Journal of Computers Vol. 29 No. 5, 2018, pp. 52-65
doi:10.3966/199115992018102905005

Algorithm of Video Semantic Classification Based on IAGA and Deep Convolution Neural Network

Min Wang[1], Ke-Xin Liu[1*], Ming-Fei Wei[1], Li-Cai Zhang[1]

[1] Department of Communication and Information Engineering, Xi'an University of Architecture and Technology,
710055, Shaanxi Province, China
{m18602907837, wangmin1329}@163.com

Received 1 June 2017; Revised 29 September 2017; Accepted 30 October 2017

Abstract. The rise of convolutional neural network (CNN) has greatly improved the disadvantage of the traditional video classification method. However, in the pre-training process of classification network, often due to over-fitting, gradient disappearance and other factors lead to training data convergence performance is poor, thus would affect the accuracy of the classifier. Aiming at the problem of network optimization, this paper proposes an algorithm to combine improved adaptive genetic algorithm (IAGA) with deep convolution neural network (DCNN) classifier. The weighting of the network is initialized by the IAGA algorithm, and the weight is corrected by combining the gradient descent (GD) algorithm. Finally, the fusion of global feature extracted by the network is input into the extreme learning machine (ELM) for classification. The results of the news video classification show that the algorithm can combine the global search ability of IAGA with the local optimization ability of gradient descent algorithm to improve the accuracy of the training network with less parameters, and the average classification accuracy rate can reach 90.03%. Compared with the three existing algorithms, the algorithm has higher classification accuracy. Compared with the four kinds of network pre-training methods, the algorithm presented in this article is more dominant.

Keywords: deep convolutional neural network, extreme learning machine, gradient descent algorithm, improved adaptive genetic algorithm, video semantic classification

1 Introduction

News video, as a composite sequence of multi-frame images, not only has a strong structure, but also has abundant semantic information. The events and scenes described in the video belong to high-level semantics, which is a high-level concept mapping of low-level features [1]. Compared with the other low-level features such as texture, shape and color, the semantic features, being able to cross the semantic gap, are closer to human thought [2]. Therefore, the extraction of the desired semantic information in the video and the completion of the precise classification will significantly facilitate the efficiency of researching and retrieval work. However, due to the inter-frame similarity among massive video frames and the complexity of semantic data, it is very difficult to obtain the video of the required category in the retrieval process. How to effectively and accurately classify the resources is the current research focus.

In recent years, plenty of classification methods for Video Semantic have been developed. Different from the method of classifying after the global feature extraction [3], the Bag-of-Visual-Words (BoVW) [4] was successfully applied by the researchers to solve the problem of video semantic classification with its high robustness. The principle is based on the extracted key frame features to construct the visual dictionary, using the clustered feature word for statistical judgment. Gaussian Mixture Model (GMM) [5], as an extension of the BoVW model, further enhances the classification efficiency by constructing super-vectors [6] to more accurately express key frame features. But these methods are still inseparable from

[*] Corresponding Author

the manual design parameters, which consume a lot of manpower.

Convolutional neural network (CNN) [7], with the advantages of weight sharing and small storage space [8], is widely used in human behavior identification [9], image semantic segmentation [10], ImageNet classification [11] and other image issues. Inspired by image classification, the CNN model and single classifier are used to classify large-scale image semantics [12] and therefore some good results are achieved. But the simple classification effect and the lack of network depth lead to over-fitting phenomenon, resulting in reduced accuracy. In 2016, the scholar proposed a pre-training model based on the structured CNN network of video frames [13], which effectively improved the accuracy of semantic classification. However, the problem of poor convergence performance still exists. The urgent matter is how to optimize the network training process.

Aiming at the above-mentioned shortcomings of the video semantic classification method, this paper presents a new algorithm to optimize the network and achieve high classification. Based on CNN, we have adopted deep convolution neural network structure, and initialized the network weights by taking advantage of the Improved Adaptive Genetic Algorithm (IAGA) and Gradient Descent (GD) method. In the meantime, Extreme Learning Machine (ELM) as network classifiers is selected, and appropriate parameters are applied to enhance the regularization of the learning process.

The rest of the paper is organized as follows: Sections 2 and 3 show the related work. Section 2 discusses the classification structure of deep convolution neural network and the training mode of network; Section 3 introduces the improved adaptive genetic algorithm. In Section 4, we elaborate on our DCNN-ELM network framework, and state the complete algorithm procedures. In Section 5, we carry out the video classification simulation work. Firstly, the stability and high precision of the model presented in this paper are verified through a large number of iterative experiments. Then the best batchsize value for enhance accuracy is selected; Secondly, three groups of comparative experiments were implemented: (1) Comparisons are made among the three kinds of classification algorithm of extracting image features manually, highlighting the high efficiency of this proposed algorithm; (2) Comparisons are made between the classification method of regard single-layer network as a classifier, highlighting the high robustness of this proposed algorithm; (3) Comparisons are made among the four kinds of classification methods of network pre-training, highlighting the classification advantage of this proposed algorithm for complex scenes semantics. Section 6 concludes the work, and proposes prospects for further research.

2 Deep Convolution Neural Network Classification Structure

Convolution neural network, as an extension of deep learning in the artificial neural network, is the model architecture with supervision, especially suitable for two-dimensional array problems [14].The weight sharing reduces the complexity of the research network, and the data processing of sub-sampling layers saves storage space and removes non-dominant features. After the input image is processed by the network, the spatial map is transformed into the corresponding feature expression.

2.1 Semantic Feature Extraction

In this paper, we will improve the CNN basic hidden layer structure [7] to obtain the depth network composed of input layer, three-layer convolution layers, three-layer sub-sampling layers (pooling layers), fully-connection layer and classification output layer structure, as shown in Fig. 1.

