EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Multichannel attention mechanisms fusion based on gate recurrent unit memory network for fine-grained image classification

Rui Yang[1,2*] and Dahai Li[1,2]

[1]School of Electrical Engineering, Zhengzhou University of Science and Technology, Zhengzhou, 450015 China
[2]Henan Intelligent Information Processing and Control Engineering Technology Research Center, Zhengzhou, 450015 China

Abstract

Attention mechanism is widely used in fine-grained image classification. Most of the existing methods are to construct an attention weight map for simple weighted processing of features, but there are problems of low efficiency and slow convergence. Therefore, this paper proposes a multi-channel attention fusion mechanism based on the deep neural network model which can be trained end-to-end. Firstly, the different regions corresponding to the object are described by the attention diagram. Then the corresponding higher order statistical characteristics are extracted to obtain the corresponding representation. In many standard fine-grained image classification test tasks, the proposed method works best compared with other methods.

Keywords: Multichannel attention mechanism, result fusion, fine-grained image classification, gate recurrent unit memory network.

Received on 14 January 2022, accepted on 20 January 2022, published on 27 January 2022

doi: 10.4108/eai.27-1-2022.173161

*Corresponding author. Email: 352772214@qq.com

1. Introduction

Fine-grained image classification is a hot research issue in the field of computer vision in recent years. Its purpose is to divide more detailed subcategories in coarse-grained large categories [1,2]. These subcategories usually have small inter class differences, and they often need to be distinguished by small local differences. For example, ring billed Gull and California gull in the bird data set are very similar, only the beak shape is quite different, and it is also difficult for humans who master relevant knowledge [3]. Compared with inter class differences, there are usually large intra-class differences in fine-grained image classification, including object pose, scale, occlusion and background. In particular, when the amount of data in each category is limited and there is no additional manual annotation information for object parts, it is a very challenging task to realize fine-grained image classification based on weak supervision information. According to the characteristics and difficulties of fine-grained image classification, introducing visual attention [4-6] mechanism to highlight the key parts of the image with distinction is a common idea in the research of fine-grained image classification in recent years. For example, Ying et al. [7] proposed the spatial transform network, which used soft attention to sample on the special map to obtain the morphological transformed features. Compared with the classical convolution network, it could extract the spatial feature information more effectively. The two level at-tension model proposed by Abdalla et al. [8] applied object level and part level attention. The

convolution network was used to obtain object level information, and then the clustering method was used to obtain key local areas, so as to make more accurate use of multi-level information. SP-DA-CNN [9] proposed by Barbier et al. used part annotation in Cub bird dataset to train the detection network to obtain hard attention corresponding to seven different parts of birds in the dataset, and cut the features at the corresponding positions for image classification. Suh et al. [10] combined visual attention with recursive structure, fused features and attention weights at each level of recursive network, and combined key regional features of multiple scales in the model.

The above methods have achieved good results in applying the attention mechanism to fine-grained classification, but there are still some limitations on the role of attention:

(1) For each attention and feature fusion process, the attention weight graph is a feature graph with the number of channels of 1, without using multi-dimensional attention features, this limits the ability of image feature extraction with complex distribution of key areas. In recent years, attention mechanism has been widely used in other fields outside the field of computer vision. Among them, the multi-head attention mechanism [11] generates multiple attention weight maps in parallel and integrates them with features at the same time, so that the model can obtain attention corresponding to different input positions, this method surpasses the previous methods based on complex models in tasks such as machine translation, and proves that multi-channel attention can provide more effective and comprehensive information.

(2) The method of attention weight map and image feature fusion is relatively simple. On the one hand, the method of multiplying the corresponding elements of attention weight map and feature map by position is adopted. On the one hand, it is unable to extract higher-order information more effective for classification, on the other hand, it is difficult to adapt to multi-channel attention features with more complex forms.

