A Survey of Fine-Grained Image Classification Based on Deep Learning

Chun-feng GUO¹, Hai-rong CUI², Kun YU¹ and Xin-ping MO¹

¹Shandong Foreign Trade Vocational College
²Qingdao University

Keywords: Deep learning, Fine-grained classification, Convolutional neural networks.

Abstract. The deep learning technology has shown impressive performance in various vision tasks such as image classification and object detection. Recent advances of deep learning techniques bring encouraging performance to fine-grained image classification which aims to distinguish subordinate-level categories. This paper reviews the major deep learning concepts pertinent to fine-grained image analysis and summarizes to the field, briefly introduces the representative method of fine-grained image classification based on deep learning. At present, the field of fine classification has developed a variety of methods including multi-network learning, target part detection and visual attention. Each method is to obtain a distinguishing area in the image, which help the network learn more effective features to complete the classification and recognition of fine-grained images.

Introduction

Fine-grained image categorization, also known as sub-category recognition, is a very popular research topic in the fields of computer vision and pattern recognition in recent years. The purpose is to make a more sub-classification of images (cars[1], dogs[2], flowers[3], birds[4], etc.) belonging to the same basic category, but fine-grained image classification is more difficult because of the sub-class differences between sub-categories and larger intra-class differences, compared to ordinary Image classification tasks.

The research of fine-grained image classification early based on algorithms on artificial features. Due to the limited ability of expression of artificial features, the classification effect often has great limitations. On the other hand, the traditional method relies on a large number of manual annotation information, which seriously restricts the practicality of the algorithm. In recent years, with the rise of deep learning, deep convolution features have promoted rapid progress in this field, and algorithms tend to no longer rely on manual annotation information, only using category labels to complete classification tasks.

Starting from the concept of fine-grained image classification, this paper introduces some excellent algorithms in this field and discusses the future research directions.

Overview of Fine-grained Image Classification

Image classification is a classic research topic in the field of computer vision. Traditional image classification mainly deals with two types: semantic level image[5,6] and instance level image[7,8]. The former includes semantic level image classification tasks such as scene recognition and object recognition. The goal is to identify different categories, such as cats and dogs. The latter is to classify different individuals, such as face recognition[9].

The fine-grained image classification is located between the two. Different from the coarse-grained image classification task such as object recognition, the fine-grained image has more detailed class precision, and the differences between the classes are more subtle, often only by small local differences. Since the classification boundaries are located on different sub-categories of the same category, such as different kinds of birds, they are called sub-category classifications. Compared with object classification tasks such as face recognition, the intra-class differences of fine-grained images are more huge, and there are many uncertain factors such as pose, light,
occlusion, and background interference. Therefore, fine-grained image classification is a challenging research task.

Unlike ordinary image classification tasks, the fine-grained image has a small signal-noise ratio, and information containing sufficient discrimination often exists only in very small local regions. Therefore, how to find and effectively use these useful local area information becomes the key to determine the success of fine-grained image classification algorithms. At present, most classification algorithms follow this process framework: first find the foreground object (bird) and its local area (head, foot, wings, etc.), then extract features from these areas separately, and appropriate the characteristics obtained. After the processing, it is used to complete the training and prediction of the classifier.

As mentioned above, fine-grained image classification is a very challenging research topic. In order to achieve satisfactory classification results, many existing classification algorithms rely heavily on manual annotation information. Some commonly used annotation information, including the label box and local area location. The detection of the foreground object can be done by means of the labeling box, thereby eliminating the interference of the background noise. The local area position can be used to locate and coordinate some useful local areas to achieve local feature extraction. However, the acquisition cost of manual annotation information is very expensive, which greatly restricts the practicability of these classification algorithms.

The extraction of features is also a key factor in determining the accuracy of image classification. Finding a more differentiated feature has always been the goal[10,11,12] pursued by researchers. Traditional artificial feature-based classification algorithms often face great limitations. Such algorithms generally extract the local features such as SIFT (Scale invariant feature transform)[13] or HOG (histogram of oriented gradient)[14] from the image, and then use VLAD (Vector of locally aggregated descriptors)[15] or Fisher vector[16,17] code model to perform feature encoding to get the final required feature representation. However, due to the limited ability to describe artificial features, the classification effect is often unsatisfactory. In the early stage of fine-grained image classification research, the ability of representation of features became the main bottleneck restricting its performance.

In recent years, the great success of deep learning in the field of computer vision has aroused people's strong research interest[18,19]. Compared to the artificial feature facet, deep learning can be seen as a process of express learning[20]. Studies have shown that features extracted from deep convolutional neural networks have more powerful description capabilities than artificial features, and deep convolution features can be applied to fine-grained image classification tasks to achieve better results[21]. The addition of deep convolution features brings new opportunities for the development of fine-grained image classification, making its research enter a new stage.

