spatial convolution module and the temporal convolution module are X and Y respectively, both of which can be expressed by $\mathbf{R}^{C \times H \times W}$, so the input feature dimension of the fusion model can be expressed by $\mathbf{R}^{2C \times H \times W}$. After the features are processed by FC layer in the 2s-FusionNet, the model can output the final classification vector.

3. Results & Discussion

This section will introduce the experimental results of the 2s-FusionNet on HMDB51. All the experiments are performed on a RTX3090 GPUs.

3.1. Implementation details

In our experiment, we use stochastic gradient descent (SGD) as optimizer, which parameter of Nesterov momentum is set to 0.9. We set the initial learning rate to 0.001. During the training, we set batch size and epoch to 32 and 30 respectively. For each RGB video in HMDB51, we take 70% of the data as the training data, 20% of the data as the verification data, and 10% of the data as the test data. Then, the RGB images of 16 to 19 frames are captured at the same time interval for each video, and the denseflow algorithm is used to extract the X and Y direction optical flow images from the RGB images of adjacent frames. Before all the image data are input into the model, they are first intercepted into 255×255 images and normalized accordingly.

3.2. Ablation studies

In order to verify the effectiveness of our method, we set baseline as 2s-cnn, and then verify the impact of CBAM-Resnet module and 2s-FusionNet on the model accuracy. The specific results are shown in table 1.

| Model                                  | Accuracy(%) |
|----------------------------------------|-------------|
| 2s-CNN(fusion by averaging)            | 58.0        |
| 2s-CBAM-Resnet(fusion by averaging)    | 61.2        |
| 2s-FusionNet                           | 62.0        |

As can be seen from table 1, after using CBAM-Resnet module, the accuracy of the model increase by 3.2%. When using the 2s-FusionNet, the effect of the model increases by 2.8%, which can prove that
our model is effective. The loss changes of 2s-CBAM-Resnet and 2s-FusionNet during training are shown in figure 4.

3.3. Comparision with state of the arts
In this paper, the traditional statistical method and deep learning method in the current mainstream action recognition research are compare with 2s-FusionNet, and the experimental results are shown in table 2.

| Method                                         | Accuracy(%) |
|------------------------------------------------|-------------|
| Improved dense trajectories(IDT)[14]           | 57.2        |
| IDT with higher-dimensional encodings[15]      | 61.1        |
| Spatio-temporal HMAX network[16]               | 22.8        |
| 2s-CNN(fusion by averaging)[3]                 | 58.0        |
| 2s-FusionNet                                   | **62.0**    |

4. Conclusions
In this research, a new end-to-end two-stream convolution network structure is proposed, which improves the traditional two-stream network on the model structure and spatio-temporal flow fusion method, and the model is verified on HMDB51. The experimental results show that the method of fusing the results of different branch networks and then extracting deeper features can improve the effect of the model.

Acknowledgments
The author is very grateful for the High Performance Computing Center of Shanghai University for providing the experimental environment.

References
[1] Poppe, R. (2010) A survey on vision-based human action recognition. Image and Vision Computing, 28: 976–990.
[2] Thakkar, A., Pareek, P. (2020) A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review, 54: 2259–2322.
[3] Simonyan, K., Zisserman, A. (2014) Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 1: 568–576.
[4] Wang, K., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L. (2016) Temporal segment networks: Towards good practices for deep action recognition. in Proceedings of the 2016 European Conference on Computer Vision(ECCV), 20–36.
[5] Gammulle, H., Denman, S., Sridharan, S., Fookes, C. (2017) Two stream LSTM: A deep fusion framework for human action recognition. in Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 177–186.
[6] Dai, C., Liu, X., Lai, J. (2020) Human action recognition using two-stream attention based lstm networks. Applied Soft Computing, 86: 1–8.
[7] Ardianto, S., Hang, H. (2018) Multi-view and multi-modal action recognition with learned fusion. in Proceedings of 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1601–1604.
[8] Yang, H., Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., Maybank, S. (2018) Asymmetric 3D convolutional neural networks for action recognition. Pattern Recognition, 8: 1-35.
[9] Huang, Y. Guo, Y., Gao, C. (2020) Efficient parallel inflated 3D convolution architecture for action recognition. IEEE Access, 8: 45753–45765.
[10] Liu, Z., Zhang, C., Tian, Y. (2016) 3D-based deep convolutional neural network for action recognition with depth sequences. Image & Vision Computing, 55: 93–100.
[11] Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y., Luo, J. (2016) Action recognition by learning deep multi-granular spatio-temporal video representation. in Proceedings of the 2016 ACM on
International Conference on Multimedia Retrieval, 159–166.
[12] Yu, J., Gao, H., Yang, W., Jiang, Y., Chin, W., Kubota, N., Ju, Z. (2020) A discriminative deep model with feature fusion and temporal attention for human action recognition. IEEE Access, 8: 43243–43255.
[13] Woo, S., Park, J., Lee, J. Y., et al. (2018) CBAM: Convolutional Block Attention Module. Springer, arXiv:1807.06521.
[14] Wang, H. Schmid. C. (2013) Action recognition with improved trajectories. in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 3551–3558.
[15] Peng, X., Wang, L., Wang, X., Qiao, Y. (2014) Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision & Image Understanding,.150: 109-125.
[16] Jhuang, H., Serre, T., Wolf, L., Poggio, T. (2007) A biologically inspired system for action recognition. in Proceedings of the IEEE International Conference on Computer Vision(ICCV), 1–8.