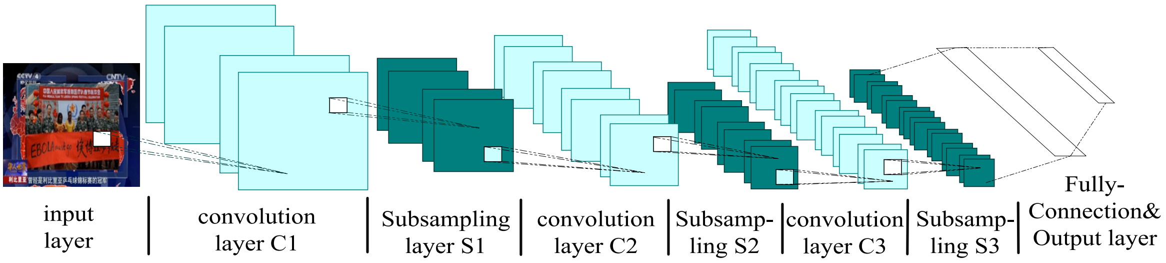


Fig. 1. Basic structure of deep convolution neural network

When a picture is taken as an input, it is necessary to select the appropriate local sensory field [15] to convolution the sample image, as follows [13]:

$$x_j^k = f\Big(\sum_{i\in M_j} W_{ij}^k * x_i^{k-1} + b_j^k\Big). \tag{1}$$

x_j^k and x_i^{k-1} represent the j-th feature graph of the K-layer and the i-th feature of the K-1-th layer, W_{ij}^k is the weight between the two layers, f is the non-linear excitation function, and b_j^k is the bias term.

The convolution operation is very good at extracting the local feature of the image due to its translation invariance, and obtaining the corresponding feature mapping group through the excitation function. Then the sub-sampling layer is used to reduce the dimensionality for the feature, the specific form is as follows [13]:

$$x_{pool\,j}^{\;k} = f\Big(\beta_j^k \sum_{i\in M_j} x_i^{k-1} + b_{pool\,j}^{\;k}\Big). \tag{2}$$

β_j^k for the training parameters, the remaining parameters are consistent with the convolution layer. To calculate the clustering statistics of the extracted feature, the local maximum (average) method are usually selected for the combination of features, whereas the number of feature graphs of the convolution operation would not be affected, as shown in Fig. 2.

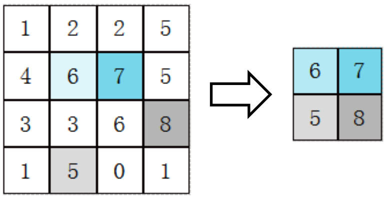


Fig. 2. The principle of max-pooling

2.2 Classifier Design

After the multi-layer convolution and pooling operation of the architecture in Fig. 1, a global feature based on image semantics could be generated from typical characteristics filtering through the full connection operation. Finally, we can get the corresponding classification results by selecting a reasonable classifier in the output layer. The classifier consists of a single-layer neural network, the multi-classification of which is often achieved through the softmax regression model [16] based on the probability classification model. Fig. 3 is a visual demonstration of the classification of picture cars using the DCNN in the CIFAR-10 dataset [17].

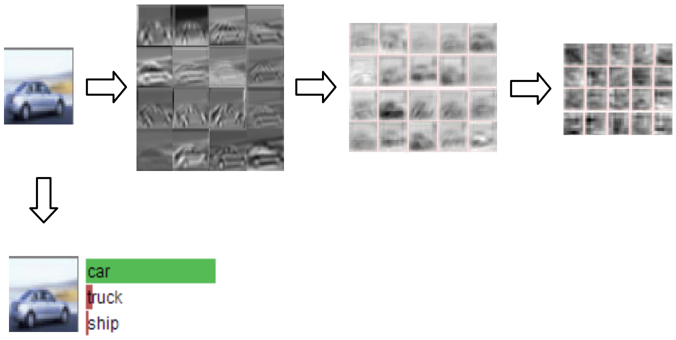


Fig. 3. Visualization of image classification based on DCNN network

After the hidden layers are increased to 7, more detailed features of the image can be extracted from each convolution layer. The classification accuracy of Fig. 3 is about 87.6%, which means the probability of misjudgment of a truck or boat is smaller. In order to save the network execution time and reduce the over-fitting phenomenon, this paper selects the extreme learning machine for classification, and for further improvement of the classification accuracy with its combination of the CNN network [18].

The advantage of the ELM method is that it does not need to be iterated after selecting the optimal parameters, and the training can be carried out efficiently merely by learning the weights between the hidden layer and the output, which is easier to get close to the global optimum structure [19]. Combined with the CNN extraction ability and the generalization ability of ELM, the classification accuracy of Fig. 3 is improved to 89.2%.

The CNN-ELM model is used to analyze the image semantics, which is more specific than traditional artificial analysis, more accurate than the subtitle extraction technique, more efficient than the underlying feature analysis algorithm of scene detection.

2.3 Training Convolution Neural Network

The loss function determines the performance of the CNN network model. Usually, the Mean Squared Error (MSE) function is used as the loss function [14], t is the classification sample label, y is the excitation function output label, l,c are the dimension and the number of categories, for each training sample error can be expressed as [20]:

$$E(W,b)=1/2\sum_{c}(t_l^n-y_l^n)^2=1/2\left\|t^n-y^n\right\|_2^2. \tag{3}$$

For the sake of minimize the loss function while preventing the weight over-fitting, on the basis of Back Propagation (BP) neural network, the gradient descent method [14] is often used to adjust the network parameters W and b. With the BP network's excellent performance in terms of video data evaluation [21], it becomes more necessary to calculate the partial derivative of each weight for the loss function according to the BP rule[20] for CNN back propagation network, the deformation and expansion [22] of BP. By combining the gradient calculation residual for different batches, the hidden layer weight gradient correction can be finally achieved.

