Video Classification Method of Multi-Way Convolutional Network Based on Deep Metric Learning

Xiaoxia Luo and Beibei Zhou*
Xi'an University of Science and Technology, College of Computer Science and Technology, Xian 710054, Shanxi, China
*Corresponding author. Email: 1215831011@qq.com

Abstract. Aiming at the significant impact of video semantic changes on video classification results, in the video classification process, which includes the large intra-class dispersion and inter-class similarity during video, this paper proposes a multi-way convolutional network video classification method based on deep metric learning. The method includes a 3D network-based multi-way convolutional network and a metric learning method based on the allocation of negative sample intervals. The network is mainly divided into three parts: segmented video feature extraction, similarity measurement based on deep metric learning, and classification. Firstly, the multi-channel convolutional network can extract the features of different periods of the video, and obtain the depth features of the video through feature fusion. Secondly, by calculating the error based on the interval function of the average semantic distance of negative samples and backpropagating, the network can learn the difference in semantic distance between samples. Finally, the network combines classification tasks with metric learning during the training process to make the network classification results better. Experiments on the data set UCF101, compared with existing methods, the multi-way convolutional network video classification method based on deep metric learning can effectively improve the accuracy of video classification.

Keywords: video classification, deep metric learning, multi-way convolutional network, interval function, multi-task learning.

1. Introduction

Traditional video classification methods are classified on the basis of manually designed features. Deep learning-based video classification methods [1-4] can automatically learn the most suitable features from video data quickly through backpropagation. The method is divided into two steps: joint optimization feature extraction and classification, to maximize the use of both Cooperate to optimize the performance of the entire system [5].

Convolutional Neural Network (CNN) [6] appeared earlier, and its convolutional characteristics have a unique role in image processing. Convolutional neural networks preserve the sift features and translation invariance of the image, which are better than Relying on artificial domain knowledge for feature extraction is more convenient and effective. Drawing on the idea of image classification, a video segment is considered as a collection sequence of many image frames, using existing network
models, such as AlexNet [7], VGGNet [8], GoogleNet [9], ResNet [10]. A pre-trained model trained on the image data set ImageNet to extract the depth features of each frame of image.

In recent years, deep convolutional networks have achieved great success in the field of computer vision. Therefore, in order to solve the above-mentioned problems about the nonlinearity and scalability of traditional metric learning methods, deep convolutional networks have been gradually introduced into metric learning. Use the depth feature to fit the non-linear mapping of the data. Deep metric learning optimizes semantic feature representations by using contrast loss functions [11-12] or triples loss functions [13], because the sub-video features at different time periods of the video have different effects on video classification results, this paper proposes a multi-way convolutional video classification network based on 3D networks, and in order to enable the network to learn intra-class dispersion and class Similarity, this paper introduces deep metric learning structure in the network. At the same time, the multi-task learning method is used in the training phase to perform metric learning and classification tasks in the network at the same time, avoiding that the network only focuses on metric learning and biases the classification network, thereby improving the result of video classification.

2. Algorithm description

2.1. Video classification method based on multi-channel convolutional network

Due to the different impacts of video content on video classification results, a multi-channel convolutional network is proposed for video classification. First, the video is segmented, and the segmented video is sent to a multi-channel convolutional network for feature extraction. Then the features of multiple video segments are fused to form the features of the video, thereby improving the classification effect. Finally, the structure of deep metric learning is added to the network. Multi-task learning is used to train the network during network training. Perform metric learning and classification tasks at the same time to avoid bias in the learning of the network and only focus on metric learning tasks. The network structure is shown in Figure 1.

![Figure 1. Multi-channel convolutional network model.](image)

Since a continuous video sequence is to be input into the network, the video sequence is normalized. Here, the nearest-neighbor interpolation method is used to process the image sequence, the purpose is to delete or repeat some video frames, and finally change the size of the graphics sequence to 16 frames. The video sequence is input to a 3DConvNet network for feature extraction. The network model is shown in Figure 2 below.
2.2. Metric learning based on negative sample-pair interval

The interval in the loss function proposed by [15] is constant, that is, the same interval is assigned to all negative samples. This makes negative samples with smaller semantic distances more likely to cause misjudgment. If you ignore the differences in semantic distance and assign a uniform constant interval to all negative samples, you obviously cannot make full use of this information. For this reason, this paper proposes an interval allocation method based on the average interval to make the network pay more attention to these negative samples, thereby improving the classification effect of the network model. Interval function:

\[
\beta = \frac{\sum_{(i,j) \in P, (i,j) \in N} D_{i,k} + D_{j,l}}{m}
\]

\[
\begin{cases}
\alpha_{i,k} = \exp \left( \frac{D_{i,k}}{\beta} \right) & D_{i,k} \leq \beta \\
\alpha_{j,l} = \exp \left( \frac{D_{j,l}}{\beta} \right) & D_{j,l} \leq \beta \\
\alpha_{i,k} = \max(D) & D_{i,k} > \beta \\
\alpha_{j,l} = \max(D) & D_{j,l} > \beta
\end{cases}
\]

When the distance between negative sample pairs is less than or equal to the average semantic distance between negative sample pairs, the interval function is the power of the ratio of the semantic distance to the average distance. The average semantic distance between sample pairs, then the interval function is \(\max(D)\).