Based on the above analysis, this paper proposes a fine-grained image classification model based on multi-channel attention: a multi-channel attention generation method based on neural network is proposed to obtain rich spatial attention information by extracting multi-channel attention weight map; At the same time, a new attention and feature fusion method is proposed. By extracting the high-order information of image features corresponding to attention, the high-level features with more descriptive ability are obtained. Finally, a deep neural network model that can be trained end-to-end is formed. In the experiments on common fine-grained image classification data sets such as Cub-200-2011, FGVC-aircraft and Stanford Cars, compared with the mainstream fine-grained image classification methods in recent years, the classification accuracy obtained by this model has been significantly improved.

2. Principle of attention mechanism

2.1. Attention extraction

The role of attention can be regarded as the process of selecting some task related information from the input information, and the attention weight is the index of these information [12-14]. $x_{1:N} = [x_1, \cdots, x_N]$ represents input information, and the attention variable $z \in [1, N]$ can be used to represent the index position of the selected information, that is, $z = i$ represents that the i-th input information is selected. For the use of soft attention, it can make α_i represents the probability of selecting the i-th input information given the current input information $x_{1:N}$, that is, the attention weight to be extracted, then,

$$\alpha_i = p(z = i \mid x_{1:N}) = soft\max(s(x_i)) = \frac{\exp(s(x_i))}{\sum_{j=1}^{N} \exp(s(x_i))} \quad (1)$$

Where $s(x_i)$ is the attention scoring function, and the corresponding model can be selected according to the actual task and situation. For example, the point product model is used.

$$s(x_i) = x_i^T W \quad (2)$$

Where W is the learnable network parameter.

2.2. The fusion of attention and features

The method of attention weight acting on features can be regarded as the process of encoding input information under an information selection mechanism [15]. For a single dimensional soft attention weight graph, the most common way of attention and feature fusion is to multiply the corresponding position elements in the form of dot product.

$$attention(x_{1:N}) = \sum_{i=1}^{N} \alpha_i \cdot x_i = E_{z \sim p(z|x_{1:N}, q)}[x] \quad (3)$$

3. Proposed fine-grained image classification

The deep neural network model described in this paper can be divided into feature extraction, attention weight graph generation, attention weight and feature fusion,

classifier and so on. The feature extraction module transforms the input image into low-level features using full convolution network; The attention weight map generation module inputs image features to obtain multi-channel attention weights; The fusion module fuses the attention weight with the low-level features of the image to obtain the feature vector as the high-level representation of the image; The classifier transforms the attention fused feature vector into the probability corresponding to each category of the data set, so as to obtain the classification result. The above parts constitute an image classification model framework that can be trained end-to-end, as shown in Figure 1.

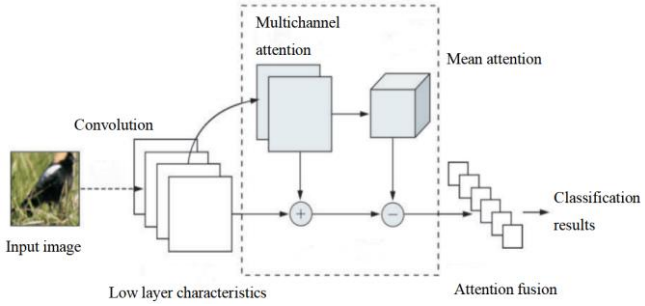

Figure 1. Structure of proposed fine-grained image classification model

3.1. Feature extraction

The feature extraction part contains a plurality of convolution layers, which can be converted from the pre trained network. For the input two-dimensional image, the output of the final convolution layer is H×W×D. The characteristic diagram of D can be regarded as group D characteristics, and each group contains N=H×W pieces of information respectively correspond to the corresponding spatial location, and the low-level features can be expressed as: $X \in R^{N \times D} = \{x_{i,d}\}$.

3.2. Generation of multi-channel attention weight

In the network model, the low-level feature x is shown above. The multi-channel attention of K dimension corresponds to K selection processes of input information, and the corresponding N×K the characteristic graph with attention weight N×K, expressed as $A \in R^{N \times K} = \{\alpha_{i,d}\}$. Where K is the number of attention weight graphs.