For the back propagation of the convolution layer, according to (1), the residual of each neuron can be expressed as [20]:

$$\delta_j^k=\beta_j^{k+1}(f'(x_j^k)\circ up(\delta_j^{k+1})). \tag{4}$$

$up(x)=x\otimes 1_{n\times n}$ represents an up-sampling operation, which uses the Kronecker product to implement, “$\circ$”means multiplied by element.

p_i^{k-1} is a template of x_i^{k-1} is multiplied by each weight element during the convolution operation; η for the learning rate, then through the chain rule to calculate the parameters can be updated to:

$$\frac{\partial E}{\partial W_{ij}^k}=\sum_{u,v}(\delta_j^k)_{u,v}(p_i^{k-1})_{u,v},\frac{\partial E}{\partial b_j}=\sum_{u,v}(\delta_j^k)_{u,v}. \tag{5}$$

$$W_j=W_j-\eta\frac{\partial E}{\partial W_{ij}^k},b_j=b_j-\eta\frac{\partial E}{\partial b_j}. \tag{6}$$

$\Diamond$ represents the matrix convolution operation. According to (2), the residuals of the sub-sampling layer can be calculated as:

$$\delta_i^k=(\sum_{j=1}^{M}\delta_j^{k+1}\Diamond(W_{ij}^{k+1}))\circ f'(x_j^k). \tag{7}$$

d_j^k is the sub-sampling operation of the previous layer x_i^{k-1}; η for the learning rate, the parameters can be updated to:

$$\frac{\partial E}{\partial \beta_j^k}=\sum_{u,v}(\delta_i^k\circ d_j^k)_{u,v},\frac{\partial E}{\partial b_{pool\,j}}=\sum_{u,v}(\delta_i^k)_{u,v}. \tag{8}$$

$$\beta_j = \beta_j - \eta \frac{\partial E}{\partial \beta_j^k}, b_{pool\,j} = b_{pool\,j} - \eta \frac{\partial E}{\partial b_{pool\,j}}. \tag{9}$$

3 Improved Adaptive Genetic Algorithm

The random weight initialization cannot satisfy the high precision requirement of the convolution neural network classification, but the convergence of the genetic algorithm provides the excellent parameters of the complex network model. Therefore, this paper will use the improved adaptive genetic algorithm to initialize the DCNN network weights.

Genetic algorithm (GA), with its global search ability, is widely used in the field of neural network model optimization. Through the three steps of selecting, crossing and mutating, the best fit individual for the study could be selected. The choice of crossover probability and mutation probability in the process of executing the cycle directly affects the convergence of the algorithm and determines the performance of the genetic algorithm.

In order to solve the problem of premature convergence and local stagnation, the literature [23] improves the adaptive genetic algorithm, proposes IAGA and confirms that the new algorithm improves the search efficiency. The optimized probability of crossover and mutation is as follows:

$$P_c = \begin{cases} \dfrac{P_{c1}}{(P_{c1} - P_{c2}) + e^{\frac{f'-f_{avg}}{f_{\max}-f_{avg}}}}, & f' \geq f_{avg} \\ k_1 P_{c1}, & f' < f_{avg}. \end{cases} \tag{10}$$

$$P_m = \begin{cases} \dfrac{P_{m1}}{(P_{m1} - P_{m2}) + e^{\frac{f-f_{avg}}{f_{\max}-f_{avg}}}}, & f \geq f_{avg} \\ k_2 P_{m1}, & f < f_{avg}. \end{cases} \tag{11}$$

$f_{\max}$ and f_{avg} are the maximum fitness and average fitness values in the population, respectively; $f^{'}$ is the greater fitness value of the individual when preparing the cross operation; f is the larger fitness value of the individual in the preparation mutation; k_1, k_2 are constants that takes the range between $[0.5,1]$.

4 Video Semantic Classification Based on IAGA-DCNN

In this paper, a new algorithm based on the combination of adaptive genetic algorithm and convolution neural network is proposed. The gradient descent method is used to calculate the training residual of DCNN network and correct the parameters during the execution cycle of IAGA choosing the best weight. After the best network is set up, the extracted features from training set and test set are fused and connected to the extreme learning machine so that the accurate classification of news video semantics can be realized.

4.1 Video Semantic Classification DCNN Model Structure and Parameter Design

Based on the structure of Fig. 1 a total of 7 layers（input and output layer not included）, each layer of the processing structure shown in Fig. 4. After the balance treatment, a plurality of key frames of each training video are converted into 64×64 size image sequences as input, and Rectified linear units (RELU) are selected as the nonlinear excitation function of neurons.

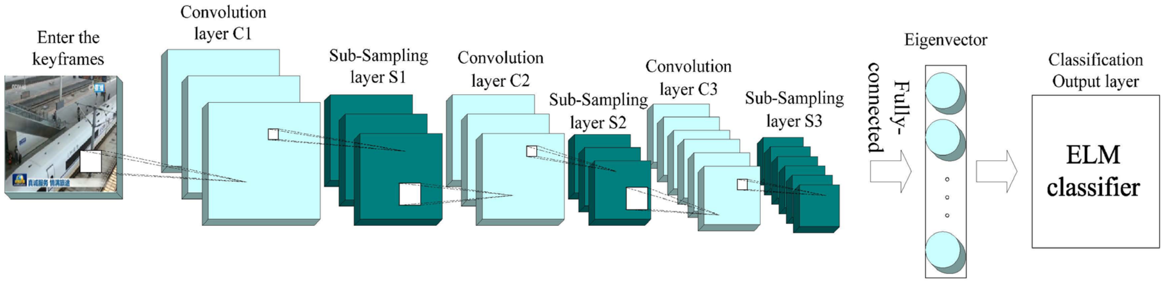


Fig. 4. DCNN-ELM model of video semantic classification

The first layer C_1 defaults to the initial convolution kernel size of $7\times7\times6$, and 6 characteristic maps with size of 58×58 could be obtained after convolution according to (1);

The second layer S_1 defaults the sub-sampling template to 2×2, the maximum pooling operation in accordance with (2) could be clustered and compressed into $1/4$ of the previous image size and obtained six characteristic maps with size of 29×29.