Research on Fine-grained Image Classification Algorithm

Since deep convolutional networks can learn very robust image feature representations, most of the methods for fine-grained image classification are based on deep convolutional networks. These methods can be roughly divided into the following four directions:

Fine-tuning Method Based on Conventional Image Classification Network

Most of these methods directly use the common deep convolution network to directly classify images, such as AlexNet [22], VGG [23], GoogleNet[24], ResNet[25], and DenseNet [26] and SENet [27].etc.

Since these classification networks have strong feature representation capabilities, good results can be obtained in conventional image classification. However, in fine-grained classification, the difference between different species is actually very subtle. If the conventional image classification network is directly used to classify fine-grained images, the effect is not satisfactory. Inspired by the theory of migration learning, one approach is to migrate a network trained on large-scale data to a fine-grained classification and recognition task. A common solution is to use the network weights
pre-trained on ImageNet as the initial weights, and then fine-tune the weights of the network on the fine-grained classification dataset to obtain the final classification network. (As shown in Figure 1)

![Figure 1. Multitask learning network structure[28].](image)

In [28], Zhang et al. further introduced the metric loss function into the fine-tuning of the fine classification network. Specifically, each time three samples (Positive, Reference, and Negative) are input into three networks of shared weights, and then the feature outputs of the three networks are used to calculate the loss function, except for the traditional softmax loss function, the three feature output also constitutes a generalized triplet loss. The last two loss functions are combined to fine tune the network to increase the ability of the network to identify similar samples in different categories.

### Method Based on Fine-grained Feature Learning

Lin et al. published in the ICCV paper [29], proposed a bilinear convolutional neural network model (Bilinear CNN, network structure shown in Figure 2) to achieve a better representation of deep convolution features. This method uses VGG-D and VGG-M networks as the reference network, and achieves 84.1% classification accuracy on the CUB200-2011 data set without using Bounding Box information. While using Bounding Box, its classification accuracy is as high as 85.1%.

![Figure 2. Bilinear CNN network structure.](image)  
![Figure 3. Part-RCNN algorithm flow chart.](image)

A bilinear model $M$ consists of a four-tuple: $M = (f_A; f_B; P; C)$. Among them, $f_A$ and $f_B$ represent the feature extraction function, that is, the convolutional network $A$ and the convolutional network $B$ in Figure 2, $P$ is a pooling function, and $C$ is a classification function.

If the feature dimensions extracted by the two characteristic Functions $f_A$, $f_B$ are $K \times M$ and $K \times N$, respectively, the output of the pooling function $P$ will be an $M \times N$ matrix, and the features need to be taken before classifying them. The matrix is stretched into a list of $MN$-sized feature vectors. Finally, the role of the classification function is to classify the extracted features, which can be implemented using logistic regression or SVM classifiers.

In general, the bilinear CNN model can effectively identify fine-grained images based on a concise network model. On the one hand, the CNN network can perform high-level semantic feature acquisition on fine-grained images, and iteratively trains convolution parameters in the network model to filter uncorrelated background information in the image. More importantly, on the other hand, Network A and Network B play complementary roles in the image recognition task,
that is, Network A can locate objects in the image, and Network B completes the feature extraction of the objects that are located on Network A. In this way, the two networks can cooperate with the process of class detection and target feature input of the fine-grained image to complete the fine-grained image recognition task.

Method Based on Object Part Detection and Alignment

The idea of the object part detection method is to first detect the location of the target in the image, then detect the location of the distinguishing region in the target, and then send the target image (ie, foreground) and the discriminative area simultaneously to the deep convolutional network for classification. However, based on the method of object part detection, it is often necessary to use the target bounding box annotation information in the training process, or even the key feature point information in the target image. In practical applications, it is very difficult to obtain the annotation information. More representative is the Part-RCNN method proposed in the 2014 ECCV [30].

The flow chart of the Part-RCNN algorithm is shown in Figure 3. Its main idea is to detect the position of the bird and the position of the bird's head and the bird's body by means of the classical method R-CNN [31] in the target detection, and then input the three parts of information into the deep convolution network for training.

