Researches advanced in fine-grained image classification based on convolutional neural network

Shiyao Xu

Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China

xsy@emails.bjut.edu.cn

Abstract. Due to practical needs, fine-grained image classification (FGIC) has been considered for many years as a direction of study in computer vision, which aims to subdivide images belonging to one coarse-grained category into multiple fine-grained classes. Traditional fine-grained image classification algorithms rely heavily on annotations. Recently, convolutional neural networks (CNN) have prefigured unprecedented opportunities for this research direction with the popularity and development in deep learning. To start, this study introduces the development history with various fine-grained image classification algorithms, as well as definition and research significance of the problem. After that, it compares and analyzes the different algorithms respectively in the aspects of strong supervision and weak supervision. This paper also compares the accuracy of these models on frequently used datasets. We conclude the paper by summarizing and evaluating the different aspects of these algorithms, and then discuss possible future research domains and challenges in this field.

Keywords: Fine-grained Image Classification, Convolutional Neural Network, Weakly Supervised Learning

1. Introduction

Image classification has remained an important research direction in the computer vision field for years. Image categorization tasks can be divided into cross-species semantic level and fine-grained categorization in accordance with the similarity between image classes. Image classification generally refers to cross-species image classification when not specifically stated. In cross-species image classification, the features vary greatly between multiple classes, which are characterized by large between-class variance and small within-class variance, or can be unrelated multiple classes, e.g., classifying images into human, cat, dog, and other classes.

Different from cross-species image classification, fine-grained image classification classifies its subclasses under a specific class with the aim of distinguishing images with smaller differences between classes that have more fine-grained features, such as classifying images of dogs into multiple dog breeds including malamute, pug, and husky. Compared with the former, fine-grained image recognition tasks have the feature of small between-class differences and large within-class differences, so it is difficult to distinguish between classes with small differences using traditional methods. Take dogs as an example, malamute and husky breeds are similar from body size to hair color, with only minor differences in bone distribution, which is generally difficult for non-specialists to distinguish. In real life, fine-grained image classification algorithms have been extensively adopted in many fields, like precise species categorization in wildlife conservation and accurate classification of product images in search and recommendation algorithms for e-commerce platforms.

Early research on fine-grained image recognition mainly rely on handcrafted features. In 2013, Berg et al. extracted POOF features by a local information feature encoding method and achieved an accuracy of 56.8% on CUB-200-2011. Benefiting from convolutional neural networks' powerful feature expression capacity, frameworks using deep learning for fine-grained image recognition have gradually taken a dominant position in recent years. According to whether the training data contains data other than image and category labels (such as the position of targets and components in the image), the previously proposed models for fine-grained image recognition can be divided into two categories: strongly supervised-based methods and weakly supervised-based methods.

(1) Strongly supervised based fine-grained image classification. In this category of classification methods, each image has category label information in addition to other manually labeled information, such as local regions in the image, that can be utilized in the classification algorithm. Strongly supervised-based methods usually improve or add other modules to convolutional neural networks (CNN) to improve the classification accuracy. Although strongly supervised learning-based methods are more effective in terms of classification accuracy, they require additional data annotation, like bounding box, for each image sample before learning, which can only rely on manual annotation by professionals in the classification field, so it largely increases the cost of data annotation before learning, resulting in a lower practical value than weakly supervised learning-based methods. This results in a lower practical value than weakly supervised learning-based methods.

(2) Weakly supervised based fine-grained image classification. Compared with strongly supervised based learning, these belong to this category of methods are more interested in reducing the manually labeled information in the data to achieve classification. With the reduction of manually labeled data, training usually becomes easier, but the classification accuracy of some weakly supervised-based algorithms is slightly weaker than that of strongly supervised learning. The two main types of weakly supervised algorithm implementations are Bilinear CNNs and algorithms implemented using attention mechanisms. The Bilinear Convolutional Neural Network (Bilinear CNN) proposed by Lin et al. [1] in 2015, the architecture of this model consists of two feature extractors, each performing feature extraction on the images, which are then bilinearly fused with the two features, and the accuracy of fine-grained classification is increased. But because of its high time complexity of bilinear process computation, it can take too much time in training and inference, and is more difficult to apply in practical usage scenarios. Many follow-ups are based on this study to improve and optimize the operation speed of the bilinear convolutional neural network. Humans do not focus directly on a certain part of the world when they perceive the outside world through vision, but scan the whole scene first and identify the area of interest before looking at it carefully, which is the principle of attention mechanism. The attentional mechanism was first used by Two-level attention [2] for this sub-direction of the task, where the algorithm finds several components with strong features in the image by attention and converges the results that are separately predicted by components and objects.