The third layer C_2 selected $7\times7\times12$ size convolution kernel, and 12 feature maps with size of 23×23 could be gained;

The fourth layer S_2 sub-sampling layer template is 2×2, 12 feature maps with size of 11×11 could be received;

The fifth layer C_3 preset convolution kernel size is $7\times7\times24$, 24 feature maps with size of 5×5 could be obtained after the operation that's similar to the convolution operation;

The sixth layer S_3 default sub-sampling template is 2×2, and finally 24 feature map with compression size of 2×2 could be gained;

The seventh layer is a fully connect layer containing 300 neurons, the output layer gives the results of the judgment using a single-layer network extreme learning machine with five neurons as a classifier.

4.2 IAGA-DCNN Optimization Algorithm

According to the flow chart, a video semantic classification network is constructed as follows:

(1) Preprocessing the video data and classify them into ~~a~~ training set and test set in a cross-validation manner. To avoid premature convergence, the order of all video sets is disrupted.

(2) As shown in Fig. 4, establish the IAGA-DCNN model based on N training video segments. Assuming that each video has m frame images. Given the local patterns' and inter-frames' similarity, we calculate the difference between each frame and set the two thresholds $a,b\,\&\,a>b$. a is set for judging the abrupt change point, and b is for judging the gradual change point of the lens, dividing the video into R parts according to the lens rate of change, and setting the intermediate frame as a key frame input execution network. Then the m/R classifier weights of the convolution network and the seed value of the fully connect layer classifier are encoded as the total population, and the current best fitness selected by the IAGA algorithm are used to initialize the weights of network.

(3) Train the network with the decoded parameters as shown in Fig. 4. All the key frame features extracted by m/R video are merged and connected to generate global features. According to the label error of classification samples, the gradient descent method is used to optimize the parameters W,b of the training set based on the BP algorithm.

(4) Perform the operation in step 2 for the local optimal solution again after training, seek the current best fitness and decode it to execute the DCNN network again.

(5) Repeat the IAGA-DCNN cycle framework to optimize the fitness until the global optimal solution is found. And the chromosomes with the best fitness after evaluating and decoding will be the weight of each classifier for the final test network.

(6) Similar to the classification of the training set, the key frames for each segment of the video data in the test set are extracted as input (like step 3). The global feature matrix formed by mapping fusion is input into the ELM classifier to classify and judge according to the probability.

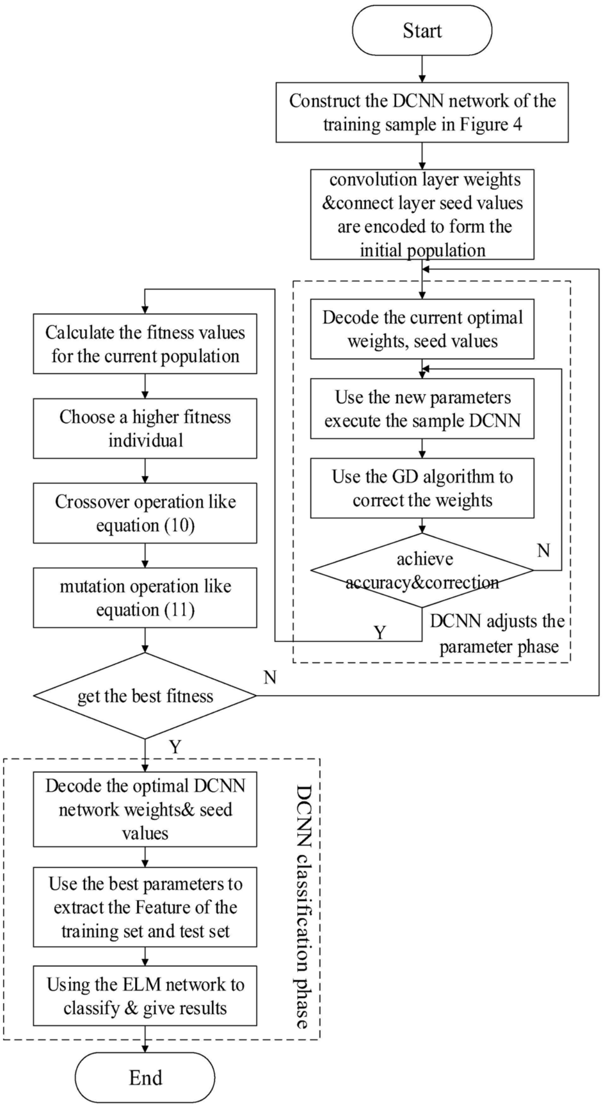

Start
Construct the DCNN network of the training sample in Figure 4
convolution layer weights &connect layer seed values are encoded to form the initial population
Calculate the fitness values for the current population
Choose a higher fitness individual
Crossover operation like equation (10)
mutation operation like equation (11)
get the best fitness
N
Y
Decode the current optimal weights, seed values
Use the new parameters execute the sample DCNN
Use the GD algorithm to correct the weights
achieve accuracy&correction
N
Y
DCNN adjusts the parameter phase
Decode the optimal DCNN network weights& seed values
Use the best parameters to extract the Feature of the training set and test set
Using the ELM network to classify & give results
DCNN classification phase
End

Fig. 5. IAGA-DCNN video semantic classification algorithm flow chart

From the above six aspects of operation, we can accurately and efficiently classify the video semantics by the following Algorithm.