First, like R-CNN, Part RCNN also uses the bottom-up region algorithm Selective Search[32] to generate region candidates, as shown in the upper left corner of Figure 3. After that, the R-CNN algorithm is used to detect these regional candidates, and the scores are given. Here, Part R-CNN only detects foreground objects and two local areas[33]. The area test results are selected based on the detection score (in the middle of Figure 3). The author believes that the detection score given by R-CNN does not accurately reflect the quality of each region. For example, the label box given for the head detection may be outside the label box of the object detection, and the result of the body detection may overlap with the result of the head detection. These phenomena all affect the final classification performance. Therefore, it is necessary to correct the detection area, mainly considering the use of the area’s border constraints and geometric constraints.

\[ X^* = \arg \max_X \Delta(X) \prod_{i=0}^n d_i(x_i) \]  
(1)

\[ \Delta_{box}(X) = \prod_{i=0}^n c_{box}(x_i) \]  
(2)

\[ \Delta_{geometric}(X) = \Delta_{box}(X) \left( \prod_{i=1}^n \delta(x_i) \right)^\alpha \]  
(3)

Among them, the formula 1 is the improved scoring function, and the formula 2 and the formula 3 are the frame constraint and the geometric constraint bar, respectively. After correcting the position information detected by the R-CNN by using the constraint conditions as described above, the convolution feature is extracted for each block region, and the features of different regions are connected to each other to form a final feature representation to train the SVM classifier. In the CUB200-2011 dataset, the algorithm uses AlexNet as the backbone network. If the target labeling frame is not used during the test, the recognition accuracy is 73.89%. If the target phase is marked in the test phase, the recognition accuracy is 76.37%.

Method Based on Visual Attention Mechanism

The Vision Attention Mechanism is a signal processing mechanism unique to human vision. The specific manifestation is that it first obtains the target area that needs attention by quickly scanning the global image when the visual system looks at things, and then suppresses other useless information to obtain the target of interest. At present, the visual attention method based on CNN network is widely applied to computer vision, including tasks such as target detection and
recognition. In deep convolutional networks, attention models can also be used to find regions of
interest or discriminative regions in an image, and for different tasks, the region of interest of the
convolutional network is different.

Since the Vision Attention Model-based method can locate differentiated regions in an image
without the need for additional annotation information, it has been Widely used in the fine-grained
classification of images. The representative work is the Recurrent Attention Convolutional Neural
Network (RA-CNN) proposed in the 17-year CVPR [34]. The model mimics the RPN (Region
Proposal Network) network in the fast-RCNN [35], proposes to use the APN (Attention Proposal
Network) network to locate the discriminative regions in the image, and by using the sorting loss
function during the training process to ensure that each time you use the attention model to locate
the area is more effective.

As shown in Figure 4, the entire model is composed of three VGG19 networks, and the feature
map of the first VGG network convolutional layer output is sent to the APN network for learning,
and the coordinates of the distinguishing region of the entire image and the radius of bounding box
are obtained. Assuming that the current VGG network input image is X, and all parameters of the
VGG convolutional layer are \( W_c \), the convolution feature \( F = W_c \ast X \), where \( \ast \) indicates all
convolutions in the network during the forward transfer of the convolution feature. Pooling and
activation function operations. The APN network can be expressed as \( [t_x, t_y, t_l] = g(W_c \ast X) \), where
\( t_x, t_y, t_l \) are the horizontal and vertical coordinates of the center of the distinguishing region in the
input image X, and the radius of the region.

![RA-CNN network algorithm flow chart.](image)

After obtaining the coordinates and radius of the discriminative region, the author uses the
coordinates to crop the region from the upper VGG input image, and then enlarges it and inputs it to
the next VGG network for further learning, and iterating in turn. Finally, the authors sent these three
scale images to the deep convolution network for training, feature fusion, classification, and
achieved 85.3% accuracy in the CUB-200 data set.

**Future Research Direction**

The research on fine-grained image classification is in the ascendant, and it is urgent to carry out
further research. Regarding possible future research directions, it can be considered from the
following aspects:

1. Build a higher quality standard database: The fine-grained image database used by current
mainstream research has a common shortcoming, although there is a lot of room for choice: the
scale and fineness of the data are not too high, the quality of labels and the number of categories are
also very limited. As we all know, the performance of deep learning is positively correlated with the
size of the database. The richer the image, the more obvious the performance improvement and the
more practical. Therefore, how to build a higher quality standard database becomes an important issue that needs to be solved urgently in future research.

2. Effective use of local area information: A feature of fine-grained image recognition that is different from ordinary image classification tasks is that information with discrimination is hidden in local areas. How to use these local information more effectively will become a breakthrough in future research. Image classification is the main direction of future research. How to effectively complete the localization detection of local areas under the premise of only category marking is undoubtedly a big challenge.

3. Image recognition in natural scenes: Fine-grained image classification is a research topic closely related to practical applications, and its ultimate goal should be to serve real life. However, the databases used in current academic research generally have the characteristics of prominent foreground objects and single background. Such pictures are not common in real life. If you want to make a fine-grained image recognition system widely used in natural scenes, you have to consider image recognition problems in complex scenes such as illumination, blur, occlusion, low-split rate, and object interference. These factors are often lacking in the current application. There is a stronger actual demand in real life, and it is worthy of the future work.

4. Expansion to other fields: In fact, fine-grained image classification is a comprehensive research topic, and should not be limited to one field of image classification. It needs to be extended to other research directions of computer vision, such as image retrieval and object detection. More research content still needs to be further explored.