Multichannel attention is equivalent to multi-terminal attention applied to two-dimensional features of convolution output [16]. When multi-channel attention acts on the model, attention corresponds to multiple separate selection processes of input information, which act on input features in parallel, that is,

$$\alpha_{i,k} = p(z_k = i \mid x_{1:N}) = soft\max(s_k(x_i)) \quad (4)$$

Where $s_k(x_i)$ is the scoring function corresponding to the k-th attention. In order to ensure that the attention weights of different channels focus on different spatial positions in the feature map, the number of channels with attention weights applied by softmax function is 1:K.

For the input s of channel K, the softmax function acting on it is expressed as follows.

$$soft\max(s) = \frac{e^{s_j}}{\sum_{k=1}^{K} s_k} \quad (5)$$

The attention scoring function $s_k(x_i)$ used in this method is based on the dot product model most commonly used in the application of attention mechanism, and normalized according to the characteristics of the input image features, as shown below:

$$s_k(x_i) = \mid x_i \mid^T W_k \quad (6)$$

Where $\mid x_i \mid$ means that L2 normalization processing is performed on the input low-level feature x_i corresponding to its dimension D, which is helpful to produce a more stable attention weight, $W_k \in R^D$, convert the D-dimensional input to the output corresponding to the K-channel.

Attention weight $\alpha_{i,k}$ can be further obtained as follows:

$$\alpha_{i,k} = soft\max(\mid x_i \mid^T W_k) \quad (7)$$

The above attention weight generation process can be realized by common operations such as convolution layer and softmax layer in neural network, which ensures that the model can be trained end-to-end as a whole.

3.3. Function of multi-channel attention weight map

For the low-level features and attention weights with the same spatial dimension and the number of channels are D and K respectively, the action process of the two can be written according to the attention fusion method shown in equation (8).

$$attention_k(x_d) = \frac{1}{N} \sum_{i=1}^{N} \alpha_{i,k} \cdot x_{i,d} \quad (8)$$

The low-level features are expressed as X in the form of matrix, and the attention weight is expressed as

$A \in R^{\{N \times K\}} = \{\alpha_{i,k}\}$, the above operation can be expressed as:

$$attention(X, A) = \frac{1}{N} X^T X \quad (9)$$

After the operation, we can get the High level characteristics of D×K dimension.

Furthermore, the characteristic mean corresponding to each group of attention weights is introduced into the model. The mean value of attention is a network parameter, which represents the low-level features corresponding to all data, and corresponds to the mean value of attention of each channel. This operation is different from Vlad in image representation

The feature extraction process has similarities, and Vlad has been proved to be an effective image representation. The introduction of feature mean can extract higher-order features more related to categories, improve the expression ability of the output fusion results and improve the classification effect.

For the features X_i , d and attention weight in the above $\alpha_{i,k}$, the mean value of attention can be expressed as $\mu_{d,k}$. Then the attention fusion method above can be rewritten as:

$$attention_k(x_d) = \frac{1}{N} \sum_{i=1}^{N} \alpha_{i,k} \cdot (x_{i,d} - \mu_{d,k}) \quad (10)$$

The mean value of attention is written as $M \in R^{\{D \times K\}} = \{\mu_{d,k}\}$, then the attention fusion method can be expressed as:

$$attention(X, A) = \frac{1}{N}(X^T A - A^T \Theta M) \quad (11)$$

Where Θ represents the operation of dimension dot product corresponding to K . In equation (11), the action process of attention fusion is mainly composed of the multiplication of matrix and vector, which can easily realize the reverse operation:

$$\frac{\partial(attention(X, A))}{\partial X} = \frac{1}{N}(A^T \cdot \frac{\partial A^T}{\partial X} - M\Theta \frac{\partial A^T}{\partial X})$$

$$\frac{\partial(attention(X, A))}{\partial M} = \frac{1}{N} \frac{\partial A^T}{\partial X} \quad (12)$$

After the fusion operation including subtracting the mean of attention, the output high-level feature dimension is still D×K.