Algorithm. Video semantic classification based on IAGA - DCNN

Step 1. The video is divided into training set D_{Train} and test set D_{Test} by cross validation;

Step 2. D_{Train} training group contains N groups of videos, each containing m/R key frames. m/R is also regarded as the number of classifiers, encoding the total number of population by combining with the classifier seed number, and evaluating fitness of IAGA chromosomes;

Step 3. Establish a network for each group of key frames as shown in Fig. 4. Initialize the network using the fitness generated by IAGA and calculate the propagation error $E(W,b)$. Update the convolution layer parameters according to Eq. (5) (6); Update the pooling layer parameters according to Eq. (8) (9);

Step 4. Repeat Step 2, calculate the current optimal fitness according to Eq. (10) (11) and execute IAGA-DCNN framework for multiple cycles;

Step 5. Harvest the optimal IAGA population, and initialize the m/R classifier weights for $1/N$ group using the final parameters;

Step 6. The key frames of all the videos in D_{Train} and D_{Test} are put into the trained network and the classification results are acquired.

5 News Video Semantic Analysis Results

Experimental environment. In order to verify the feasibility of the proposed algorithm, we choose the news video with strong structure as the research object, the experimental environment is Intel core i7, NVIDIA Geforce 840M, operating system for 64-bit Windows 8.1, CNN network dependent configuration is based on Anaconda Python 2.7.10, CUDA 6.5, Theano 0.7, VS 2013 ultimate, Matlab 2014a and deep Learning Toolbox, genetic algorithm toolbox developed by Sheffield University.

Object of study. Firstly, we selected 988 news video from the CCTV official website. Due to CNN semantic identification of the integrity, the news network video will be divided into leaders meet (leader), traffic, meeting, fire, water conservancy construction (water) five categories. In the fitness assessment, we use the 5-fold cross validation method to group the data sets into groups $E_1, E_2, ..., E_5$, and then $E_q, q = 1,2,...,5$ as the test set, and the other group $q-1$ is used as the training set to verify, and some training set keyframes are shown in Fig. 6.

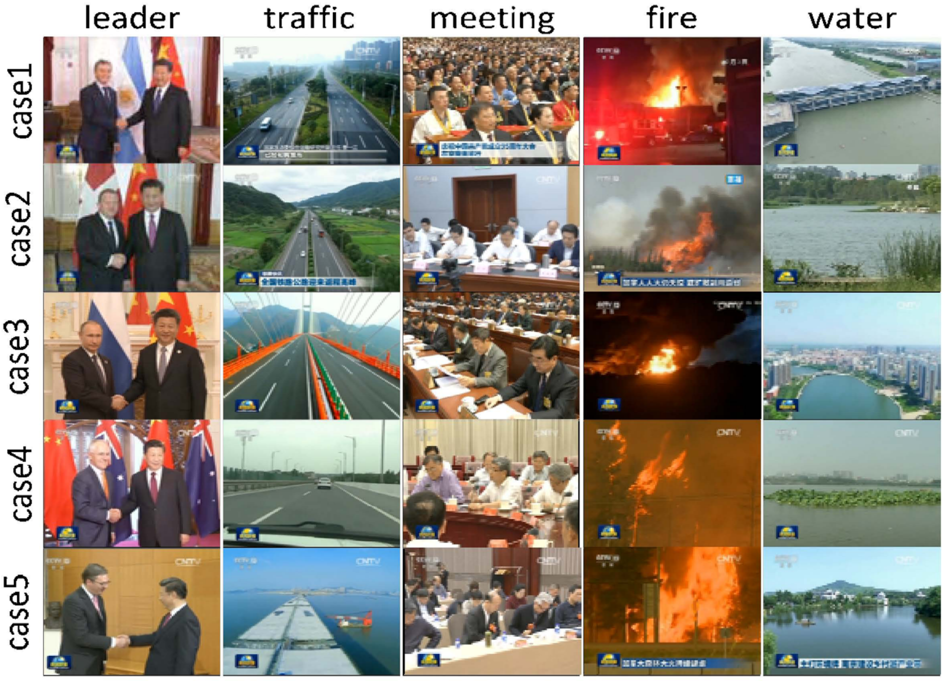


Fig. 6. Keyframes of five categories of news semantics

Secondly, we selected the National Institute of Standards and Technology Trecvid 2012 data set as a research object to test the classification performance of the optimized network. The study subjects were divided into 10 categories: Airplane Flying (AF), Baby (Ba), Building (Bu), Car, Dog, Flower (Fl), Instrumental Musician (IM), Mountain (Mo), Scene Text (ST), Speech (Sp), and each category contains 20 videos.

Parameter selection. According to the parameters of [18], the weight range of the convolution layer mask is $[-100,100]$, the activation value range of the fully-connect layer is *[0,5000]*. From the number of network classifier in Fig. 4, population size selected 40; In order to reduce the disturbance probability of the mutation probability, the crossover parameters are: $P_{c1} = 0.9$, $P_{c2} = 0.6$, $K_1 = 1$; The parameters of the mutation are: $P_{m1} = 0.1$, $P_{m2} = 0.05$, $K_2 = 0.5$; CNN network learning rate η is selected 0.01, the execution period epoch is 50 times. And the effect of batchsize selection on performance is recorded in Table 1.

