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	87.5

5. Conclusion

In the first mock exam, a deep neural network model for fine grained image classification is proposed and verified. This model applies multi-channel 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, Jahn A, Fine J M, et al. Neural correlates of visual attention during risky decision evidence integration[J]. NeuroImage, 2021, 234(33):117979.

[5] Yin, S., Li, H. GSAPSO-MQC:medical image encryption based on genetic simulated annealing particle swarm optimization and modified quantum chaos system. Evolutionary Intelligence, 14: 1817-1829, 2021. doi: 10.1007/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/s12530-021-09392-3

[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