The generation parameter $W = \{W_k\}$ of attention weight in the above is a key parameter in the network model. This parameter can be initialized randomly according to the traditional method, but the introduction of category information that is not related to the image category plays a strong role in promoting a more descriptive attention weight map. At the same time, initializing this parameter helps to accelerate the network convergence, It can reduce training time. Therefore, a certain external category information is introduced to initialize the parameter W by clustering. In this method, the orthogonal matching pursuit (OMP-k) algorithm [17] is used to initialize the parameter W and obtain the minimum value of the following operation.

$$\min_{W,s} \sum_{i=1}^{N_{data}} \| Ws(x_i)^T - x_i \|$$
$$s.t. \| W_j \|_2^2 = 1, \forall j, \| s(x_i) \|_0 \le k, \forall i \quad (13)$$

Where $\| s(x_j) \|_0$ is the number of non-zero elements in $s(x_j)$. N_{data} corresponds to all data used for initialization. When K is 1, omp-1 algorithm is also called gain shape vector quantization or spherical k-means algorithm, which can be regarded as a special form of K-means algorithm. This algorithm can cluster the normalized features.

When initializing the model parameters, input the training data into the network convolution layer, collect the corresponding output characteristics through the network forward operation, and then calculate the weight according to the omp-1 algorithm. Meanwhile, the data mean $M = \{\mu_{d,k}\}$ can also be initialized using the features and attention weights collected from the initialization data, and the expression is:

$$\mu_{d,k} = \frac{1}{N} \sum_{i=1}^{N_{data}} x_{i,d} \cdot \alpha_{i,k} \quad (14)$$

3.4. Attention-based gate recurrent unit

The purpose of memory network is to retrieve the information needed to answer questions from the input visual information and store the valid information in memory. In order to improve the understanding of questions and images, especially when the problem requires transmission of reasoning, the memory network needs to transmit inputs multiple times, updating the remembered information after each transmission. The memory network is composed of the attention mechanism module and the memory update module. Each iteration will calculate the weight of the input vector through the attention mechanism to generate new memories. Then update the remembered information through the memory update module. AttnGRU refers to attention-based GRU model.

The original definition of GRU is as follows: For the input x_i of each time step i and the hidden state h_{i-1} of the previous time step, the updated hidden state $h_i = GRU(x_i, h_{i-1})$ can be calculated by the following formula.

$$u_i = \sigma(W^u x_i + U^u h_{i-1} + b^u) \quad (15)$$

$$r_i = \sigma(W^r x_i + U^r h_{i-1} + b^r) \quad (16)$$

$$\tilde{h}_i = \tanh(W x_i + r_i \circ U h_{i-1} + b^h) \quad (17)$$

$$h_i = u_i \circ \tilde{h}_i + (1 - u_i) \circ h_{i-1} \quad (18)$$

Where u_i is used to control the degree to which the state information of the previous moment is substituted into the current moment. r_i is used to control the degree to which the state information of the previous moment is ignored. h_i represents the updated status of the hidden layer. σ stands for Sigmoid activation function. x_i represents the current input at time i. h_{i-1} represents the hidden layer state of the previous time step. b is the offset term. $\tanh$ stands for hyperbolic tangent function. The symbol $\circ$ means multiply by the corresponding elements. $W^u, W^r, W \in R^{n_H \times n_1}$. n_H indicates the size of the hidden layer. n_I represents the size of the input.

In formula (15), the update gate u_i only uses the current input and the state of the hidden layer at the previous moment to calculate. Hence, the lack of any information from the question and memory of each prior moment.