Table 1. The effect of batchsize selection on the optimal fitness of the network (classification accuracy) (%)

Batch-size	k1 (1st group)		k (2nd group)		k3 (3rd group)		k4 (4th group)		k5 (5th group)	
	Gene	Rate	Gene	Rate	Gene	Rate	Gene	Rate	Gene	Rate
6	1	47.78	1	50.5	1	52.6	1	43.1	1	55.7
	2	67.65	2	69.63	2	68.4	2	69.3	2	70.6
	3	79.1	3	86.4	3	70.7	3	79.2	3	90.09
	4	87.4	4	87.99	4	90.1	4	89.7	4	90.1
	5	88.7	5	88.91	5	90.1	5	89.9	5	90.1
8	1	59.72	1	57.8	1	52.1	1	42.6	1	47.03
	2	68.93	2	72.3	2	67.63	2	57.8	2	68.6
	3	86.6	3	79.19	3	89.4	3	84.7	3	89.44
	4	89.99	4	88.9	4	89.69	4	90.2	4	90.02
	5	90.1	5	89.2	5	90.1	5	90.3	5	90.02
10	1	56.9	1	58.1	1	57.7	1	48.2	1	55.5
	2	74.05	2	78.11	2	80.2	2	81.6	2	88.3
	3	89.4	3	89.7	3	89.4	3	89.2	3	91.2
	4	90.2	4	89.8	4	90.3	4	89.4	4	91.2
	5	90.2	5	89.9	5	90.32	5	90.1	5	91.2
20	1	49.1	1	52.7	1	37.6	1	52.4	1	54.61
	2	64.63	2	60.3	2	50.4	2	67.63	2	77.9
	3	87.2	3	88.1	3	89.9	3	89.7	3	88.4
	4	87.6	4	89.3	4	89.9	4	90.02	4	89.6
	5	87.6	5	89.3	5	89.9	5	90.02	5	89.6
50	1	27.6	1	30.2	1	12.7	1	29.8	1	30.7
	2	38.8	2	41.2	2	43.6	2	36.7	2	48.3
	3	69.2	3	66.1	3	59.7	3	68.9	3	69.2
	4	70.72	4	70.2	4	69.11	4	69.8	4	69.4
	5	71.5	5	70.9	5	70.3	5	69.8	5	69.7

Gene represents the number of genetic cycles and found that the accuracy of the 5-generation cycle tends to be stable. Rate represents the average accuracy rate of each group of videos after classification by the network of Fig. 4. The number of videos per group is 198,202,197,200,191, and the weights of each group are initialized by the IAGA algorithm. The experimental results show that the higher accuracy can be obtained under the condition of fewer parameters. With the increase of batchsize, the running time of the network is decreasing, but the classification accuracy is gradually declines. When batchsize is 10, the accuracy rate tends to be optimal during the genetic cycle.

For the classification of five categories of video, this paper assesses the classification performance in the form of confusion matrix as shown in Fig. 7. We can find that the accuracy of single scene recognition is higher. The recognition rate of the news clips concerning traffic, fire and water conservancy construction are all higher than 90%. Whereas the recognition rate for meet and meetings which contains more people and confusing background is slightly lower.

Comparative Experiment. Firstly, the algorithm of this paper is compared with the classification results of The SIFT feature combines SVM with the method (SIFT-SVM), the BoVW, and the Method for SIFT feature to applying Gaussian mixture model (SIFT-GMM) for the news video semantics respectively. SIFT dimension is selected 128, SVM selects the RBF kernel function, each video segment is specified as a GMM super vector, and the result is recorded in Fig. 8.

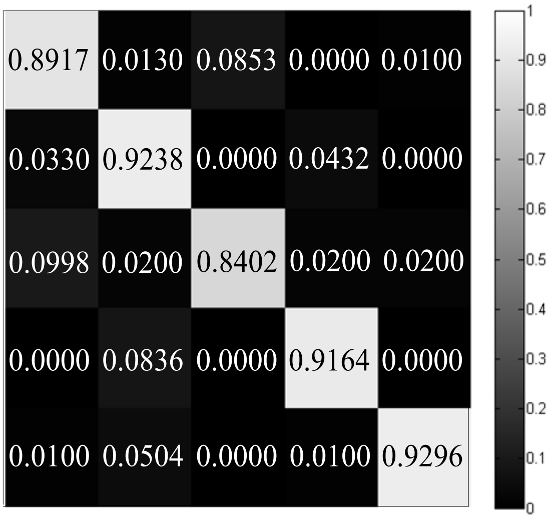
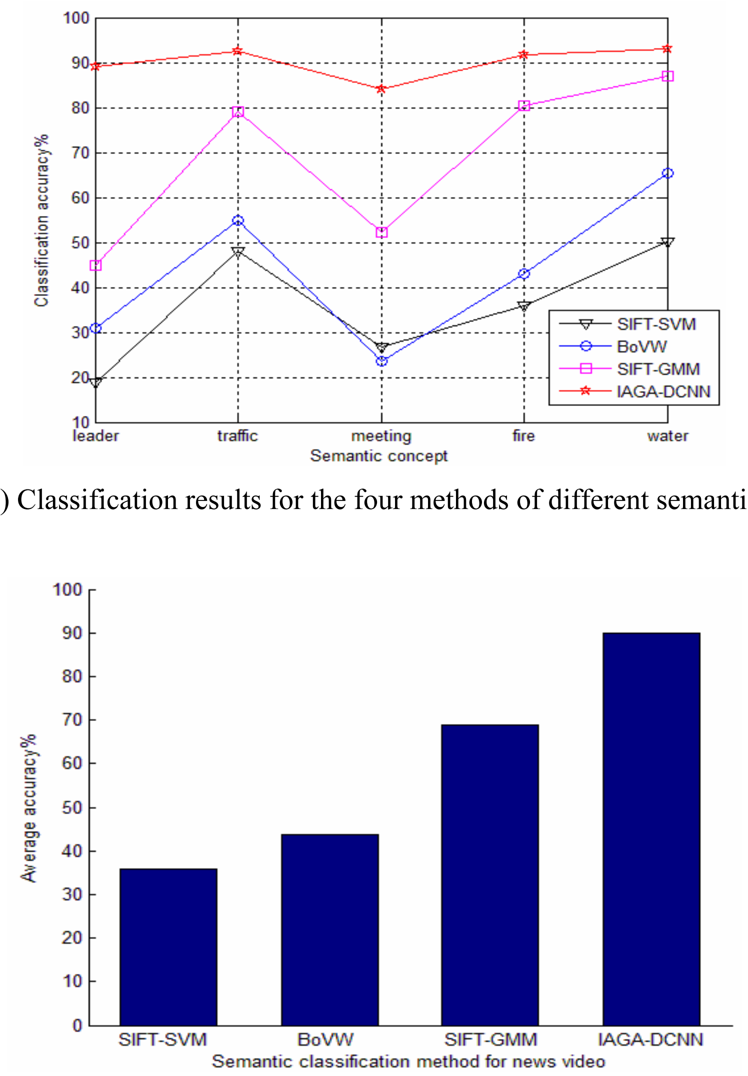


Fig. 7. News video semantic classification performance confusion matrix



(a) Classification results for the four methods of different semantics

(b) The average accuracy of the four different semantic classification methods

Fig. 8.