Focusing on the above two broad categories of classification algorithms for fine-grained images, this paper first introduces the current representative algorithms in this field, as well as the basic processes and advantages and disadvantages of the algorithms; secondly, the datasets and evaluation metrics commonly used in this field are presented; then, the top-1 accuracy of representative algorithms on the datasets is analyzed and compared; finally, the problems in the field are summarized and the future development of the field is given an outlook.

2. Methods based on strongly supervised learning

Part-based R-CNN proposed by Zhang et al. [3] is an improvement on the R-CNN, and by bottom-up detection, local features can be extracted and features from different regions can be fused to form the global target features. The algorithm can effectively extract the key information of the target local area. Branson et al. [4] proposed an algorithm that was obtained by improving Part-based R-CNN [3], on the basis of which pose recognition is performed on local images for computing the local features that are needed for classification, so the model's name is Pose Normalized CNN. The pose localization and normalization to obtain the features are computed using a deep convolutional network. The existing pose normalization scheme was also investigated in the study, and a new graph-based scheme was proposed, which could lead to a more accurate classification of bird species. Wei et al. [5] proposed the Mask-CNN, where the whole model is designed based on full CNN. Based on the part-level annotation, localization and classification are performed by the full convolutional network, and a mask of the object or part is generated. Then, in order to fuse object-level descriptors with part-level descriptors, a model with three-stream Mask-CNN structure is constructed.

Part-stacked CNN algorithm was proposed by Huang et al. [6] with fine-grained component-level supervision used in the model and multiple components on the target labeled with key points in the data. The algorithm structure can be decomposed into two parts, the network for locating the components' positions and the network for categorization. In order for the model to have the function of generating feature maps, a full convolutional network is used, which is directly used for the part of the network that has a categorization function to position the components, and a part with a two-stream structure is included in the categorization network to capture the features of each target and component at the same time, and to encode them as features. Lam et al. [7] proposed HSnet, which is a strongly supervised based model. The model first extracts part information through the network, then performs sequential retrieval to obtain parts that are valuable for fine-grained classification, by fusing the regions of useful parts with the original candidate frames.

3. Methods based on weakly supervised learning

3.1. Methods based on Bilinear CNN

Lin et al [1] proposed Bilinear CNN in 2015, in which two CNNs are used in the network to capture the features of the samples and the features obtained from the two extractions are combined bilinearly. The introduction of bilinearity simplifies the computation, while the network has end-to-end characteristics under the premise of relying only on image category labels. The Bilinear convolutional neural networks can effectively mine discriminative features closely related to categories in images, which has become a classic framework for weakly supervised fine-grained image recognition. Bilinear CNNs are effective in classification, but the number of model parameters usually reaches hundreds of thousands to millions due to the characteristics of the model using bilinear operations, which makes training and deployment more difficult and impractical. A series of subsequent studies have been obtained by improving on Bilinear CNNs, such as Compact Bilinear Pooling, Low-rank Bilinear Pooling [8], and Improved Bilinear Pooling with CNNs [9].

Kong et al. [8] proposed Low-rank Bilinear Pooling, a model that avoids the direct computation of bilinear features by matrix decomposition with reduced rank of the bilinear classifier, which reduces the parameters of the model to a great extent and shortens the time spent on classification computation at the same time. Lin et al. [9] proposed Improved Bilinear Pooling with CNNs, attempted to improve the performance of Bilinear CNN related methods by investigating the normalization method with matrix square root algorithm, improved the accuracy of the algorithm classification by matrix square root normalization, and improved the accuracy of calculating the gradient by replacing the singular value decomposition (SVD) method with a method that solves Lyapunov's equation to estimate the gradient of matrix square root.

3.2. Methods based on attention mechanisms

In 2015, Xiao et al. [2] proposed the two-level attention model, by trying to utilize attention theory in fine-grained image classification, where a novel combination of bottom-up attention, target-level and top-down attention at the component level for a total of three types of attention, allowed the model to extract regions of objects in the image with the more characteristic component regions, and fuse the extracted object-level and component-level predictions to obtain classification results. In 2017, Fu et al.