To solve the above problems, update gate u_i in equation (15) is replaced with gate value g_i^t as the attention mechanism. Then the GRU can use the attention gate g_i^t to update the internal state, that is,

$$z_i^t = [v_i \circ q; v_i \circ m^{t-1}; |v_i - q|; |v_i - m^{t-1}|] \quad (19)$$

$$Z_i^t = W^2 \tanh(W^1 z_i^t + b^1) + b^2 \quad (20)$$

$$g_i^t = \frac{\exp(Z_i^t)}{\sum_{k=1}^{K} \exp(Z_i^t)} \quad (21)$$

$$h_i = g_i^t \circ \tilde{h}_i + (1 - g_i^t) \circ h_{i-1} \quad (22)$$

Where t represents the times of transmission and input of memory network. v_i refers to the object region represented by K 2048-dimensional vectors extracted from the image. m^{t-1} represents the memory of the last passed output (initial memory $m^0 = q$). q represents the problem embedding vector. The symbol $\circ$ represents the product of elements. $|\cdot|$ is the absolute value of the product of the elements. W is the weight matrix. b is the offset term. h_{i-1} represents the state of the hidden layer at the previous moment. h_i represents the state of the hidden layer at the current moment.

In order to generate the context vector for updating the episodic memory state m^t, the final hidden state of the GRU based on attention gate is used.

4. Experimental results and analysis

4.1. Model setting

This paper proposes that the convolution network of the feature extraction part of the model can be obtained from the pre-training model. In the experiment, VGG-16 network [18, 19] pre-trained in ImageNet data set is selected as the basic network of this part. The pre-trained vgg-16 network can effectively obtain rich convolution features, which is used as the basis in many depth neural network models. In this model, the last convolution layer output of the pre-training network, namely conv5_3. As a low-level feature in the model, the feature dimension is 512. In the experiment, the input image size of network convolution layer is 512×512 pixels. The image goes through several times before entering the network

Common data enhancement operations, including cutting part of the image at the ratio of 224/256, randomly mirroring the image, subtracting the image mean, etc. As described above, the attention weight map generation section may be composed of a convolution kernel with a size of 1×1 and Softmax. The parameters of the convolution layer are initialized by the OMP-1 method. In the experiment, the number of channels K of the attention weight graph is the key parameter of the network model, and the relationship between its value and classification accuracy can be determined through experiments. When the channel number of the basic attention weight graph is K=32, the channel number of the low-level features output from the convolution layer is 512, and the high-level features output after the fusion of attention and low-level features are a long vector with a dimension of 32×512=16384. In order to enhance the stability of this high-level feature as an image representation, L2 normalization is performed to obtain the final high-level feature.

The high-level features obtained by attention and feature fusion are input into the full connection layer, and the output dimension corresponds to the category

corresponding to the data set. After passing through the Softmax layer, the probability output of each category can be obtained. In this classifier, the input dimension of the full connection layer is high. In order to accelerate the network training speed, the high-level feature vectors corresponding to the training images can be collected to train the linear SVM classifier, and the parameters of the full connection layer can be initialized with the parameters of the SVM model.

4.2. Data sets

In order to comprehensively evaluate the performance of this method for fine-grained image classification, CUB-200-2011 bird dataset and FGVC are used Aircraft data set and Stanford cars data set [20] and other data sets commonly used in fine-grained image classification.

Caltech-UCSD birds-200-2011 fine-grained image data set, referred to as CUB-200-2011, is the most classic and commonly used data set in fine-grained image classification research at this stage. The CUB-200-2011 data set contains a total of 11788 images of 200 species of North American birds. According to the division provided by the data set, there are 5994 training images and 5794 test images. This data set has the characteristics of small difference between categories and small difference in images. Birds have challenging characteristics, such as diverse posture positions and limited training data.

FGVC aircraft fine-grained image classification data set contains 102 aircraft images of different models, each model contains 100 images, a total of 10200 images, about one third of which are used as tests. The main objects in the images in this dataset are different types of aircraft. Because many aircraft types in the dataset are divided in detail, the similarity between some categories is very high; The aircraft coating and the environment are different. The same results in large changes within the category, making FGVC aircraft a challenging fine-grained image data set.

Stanford cars fine-grained image classification data set contains 196 different types of car images, a total of 16185, of which 8144 images are used as training and others as testing. Cars dataset has the same vehicle model and manufacturer corresponding to many categories, and the perspective and coating of vehicles in the same category have great changes, which has strong fine-grained image classification characteristics.