It can be concluded from Fig. 8(a) that the accuracy rate of DCNN-based classification method is significantly higher than that based on the image features manually extracting method.

For the single scene semantics of traffic, fire and water, better performance can be achieved by using IAGA-DCNN compared to others. But for the meeting semantics, the classification is put at a disadvantage with less than 50% accuracy. What's worse, the classification accuracy of SIFT-SVM and BoVW can't even reach 27%. All this can be attributed to the semantics of the meeting combining the background and the characters, which added a certain degree of difficulty in identification. However, the accuracy of the IAGA-DCNN for complex scenes is still 84.02%, much higher than the rest of classification model. This could be due to the convergence and efficiency of the model proposed in this paper.

It can be observed from Fig. 8(b) that the classification accuracy of IAGA-DCNN is 21.31% higher than the second ranked SIFT-GMM model, 46.46% higher than that of BoVW, 54.13% higher than that of SIFT-SVM, which fully reflects the advantages and high robustness of CNN to semantic classification.

Secondly, we explored the impact of classifier selection on classification performance and conducted the following experiment: Based on the feature extraction structure of IAGA-DCNN, the classifiers of this paper adopted neural network and Extreme Learning Machine respectively, and the performance tests for five sets of data sets are recorded in Table 2.

Table 2. Performance of classifiers using neural network and extreme learning machine for the 5-folds of news dataset (%)

Group	Number of observations	Classifier selection	
		DCNN-NN	DCNN-ELM
K1	198	87.6	90.2
K2	202	86.5	89.9
K3	197	88.11	90.32
K4	200	87.2	90.1
K5	191	89.0	91.2
Avg		87.67	90.34

It can be observed that the accuracy of the model classification proposed in this paper is very little affected by the number of videos. The accuracy rate of the experimental sample with a large number of videos is still maintained at a high level, this may be due to the generalization ability of ELM. The average correct rate of using ELM classification is 2.66% higher than that of neural network, which demonstrates the classification effectiveness of ELM.

Finally, in order to further study the classification performance of DCNN model training method, the semantics of news video will be classified and analyzed from the following five cases:

(1) Initialize DCNN weights using only IAGA (IAGA);

(2) Based on the BP algorithm, only the gradient descent method is used to train DCNN (BP-DCNN);

(3) Based on BP algorithm, DCNN is adjusted by GA and gradient descent method (GA-BP);

(4) Use the TLD model parameters in [13] to train the depth network. For the constraint analysis of different sparse regular terms, using the sparse penalty term of the L1 paradigm, the topology is grouped by the neurons of the hidden layer. (TLD-DCNN);

(5) Based on the BP algorithm, DCNN is adjusted by IAGA and gradient descent method (IAGA-DCNN).

In the comparative experiment, the crossover probability of the traditional genetic algorithm was chosen to be 0.8, the mutation probability was chosen to be 0.01. The average accuracy of the five sets of video data was recorded in Table 3. As a challenging task with extensional, the performance of test results for Trecvid 2012 are shown in Table 4.

It can be seen from Table 3 that the algorithm proposed in this paper gives better performance than the other four approaches. For the composite scene, IAGA-DCNN performance is still stand out, and for each type of semantic classification are achieved good results.

Table 3. Performance comparison of IAGA, BP, GA-BP, IAGA-BP training networks for news video semantics (%)

Semantic concept	DCNN training method				
	IAGA	BP-DCNN	GA-BP	TLD-DCNN	IAGA-DCNN
leader	2.45	69.70	87.96	87.24	**89.17**
traffic	3.19	82.40	91.40	91.53	**92.38**
meeting	2.20	78.11	82.33	82.67	**84.02**
fire	2.29	80.24	90.85	91.02	**91.64**
water	2.98	90.53	92.49	92.38	**92.96**
Avg	2.62	80.20	89.01	88.97	**90.03**

Table 4. Performance comparison of IAGA, BP, GA-BP, TLD, IAGA-BP training networks for Trecvid 2012 (%)

Semantic concept	DCNN training method				
	IAGA	BP-DCNN	GA-BP	TLD-DCNN	IAGA-DCNN
AF	2.59	69.7	75.96	65.72	**76.01**
Ba	2.27	90.11	95.29	94.45	**96.36**
Bu	1.84	79.44	87.73	**89.97**	88.28
Car	2.39	82.24	93.05	84.92	**93.14**
Dog	2.48	85.53	91.49	88.38	**92.60**
Fl	2.13	72.76	80.99	70.84	**81.45**
IM	1.9	72.33	89.13	87.90	**89.87**
Mo	2.65	86.98	90.52	85.14	**91.76**
ST	2.23	87.04	89.67	82.67	**89.83**
Sp	1.97	95.75	95.54	**96.77**	96.11
Avg	2.25	82.19	88.94	84.68	**89.55**

From the average accuracy given in this table, it can be seen that the accuracy is improved by 1.02% on the basis of GA-BP, which 1.06% higher than the correct rate of using the TLD training network. This indicates that the algorithm proposed in this paper not only make up for local stagnation defects of the traditional genetic algorithm, but also fully validates the effectiveness in semantic classification for news video.