[10] proposed an RA-CNN model structure with three layers of subnetworks, which extract key regions in images by convolutional neural and attentional suggestion subnetworks (APN), and then crop and amplify the region and output it to the next layer of subnetworks for further prediction, to make the algorithm classification better, the feature regions are learned recursively at more than one scale-wise. Zhao et al. [11] proposed Diversified Visual Attention Networks (DVAN) in 2017, which, unlike previous studies where attention can only be extracted from a certain region, improves the diversity of attention to obtain multiple attention regions, and then achieves joint determination in multiple regions by adjusting their loss function. In 2017, Zheng et al [12] proposed a model with a similar design idea to DVAN [11], both of which attempt to discover the identity of several regions or part of target in the picture by attention, to then classify and discriminate each part separately, so the model is named as Multi-Attention Convolutional Neural Network (MA-CNN), so it can learn more characteristic fine-grained features. Zheng et al. [13] proposed TASN. The innovation of the TASN algorithm is the trilinear attention module, and the functional block for sampling using attention theory, resulting in a model that can display feature-rich regions with high resolution and better accuracy than MA-CNN.

In 2020, Ji et al. [14] proposed ACNet, a network with a binomial tree structure using the attention principle that includes convolutional operations to solve fine-grained image classification. The structure of the binary tree from the root node to the leaf nodes is consistent with the directional characteristics of feature granularity in images, and the convolutional computation occurs at the edges of the tree, where each node gets the route beginning with the leaf and ending with the root through a function, while the sum of the leaf node classification speculations is the final prediction. In 2022, Zhu et al. [15] proposed Dual Cross-Attention Learning (DCAL), in which the model DCAL contains two attention modules, "global-local cross-attention (GLCA)" and "pairwise cross-attention (PWCA)", in which GLCA enhances the interaction between global and local feature regions of an image, and PWCA constructs the interaction between images. DCAL reduces misleading attention and allows the network to discover more complementary regions.

4. Experiments & Performance analysis

4.1. Common dataset

Common fine-grained image recognition datasets mainly include the CUB-200-2011 [16], Stanford Cars [17] and FGVC-aircraft [18]. Birds [16] includes 200 bird classes, each species can be found in the corresponding Wikipedia article, with 11788 images, which are divided into two parts, 5994 and 5794, where the part with more samples is for training and the other with fewer samples is for testing. Each sample has annotated data of bounding boxes, partial locations, and attribute labels. For the Stanford Cars [17], there are 196 categories in the Stanford Cars dataset, including cars, pickup trucks, SUVs, and other models, with 16185 images, which are divided into two parts, 8144 and 8041, where the part with more samples is used for training and the part with fewer samples is used for testing. The labels of the categories are determined by the brand of the car, manufacturing-related information, and so on, and all images in each category originate from the network, and duplicate identical images are removed. In the FGVC-aircraft dataset [18], there are 100 categories, each corresponding to an aircraft variant, with a total of 10,000 images, divided into two different large parts with 6667 and 3333 samples, respectively, where the part with more samples is used to train the model and the part with fewer samples is used to test the model, with the labels of the categories determined according to model, variant, family and manufacturer.

4.2. Evaluation metrics

In image classification tasks top-1 accuracy is often used to compare the accuracy between models. Suppose a model is asked to classify images into N classes, and let there be n samples within the testset, and there are n_t samples within the testset with the highest probability that the class predicted by the model matches the actual class, then top-1 Accuracy can be expressed as $\frac{n_t}{n}$.

4.3. Performance analysis

The classification results of those representative methods mentioned above were compared on the datasets birds [16], cars [17] and planes [18], which are several datasets commonly used in the field, where the accuracy results can be seen in Table 1. Strongly supervised learning-based models usually require training data other than category labels, and usually use bounding boxes with partial locations as additional training data in the training process. The only common dataset used to evaluate several strongly supervised models mentioned in the previous section is CUB-200-2011 [16], so the datasets mentioned in this paragraph are CUB-200-2011 unless otherwise specified. Part-based R-CNN [3] with additional modeling of the pose of the target, which is beneficial for fine-grained classification results. Inspired by the pose modeling operation in [3], Pose Normalized CNN [4] was proposed, which does the pose alignment operation locally on top of the former model, enabling the algorithm to improve the classification accuracy up to 75.7% on CUB-200-2011 by 1.2%. Part-stacked CNN [9] and Mask-CNN [5] achieved good results by using full convolutional networks for localization, with classification accuracies of 76.2% and 87.3%, respectively. HSnet [7] achieved 87.5% accuracy by integrating the retrieved part regions into the original candidate frames for classification, and achieved 93.9% accuracy on Stanford Cars [17].