For these fine-grained image classification data sets, this paper uses them according to the standard training and testing provided by the data set. There is no duplicate data between them, which ensures the effectiveness of the model and is easy to compare with other methods.

4.3. Experimental results and analysis

In Experiment 1, the classification results of cub-200-2011 dataset are configured according to the model described above. In this paper, the fine-grained classification model based on multi-channel attention obtains 87.7% classification accuracy in cub-200-2011 dataset, as shown in Table 1. Some of the comparison methods use additional supervision information outside the image category, including bounding boxes and location labels provided by the data set. In the control method, SPDA-CNN, mask CNN, CBCNN, B-CNN and RA-CNN all use VGG-16 as the basic network as the method in this paper, which is more helpful to compare the ability of the model to extract effective classification information based on the low-level features of the image. According to the experimental results in Table 1, the classification accuracy of this method is significantly improved compared with the previous weak supervised classification method without additional annotation; At the same time, compared with the labeling method of data sets such as parts, this method achieves the same level of classification accuracy. This result proves that the model based on multi-channel attention has the ability to effectively extract classification related features and distinguish fine-grained images.

Table 1. Classification accuracy of different methods on CUB-200-2011 dataset

Method	Classification accuracy /%
PB R-CNN	74.1
SPDA-CNN	85.3
Mask-CNN (VGG-16)	85.6
Mask-CNN (ResNet-50)	87.5
Two-level	78.2
CB-CNN	84.1
B-CNN	84.3
ST-CNN	84.3
PDFS	84.7
RA-CNN	85.6
Our method	87.7

Experiment 2 classification results of FGVC aircraft dataset and cars dataset.

According to the above configuration, the fine-grained classification model based on multi-channel attention obtains 88.4% classification accuracy in FGVC-aircraft data set; A classification accuracy of 92.5% was obtained in the cars dataset. Table 2 shows the comparison results of different methods in the two data sets. B-CNN [D,D] uses vgg-16 network as the basic network as the method based on depth neural network, which is the same as the method in this paper; B-CNN [D,M] combines the features extracted by VGG-16 and VGG-M [21]. It can be seen from the results in the table that the classification accuracy of this method is significantly improved compared with the previous methods. At the same time, combined with the complexity of the network model, it can be seen that when using the basic model with the same or smaller scale, the multi-channel attention model used in this method can extract the features related to fine-grained image classification more effectively.

Table 2. Classification accuracy of different methods on FGVC-Aircraft and Cars datasets

Method	Aircraft dataset /%	Cars dataset /%
Chai et al. [19]	72.5	78.0
Fisher Vector[20]	80.7	82.7
B-CNN[17] [D, M]	83.9	91.3
B-CNN [D, D]	84.1	90.6
Our method	88.4	92.5

Experiment 3 number of attention weight channels.

For the multi-channel attention model described in this paper, the channel dimension k of the multi-channel attention weight graph a in equation (11) is a key parameter. When the number of attention weight channels is low, it may be difficult to provide sufficient attention information and affect the classification results; When there are many attention weight channels, it will increase the model parameters and then increase the computational complexity of the model. At the same time, it will increase the dimension of the output image representation vector after attention, so it is difficult to obtain a compact image representation [22-24]. Table 3 shows the model classification accuracy obtained by training in the CUB-200-2011 dataset according to the model configuration described above when the number of channels of the attention weight map gradually increases and takes 4, 8, 16, 32 and 64 equivalents respectively. In the experimental results, when the number of attention channels is 4, the classification accuracy is significantly different from that when the number of attention channels is 8, up to 7.2%, which proves that the attention weight feature is not enough to provide sufficient information and has a great impact on the classification accuracy. When the number of channels in the attention weight map is not less than 16, the classification accuracy is close, and the model contains sufficient attention information. The experimental results show that taking the number of channels of attention weight map as 16 or 32 can achieve a good balance between classification accuracy and model complexity.