Using the above interval allocation function, the final loss metric function is
According to formula (2), the network will allocate a larger interval for the negative samples that are closer. In this way, when calculating the loss, the calculated residual is large. During the back propagation process, the network parameters learned The greater the contribution, the more the network pays attention to these difficult samples.

Therefore, the gradient of the final measurement loss function $J_m$ is:

$$
\frac{\partial J_m}{\partial D_{i,j}} = \frac{1}{|P|} \sum_{i,j} \frac{\partial}{\partial D_{i,j}} \left[ J_{i,j} \right] 
$$

(3)

$$
\frac{\partial J_m}{\partial D_{i,k}} = \frac{1}{|P|} \sum_{i,j} \frac{\partial}{\partial D_{i,k}} \left[ J_{i,j} \right] 
$$

(4)

$$
\frac{\partial J_m}{\partial D_{j,l}} = \frac{1}{|P|} \sum_{i,j} \frac{\partial}{\partial D_{j,l}} \left[ J_{i,j} \right] 
$$

(5)

In the process of multi-way convolutional network training, the network adds the gradient of the metric loss and the classification loss, and then back-propagates to other parts of the network for training again.

### 3. Experimental results and analysis

(1) Effectiveness of multi-way convolutional networks

In order to improve the accuracy of classification, the multi-channel convolutional network uses the influence of the characteristics of different video periods on the classification. In this paper, experiments are performed on the data set UCF101 for the 3DConvNet network and the multi-channel convolutional network proposed in this paper. It can be seen from Table 1 that the 3DConvNet-based multi-channel convolutional network can effectively extract segmented video features and classify them. Compared with the original 3DConvNet network, the classification accuracy is improved by 0.204%.

The segmented features extracted by the multi-way convolutional network can effectively improve the classification accuracy.

| internet data set | Accuracy |
|-------------------|----------|
| 3DConvNet         | 0.693    |
| Multiway convolutional network | 0.897 |

Table 1. Accuracy of the network in the UCF101 dataset (%).

(2) The impact of the size of $\lambda_1$ and $\lambda_2$ on the classification results

The loss function of a multi-way convolutional network consists of two parts: classification loss and measurement loss. In order to explore the relative weight of classification loss and measurement loss on the results of network learning, experiments are performed by adjusting the magnitude of $\lambda_1$ and $\lambda_2$. Because the main task of multi-way convolutional networks is for classification, and the
classification loss between data samples is smaller than the measurement loss between samples during the classification task. Therefore, in order to make the network pay more attention to the classification task, $\lambda_1$ is set to 1 and $\lambda_2$ the value of is 0 to 1 in the parameter adjustment process. The experimental results are shown in the table. Between 0 and 0.3, the network does not converge when $\lambda_2$ is greater than 0.3. But in the experiment, even if the parameter of the measurement loss is one order of magnitude less than the parameter of the classification loss, the change of the parameters has a significant impact on the results of network classification.

Table 2. Impact of metric learning on deep learning networks (%).

| $\lambda_2$ | 0 | 0.1 | 0.2 | 0.3 | 0.5 |
|------------|---|-----|-----|-----|-----|
| mAP        | 91.3 | 91.5 | 91.7 | 不收敛 | 不收敛 |

(3) The effect of different interval allocation functions on classification results

In this paper, three interval allocation functions a, b, and c are used, and experiments are performed on the UCF101 data set. The results are shown in Table 3.

a. $\log\left(D_{i,j} / D_{i,k}\right)$  b. $D_{i,j} / D_{i,k}$  c. $\exp\left(D_{i,j} / D_{i,k}\right)$

Table 3. Impact of metric learning on deep learning networks (%).

| math | a | b | c |
|------|---|---|---|
| Map  | 89.4 | 90.9 | 91.7 |

From Table 3, it can be concluded that when the interval function is a, the classification accuracy of multi-way convolutional networks with metric learning is lower than that of multi-way convolutional networks without metric learning. It can be seen that improper setting of the interval function will reduce network classification the result of. The classification accuracy of functions b and c is higher than that of multi-way convolutional networks, and the classification accuracy increases continuously as the interval function increases.

This shows that the distribution method of the interval function has an important impact on video classification. A good interval function can improve the intra-class dispersion and the similarity between the classes, thereby making the classification result better.