From Table 4, it can be observed that the accuracy of classification for different semantics is significantly improved after using the method of network pre-training, especially the IAGA-DCNN structure makes the average classification accuracy rate reach 89.55%.

But for building and speech semantics, the performance of this algorithm is less than that of TLD training network. This may be due to the fact that such semantics contain more non-solid color regions, and TLD structure can extract more topology information. However, the classification accuracy of the remaining mixed scene is the highest, and the classification effect of different types of semantics is more stable when we use the IAGA-DCNN model.

6 Conclusion and Future Work

In this paper, a classification method combining genetic algorithm with convolution neural network is proposed. There are two advantages for initializing the multiple convolution neural network classifiers by the improved adaptive genetic algorithm and correcting the optimal weight of each unit by the gradient weight method. One is avoiding the over-fitting of the training set due to the depth of the network and excessive amount of data; and the other is avoiding the local training to be partially optimal. Finally, the global feature extracted by the optimal network model is input into the extreme learning machine, and the classification result can be concluded for news video semantics.

Three groups of experimental results show that: ①The network model proposed in this paper can achieve effective and accurate classification, and it has certain feasibility; ② The classification accuracy of this algorithm is superior to the traditional classification method, and the training performance is superior to the existing parameter optimization method; ③This algorithm still maintains the leading level of classification accuracy of complex scene semantics. However, the problem still exists for a small number of news video clips in which the subtitles without scene could not be properly identified. As our future work, we plan to build the dual-mode convolution network classification model combined with both scene semantics and subtitle meaning to improve accuracy. The performance improvement of the next stage will be aiming at specific convolution kernel number, the network depth and other parameters.

Acknowledgements

Thanks for the support of the following fund projects: (1) the National Natural Science Foundation of China under Grant No.61373112; (2) the Science and Technology Project Plan of Housing and urban-rural Ministry of Construction under Grant No. 2016-R2-045; (3) the Natural Science Basic Research Foundation of Shaanxi Province under Grant No. 2014JM8343.

References

[1] Z. Wu, Y. Fu, Y.-G. Jiang, L. Sigal, Harnessing object and scene semantics for large-scale video understanding, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] X.-C. Fan, Research on image semantic communication and retrieval based on massive resource library of internet, Journal of Computers (Taiwan) 28(1)(2017) 225-231.

[3] M. Wang, Z.-Z. Zhou, C.-H. Li, M.-F. Wei, L. Mao, Harris corner detection algorithm based on pixel point gray difference, Computer Engineering 41(6)(2015) 227-230.

[4] D.A. Rojas Vigo, F. Shahbaz Khan, J. van de Weijer, T. Gevers, The impact of color on bag-of-words based object recognition, in: Proc. International Conference on Pattern Recognition, 2010.

[5] W. Kong, Y. Zhan, Video semantic detection based on topographic independent component analysis and Gaussian mixture model, Journal of Computer Applications 36(3)(2016) 770-773.

[6] Y. Kamishima, N. Inoue, K. Shinoda, Event detection in consumer videos using GMM supervectors and SVMs, Eurasip Journal on Image & Video Processing 2013(1)(2013) 51.

[7] D3 Teney, M. Hebert, Learning to extract motion from videos in convolutional neural networks, in: Proc. Asian Conference on Computer Vision (ACCV), 2016.

[8] P. Peng, H. Chen, L. Shou, K. Chen, G. Chen, C. Xu, DeepCamera: a unified framework for recognizing places-of-Interest based on deep ConvNets, in: Proc. the 24th ACM International on Conference on Information and Knowledge Management, 2015.

[9] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis & Machine Intelligence 35(1)(2013) 221-231.

[10] S. Gupta, P. Arbelaez, J. Malik, Perceptual organization and recognition of indoor scenes from RGB-D images, in: Proc. of IEEE Conference on Computer Vision and Patten Recognition, 2013.

[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proc. the 25th International Conference on Neural Information Processing Systems, 2012.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.-F. Li, Large-scale video classification with convolutional neural networks, in: Proc. the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[13] Z. Zhan, Video semantic concept analysis based on convolution neural network, [dissertation] Jiangsu: Jiangsu University, 2016.

[14] b. Zhang, W. Yang, N.-N. Lin, Machine Learning and Visual Perception, Tsinghua University Press, Beijing, 2016.

[15] Y. Li, Z. Hao, H. Lei, Survey of convolutional neural networks, Journal of Computer Applications 36(9)(2016) 2509-2514.

[16] S. Tao, T. Zhang, J. Yang, X. Wang, W. Lu, Bearing fault diagnosis method based on stacked autoencoder and softmax regression, in: Proc. the 34th Chinese Conference Control Conference, 2015.

[17] The CIFAR-10 dataset and visual demonstration of classification from Stanford. <http://cs231n.stanford.edu/> 2017 (accessed 17.01.05).

[18] L. Guo, S. Ding, A. hybrid deep learning CNN-ELM model and its application in handwritten numeral recognition, Journal of Computational Information Systems 11(7)(2015)2673-2680.

[19] E.P. Ijjina, K.M. Chalavadi, Human action recognition using genetic algorithms and convolutional neural networks, Pattern Recognition 59(11)(2016) 199-212.

[20] Note on Convolutional Neural Networks from Cambridge. <https://pdfs.semanticscholar.org/714a/c6c7dbb83d69b8118e5138b3a50d8feb789b.pdf> 2017 (accessed 17.01.20).

[21] H. Yao, Y. Huang, BP-based estimate on network video QoE, Computer Engineering and Design 38(01)(2017) 1-6.

[22] J.Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[23] J. Zhang, T. Jiang, Improved adaptive genetic algorithm, Computer Engineering and Applications 46(11)(2010) 53-55.