Models based on weakly supervised learning are usually trained only by the labels of the images, without extra annotation data like bounding boxes. The main if-supervised learning-based ones introduced above are bilinear convolutional neural network-based methods and attention-based methods. Bilinear CNN [1] extracts feature by two CNNs, which makes the acquired features more diverse and rich, and achieves 84.1%, 91.3%, and 84.1% accuracy on birds [16], cars [17], planes[18], respectively. However, due to the large amount of parameters that are difficult to apply practically, Low-rank Bilinear Pooling [8] and improved methods [9] make improvements in the counts of variables to enhance the speed of training and inference, and both schemes basically reach or even mostly exceed the performance of Bilinear CNN [1] on all three datasets. Attention-based classification methods can obtain the location of key parts without additional labeling. The attention-based methods mentioned above include Two-level attention [2], DVAN [11], RA-CNN [10], MA-CNN [12], TASN [13], ACNet [14], DCAL [15], all of these methods basically use attention to partial localization to improve the classification effect, and some of them further improve the classification on accuracy by adding multiple kinds of attention to improve the diversity of features acquired by attention. The classification accuracy on CUB-200-2011 reached 92.0% for DCAL [15] model from 77.9% for Two-level attention [2] model. Higher accuracy was also achieved on Stanford Cars, FGVC-Aircraft at DCAL [15] with 95.3% and 93.3%, respectively.

Table 1. Performance comparison (Top-1 accuracy) of different methods on various datasets.

Methods	Year	Category	Top-1 accuracy on Datasets		
			Birds [16]	Cars [17]	Planes [18]
Part-based R-CNN [3]	2014	Strongly supervised	73.9%	-	-
Pose Normalized CNN [4]	2014		75.7%	-	-
Part-stacked CNN [6]	2016		76.2%	-	-
Mask-CNN [5]	2018		87.3%	-	-
HSnet [7]	2017		87.5%	93.9%	-
Bilinear CNN [1]	2015	weakly supervised +B-CNN [1]	84.1%	91.3%	84.1%
Low-rank Bilinear Pooling [8]	2017		84.2%	90.9%	87.3%
Improved Bilinear Pooling with CNNs [9]	2017		85.8%	92.0%	88.5%
Two-level attention [2]	2015	weakly supervised +Attention Mechanism	77.9%	-	-
DVAN [11]	2017		79.0%	87.1%	-
RA-CNN [10]	2017		85.3%	92.5%	-
MA-CNN [12]	2017		86.5%	92.8%	89.9%
TASN [13]	2019		87.9%	93.8%	-
ACNet [14]	2020		88.1%	94.6%	92.4%
DCAL [15]	2022		92.0%	95.3%	93.3%

As seen in Table 1, the strongly supervised model HSnet achieved an accuracy of 93.9% on Cars [17] in 2017, while the weakly supervised TASN [13] achieved 93.8% accuracy on the same dataset in 2019, almost at the same level, and TASN used training data containing only image and category labels. The recent model DCAL [15], which uses a weakly supervised approach, performs even far better than HSnet [7] on the three datasets mentioned in Table 2. Moreover, because of the expensive data acquisition, strongly supervised-based algorithms have low utility. Weakly supervised learning-based algorithms are indeed the mainstream direction in the current field, aiming at minimizing the reliance on manual data annotation while accomplishing fine-grained image classification with higher accuracy.

5. Discussion

Although existing research has greatly improved the performance of fine-grained image classification, there remain the following challenges to be solved in this field for practical applications:

(1) There are currently insufficient high-quality datasets available for FGIC. Since the sub-direction of categorization began to be studied, the datasets which have been frequently applied in studies so far are basically the datasets which were made public nearly 10 years ago (citing the above three datasets), and many of the datasets have a large number of samples from the Internet, and the people who perform data labeling are not experts and have only been trained in simple classification labeling, so the correctness of data labeling may be problematic. And the existing datasets are basically limited to samples of birds, cars, dogs, airplanes, etc., and lack of datasets for other objects with a large number of subclasses. With the advancement of deep learning models and computer computing power becoming faster, high-quality sample collections usually enhance the quality of model training and inference, and the production of high-quality datasets is something which requires research in the future.