Conclusions

Fine-grained image classification algorithm is a hot research topic in the field of computer vision. The emergence of deep learning has brought new development opportunities. This paper introduces the development of fine-grained image classification algorithms based on convolution features in recent years. The detection and feature extraction of the local information of the two core tasks in the fine-grained classification are discussed in detail, and the possible future development opportunities in this field are summarized.

Acknowledgement

This research was financially supported by the Project of Shandong Province Higher Educational Science and Technology Program Foundation under grant number J18KA395.

References

[1] Krause J, Stark M. Deng J, Li FF. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (IICCW). Sydney, Australia, IEEE, 2013.554-561.

[2] Khosla A. layadevaprkash N. Yao B P. Li FF. Novel dataset for fine-grained image categorization. In: Proceedings of the 1st Workshop on Fine-grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Springs, USA: IEEE, 2011.

[3] Nilsback M E.Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the 6th Indian Conference on Computer Vision. Graphics Image Proceasing. Bhubaneswar, India: IEEE. 2008. 722-729.

[4] Wah C. Branson S, Welinder P. Perona P. Belongie s. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-201 1001, California Institute of Technology Pasadena CA. USA. 2011.
[5] Bosch A, Zisserman A, Munoz X. Scene classification using a hybrid generative/discriminative approach. IEEE Trans. actions on Pattern Analysis and Machine Intelligence. 2008.30(4):712-727.

[6] Wu X Rehg J M. Centrist: a visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011. 33(8): 1459-1501.

[7] Gehler P, Nowozin S. On feature combination for multiclass object classification. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan IEEE. 2009.221-228.

[8] Jarnett K, Kavukcuoglu K, Ranzato M, Lecun Y. What is the best multi-stage architecture for object recognition? In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE. 2009. 2146-2153.

[9] Wright J, Yang A Y, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009. 1(2):210-227.

[10] Zhang Lin-bo, Wang Chun-heng, Xiao Bai-hua, Shao Yun Xue. Image representation using bag-of-phrases. Acta Automatica Sinica, 2012. 38(1): 46-5.

[11] Yu Wang-sheng, Tian Xiao-hua, Hou Zhi-qi. A new image feature descriptor based on region edge statistical. Chinese Journal of Computers, 2014, 37(6): 1398-1110.

[12] Yan Xue-Jun, Zhao Chun-xia, Yuan Xia, 2DPCA-SIFT: an efficient local feature descriptor. Acta Automatica Sinica. 2014, 40(4):675-682.

[13] Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 1099. 1150-1157.

[14] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE. 2005. 886-893.

[15] Jegou H, Douze M, Schmid C, Perez P. Aggregating local descriptors into a compact image representation. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010.

[16] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition Minneapolis, USA: IEEE. 2007. 1-8.

[17] Sanchez Perronnin E, Mensink I, Verbeek Image classification with the Fisher vector: theory and practice. International Journal of Computer Vision. 2013, 105(3): 222-245.

[18] Goering C, Rodner E, Freytag A, and Denzler J. Nonparametric part transfer for fine-grained recognition. In: CVPR, pages 2489–2496. IEEE, 2014.

[19] Gao Ying-ying, Zhu Wei-bin. Deep neural networks with visible intermediate layers. Acta Automatica Sinica. 2015. 41(9): 1627-1637.

[20] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015.521(753):436-44.

[21] Donahue J, Jia Y Q, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. DECAF: a deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: ACM, 2014. 647-655.

[22] Krizhevsky A, Sutskever I, Hinton G E, et al. ImageNet Classification with Deep Convolutional Neural Networks[C]. Neural Information Processing Systems (NIPS), 2012: 1097-1105.
[23] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]. International Conference on Learning Representations (ICLR), 2015.

[24] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1-9.

[25] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

[26] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 1(2): 3.

[27] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

[28] ZHANG Xiaofan, et al. Embedding label structures for fine-grained feature representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 1114-1123.

[29] Lin T Y, Roychowdhury A, Maji S. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, pp. (99):1-1.

[30] Zhang N, Donahue J, Girshick R B, et al. Part-Based R-CNNs for Fine-Grained Category Detection[J]. European Conference on Computer Vision (ECCV), 2014: 834-849.

[31] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2014: 580-587.

[32] Uijlings J R, De Sande K E, Gevers T, et al. Selective Search for Object Recognition[J]. International Journal of Computer, 2013, 104(2): 154-171.

[33] Wei, X., Zhang, Y., Gong, Y., Zheng, N.: Kernelized subspace pooling for deep local descriptors. In: CVPR. 2018.

[34] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Advances in neural information processing systems (NIPS). 2015: 91-99.

[35] Jianlong Fu, Heliang Zheng, Tao Mei. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2017, 4476-4484.