(4) Comparison with existing classification methods

This paper verifies the improvement of video classification accuracy by multi-channel convolutional network experiments, and experiments on various parameters to make the network classification results better. The network was tested on the data set UCF101 and compared with mainstream methods. The results are shown in the table.

Table 4. Comparison with existing mainstream methods (%).

| Sign | Network | Accuracy |
|------|---------|----------|
| Hand-crafted | DT+MVSV$^{[16]}$ | 82.5 |
| | iDT$^{[17]}$ | 85.9 |
| | iDT+HSV$^{[18]}$ | 87.9 |
| LSTM | LRCN$^{[19]}$ | 82.9 |
| CNN | C3D(1 net)$^{[20]}$ | 82.3 |
| | C3D(3 net)$^{[20]}$ | 85.2 |
| | Two-Stream$^{[21]}$ | 88.0 |
| | Deep metric learning | 91.7 |

From Table 4, we can see that the multi-way convolutional network based on deep metric learning proposed in this paper can effectively classify videos. The mainstream video classification methods are tested under the same data set and experimental conditions. Improved classification accuracy.
4. Conclusion

A multi-way convolutional network video classification method based on deep metric learning can extract sub-video features at different periods of the video, and then fuse the sub-video features to obtain the features of the entire video. The characteristics of each video segment can affect the classification result, thereby improving the accuracy of classification. This paper proposes a deep metric learning method based on the interval distribution function interval of the average interval of negative sample pairs. Based on the semantic distance between the negative sample pairs, the interval calculation function is used to calculate the interval, so that the classification network can pay more attention to difficult samples during the training phase. To improve classification accuracy. A large number of experiments have been performed on the data set UCF101 to verify the effectiveness of the multi-channel convolutional network, especially the effectiveness of joining a deep metric learning network. This method effectively aggregates similar samples and classifies different samples. The experimental results are compared with the existing mainstream methods. Multi-way convolutional networks with deep metric learning have better classification accuracy. However, it is found in the experiments that the deep metric learning method proposed in this paper is more sensitive to network hyperparameters, causing unstable network convergence. The next step is to enhance the robustness and versatility of the method.

References

[1] Branson S, Van Horn G, Perona P, et al. Improved bird species recognition using pose normalized deep convolutional nets[C]. Proceedings of the British Machine Vision Conference, Nottingham, British, 2014: 197–211. doi:10.5244/C.28.87.

[2] Zhang Ning, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014, 8689: 834–849. doi: 10.1007/978-3-319-10590-1_54.

[3] Krause J, Jin Hailin, Yang Jianchao, et al. Fine-grained recognition without part annotations [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 5546–5555. doi: 10.1109/CVPR.2015.7299194.

[4] Lin Tsungyu, Roychowdhury A, and Maji S. Bilinear CNN models for fine-grained visual recognition[C]. Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1449–1457.

[5] Cui Yin, Zhou Feng, Lin Yuanqing, et al. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1153–1162.

[6] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]//Computer Vision and Pattern Recognition, 2005.CVOR 2005.IEEE Computer Society Conference on. IEEE, 2005, 1:539-546.

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems, 2012:1097-1105.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv: 1409.1556, 2014.

[9] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. CVPR, 2015.

[10] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; 770-778.

[11] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]// Computer vision and pattern recognition. 2006 IEEE computer society conference on. IEEE, 2006, 2:1735-1742.

[12] Chechik G, Sharma V, Shalit U, et al. Large scale online learning of image similarity through ranking[J]. Journal of Machine Learning Research, 2010, 11(Mar):1109-1135.
[13] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor
classification[J]. Journal of Machine Learning Research, 2009, 10(Feb):207-244.
[14] Tomar S. Converting video formats with FFmpeg[J]. Linux Journal, 2006, 2006(146):10.
[15] Song Hyunoh, Xiang Yu, et al. Deep metric learning via lifted structured feature
embedding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern recognition.
Las Vegas, USA, 2016:4004-4012. DOI:10.1109/cvpr:2016.434.
[16] Cai Z, Wang L, Peng X, et al. Multi-view super vector for action recognition[C]. Proceedings of the
IEEE Conference on Computer Vision and Pattern recognition. 596-603.
[17] Peng X, Wang L, Wang X, et al. Bag of visual words and methods for action recognition:
Comprehensive study and good practice[J]. Computer Vision and Image Understanding, 2016,
150:109-125.
[18] Wang H, Schmid C. Action recognition with improved trajectories[C]. Computer Vision (ICCV),
2013 IEEE International Conference on. IEEE, 2013:3551-3558.
[19] Donahue J, Anne Hendricks L, et al. Long-term recurrent convolutional networks for visual
recognition and description[C]. Proceedings of the IEEE conference on computer vision and
pattern recognition. 2015:2625-2634.
[20] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional
networks[C]. Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE,
2015:4489-4497.
[21] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in
videos[C], Advances in neural information processing systems. 2014:568-576.