(2) Strongly supervised based models are dependent on additional data annotation of samples, which leads to high data collection costs and poor utility. Most of the strongly supervised based models rely on information such as bounding boxes, partial locations, attribute labels, etc., and the labeling of these information requires a large amount of human cost, and even a large number of experts for labeling if high quality labeling is desired. With the increasing volume of data and also the difficulty of manual labeling data tasks, strongly supervised learning based on strong supervision is becoming less and less popular [19]. However, some design ideas or functional modules are useful in some strongly supervised-based models, so strongly supervised models can be considered to be modified to be weakly supervised in order to move away from the reliance on data labeling.

(3) The sample imbalance problem in some datasets may cause the model to have difficulty in learning the features of classes with few samples, which affects the model classification accuracy. There are two potential solutions to this problem. Expansion of the existing dataset by adding samples can be considered, or to equalize the sample counts of classes as much as possible during the construction of the new dataset. The dataset can also be augmented by data enhancement methods to make the size of all categories of image sets as balanced as possible. The amount of images in the categories is close to each other before training, so that the model can learn more balanced for each category and thus increase its classification accuracy. Also finding more effective data augmentation methods may enable the new samples obtained from augmentation to be learned better by the model.

(4) Some of the models have a greater amount of variables, which makes it unlikely that they can be used practically in realistic scenarios. Some models improve the classification accuracy but do not pay attention to the speed of training and inference. A large number of parameters will lead to slower speed in training and inference, and too large will lead to too slow speed for practical use. To enhance the temporal performance of existing models, simplification or compression of the models is essential. Simplifying and compressing the model to shorten the training and inference time and reduce the hardware requirements of the model while maintaining the classification accuracy of the model is a problem worth investigating in the future.

6. Conclusion

With the advancement of deep learning and pattern recognition techniques, weakly supervised and convolutional neural network based approaches have become the mainstream framework for fine-grained image classification tasks. This paper briefly introduces the development history of algorithms related to fine-grained image classification, and compares as well as analyzes representative fine-grained recognition algorithms from both strongly and weakly supervised aspects, respectively. In addition, the results of these algorithmic models on several commonly used datasets are compared. Finally, the rationale of these algorithms is summarized, and potential future study directions and current challenges encountered in this field are discussed and concluded.

References

[1] Lin, T.Y., RoyChowdhury, A., Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 1449-1457.

[2] Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 842-850.

[3] Zhang, N., Donahue, J., Girshick, R., Darrell, T. (2014). Part-based R-CNNs for fine-grained category detection. In European conference on computer vision. pp. 834-849.

[4] Branson, S., Van Horn, G., Belongie, S., Perona, P. (2014). Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952.

[5] Wei, X. S., Xie, C. W., Wu, J., Shen, C. (2018). Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition, 76: 704-714.

[6] Huang, S., Xu, Z., Tao, D., Zhang, Y. (2016). Part-stacked CNN for fine-grained visual categorization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1173-1182.

[7] Lam, M., Mahasseni, B., Todorovic, S. (2017). Fine-grained recognition as hsnet search for informative image parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2520-2529.

[8] Kong, S., Fowlkes, C. (2017). Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 365-374.

[9] Lin, T. Y., Maji, S. (2017). Improved bilinear pooling with cnns. arXiv preprint arXiv:1707.06772.

[10] Fu, J., Zheng, H., Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4438-4446.

[11] Zhao, B., Wu, X., Feng, J., Peng, Q., Yan, S. (2017). Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia, 19(6): 1245-1256.

[12] Zheng, H., Fu, J., Mei, T., Luo, J. (2017). Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 5209-5217.

[13] Zheng, H., Fu, J., Zha, Z. J., Luo, J. (2019). Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5012-5021.

[14] Ji, R., Wen, L., Zhang, L., Du, D., Wu, Y., Zhao, C., ... Huang, F. (2020). Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10468-10477.

[15] Zhu, H., Ke, W., Li, D., Liu, J., Tian, L., Shan, Y. (2022). Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4692-4702.

[16] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. Computation & Neural Systems Technical Report, 2010-001.

[17] Krause, J., Stark, M., Deng, J., Fei-Fei, L. (2013). 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554-561.

[18] Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A. (2013). Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151.

[19] Jiang, Y., Li, X., Luo, H., Yin, S., Kaynak, O. (2022). Quo vadis artificial intelligence?. Discover Artificial Intelligence, 2(1), 1-19.