Table 3. Classification accuracy for the proposed model with different number of channels of the attention weight on CUB-200-2011 dataset

Number of channels of attention weight graph	Classification accuracy /%
4	78.4
8	85.6
16	87.0
32	87.5
64	87.6

Experiment 4 image representation features

In the model described in this paper, the high-level feature vector output after the action of attention in equation (11) can be used as a feature representation of the input image. At this time, taking this layer of the model as the output, the high-level vector is obtained as the feature extractor of the image. The dimension of this vector and the accuracy of image classification are the key factors to evaluate the performance of the model. Table 4 compares different image classification models with the ability to extract image feature vectors, and takes the feature vector dimension and the classification accuracy on the cub-200-2011 data set as the evaluation results. Among them, this method uses two configurations: the number of attention weight channels is 16 and 32 respectively. In the comparison method, CNN-FC uses the 4096 dimensional output of fc7 layer of vgg-16 as the representation vector, and vgg-16 is also the basic network of all methods in the table; CNN-IFV reduces the dimension from the output of fc7 and fc8 layers of neural network to obtain Fisher vector as image representation vector; B-CNN uses bilinear pooling to fuse the 512 dimensional outputs of two groups of convolution to

obtain a very high dimensional representation vector; CB-CNN method improves FB-CNN and reduces the dimension of representation vector while maintaining the classification accuracy. It can be seen from the results in the table that this method achieves better classification results in fine-grained image classification task while maintaining the representation vector with low dimension. This proves that the attention function method used in the model can extract important information helpful to classification more effectively.

Table 4. Comparison of different models'feature vector length and classification accuracy

Method	Eigenvector dimension	Classification accuracy /%
CNN-FC	4096	66.1
CNN-IFV[22]	51200	64.2
B-CNN[17]	2. 6e5	84.0
CB-CNN-RM[16]	8192	83.8
CB-CNN-TS[16]	8192	84.0
Our method (K=16)	8192	87.0
Our method (K=32)	16384	[illegible]

5. Conclusion

In the first mock exa[illegible] a d[illegible]ral network model for fine grained image classi[illegible]cation is proposed and verified. This model applies multi-[illegible]hannel visual attention, and extracts the higher-order information from the attention correspondence mean in the process of attention and image fusion. At the same time, a method of initializing attention parameters is proposed, which forms a set of image classification framework for training with end to end training. At the same time, it can be used to extract compact image representation. Experiments on a variety of fine-grained image classification data sets such as CUB-200-2011 show that compared with the traditional attention model and other classical fine-grained image classification frameworks, the fine-grained image classification model based on multi-channel visual attention has significant advantages in classification accuracy.

Acknowledgements.

The author greatly appreciates the anonymous comments of the reviewers.

References

[1] He G, Li F, Wang Q, et al. A Hierarchical Sampling Based Triplet Network for Fine-grained Image Classification[J]. Pattern Recognition, 2021, 115(3):107889.

[2] Liu X, Zhang L, Li T, et al. Dual attention guided multi-scale CNN for fine-grained image classification[J]. Information Sciences, 2021, 573(4).

[3] Liu C, Huang L, Wei Z, et al. Subtler mixed attention network on fine-grained image classification[J]. Applied Intelligence, 2021:1-14.

[4] Purcell J R, Rahn A, Fi[illegible] J M, et al. Neural correlates of visual [illegible]ention during risky decision evidence integr[illegible]on[J]. NeuroImage, 2021, 234(33):117979.

[5] Yan, [illegible] Li, H. G[illegible]PSO-MQC:medical image encryption based [illegible] genetic simulated annealing particle swarm optimizatio[illegible] and modified quantum chaos system. Evolutionary Intelligence, 14: 1817-1829, 2021. doi: 10.1[illegible]07/s12065-020-00440-6

[6] Khan A A, Uddin M, Shaikh A, et al. MF-Ledger: Blockchain Hyperledger Sawtooth-enabled Novel and Secure Multimedia Chain of Custody Forensic Investigation Architecture[J]. IEEE Access, 2021, PP(99):1-1.

[7] C. Ying, C. Hengshi and L. Guoqing, "Remote Sensing Image Registration Based on Spatial Transform Network and Phase Correlation Method," 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2019, pp. 125-128, doi: 10.1109/ICIIBMS46890.2019.8991540.

[8] Abdalla M, Silva J, Rocha R D. Notes on the Two-brane Model with Variable Tension[J]. Physical Review D, 2009, 80:046003.

[9] J. Barbier et al., "MAAP Annotate: When archaeology meets augmented reality for annotation of megalithic art," 2017 23rd International Conference on Virtual System & Multimedia (VSMM), 2017, pp. 1-8, doi: 10.1109/VSMM.2017.8346282.

[10] Suh T, Wilson R T, On S. Gender difference in visual attention to digital content of place-based advertising: a data-driven scientific approach[J]. Electronic Commerce Research, 2021:1-21.

[11] Kumar A, Seth S, Gupta S, et al. Sentic Computing for Aspect-Based Opinion Summarization Using Multi-Head Attention with Feature Pooled Pointer Generator Network[J]. Cognitive Computation, 2021:1-19.

[12] Zhou Z, Liu F. Filter Gate Network Based on Multi-head attention for Aspect-level Sentiment Classification[J]. Neurocomputing, 2021, 441(2).

[13] Yin Lyu, Lin Teng. Parallax information fusion-based for dance moving image posture extraction[J]. EAI Endorsed Transactions on Scalable Information Systems. 21(33), e8, 2021. http://dx.doi.org/10.4108/eai.6-10-2021.171247

[14] Laghari, A.A., Wu, K., Laghari, R.A. et al. A Review and State of Art of Internet of Things (IoT). Arch Computat Methods Eng (2021). https://doi.org/10.1007/s11831-021-09622-6

[15] Wang H, Wang W, Xiao S, et al. Improving Artificial Bee Colony Algorithm Using a New Neighborhood Selection Mechanism[J]. Information Sciences, 2020, 527.

[16] Liu, J., Zhang, J. & Yin, S. Hybrid chaotic system-oriented artificial fish swarm neural network for image encryption. Evolutionary Intelligence (2021). https://doi.org/10.1007/s12065-021-00643-5

[17] Zarei A, Asl B M. Automatic Seizure Detection Using Orthogonal Matching Pursuit, Discrete Wavelet Transform, and Entropy Based Features of EEG Signals[J]. Computers in Biology and Medicine, 2021, 131(5):104250.

[18] Laghari A A, Laghari M A. Quality of experience assessment of calling services in social network[J]. ICT Express, 2021(2).

[19] Qingwu Shi, Shoulin Yin, Kun Wang, Lin Teng and Hang Li. Multichannel convolutional neural network based fuzzy active contour model for medical image segmentation. Evolving Systems (2021). https://doi.org/10.1007/s12530-021-09393-3

[20] Peisen Wang, Yan Song, Lirong Dai. Fine-Grained Image Classification with Multi-channel Visual Attention [J]. Journal of Data Acquisition and Processing Vol. 34, No. 1, Jan. 2019, pp. 157-166.

[21] S. Yin and H. Li. Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582.

[22] Karim, S., He, H., Laghari, A.A. et al. Quality of service (QoS): measurements of image formats in social cloud computing. Multimed Tools Appl 80, 4507–4532 (2021). https://doi.org/10.1007/s11042-020-09959-3

[23] Karim S, He H, Laghari A A, et al. The Evaluation Video Quality in Social Clouds[J]. Entertainment Computing, 2020 (35):100370.

[24] Laghari, R.A., Li, J., Laghari, A.A. et al. A Review on Application of Soft Computing Techniques in Machining of Particle Reinforcement Metal Matrix Composites. Arch Computat Methods Eng 27, 1363–1377 (2020). https://doi.org/10.1007/s11831-019-09340-0