Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070208
ISSN 2371-8838 Vol. 7 Num. 2

Research on Fine-grained Image Recognition

Jingyuan He[1,2], Bailong Yang[1,*]

[1]Department of Operational Support, Xi'an Research Institute of Hi-Tech, Xi'an, Shaanxi, 710025, China

[2]School of Mathematics and Computer Science, Yan'an University, Yan'an, Shaanxi, 716000, China

**Corresponding author*

Keywords: Fine-grained image recognition, localization-classification subnetworks, feature encoding, attention mechanism, external auxiliary information

Abstract: Fine-grained image recognition (FGIR) has developed rapidly in past decade. It has important scientific significance and application value in fields of intelligent new economy and industrial internet of things. In this paper, the challenges of FGIR, methods of FGIR and FGIR datasets are introduced. Finally, we summarized and discussed the research direction of further exploration in this field.

1. Introduction

Fine-grained image recognition (FGIR) aims at visual recognition of different subcategories [1], such as different subcategories of birds, different subcategories of flowers, different types of aircraft, etc. Fine-grained image recognition has been widely used in commodity recognition in smart retail scenarios, vehicle and pedestrian rerecognition in public security scenarios, dangerous goods detection and recognition, biodiversity monitoring and many other fields, especially in the industrial application of intelligent new economy and industrial Internet shows great practical value.

FGIR has developed a series of recognition methods with good performance, which have made deep research progress and wide practical application, focusing on how to explore the fine but resolution object component level information in the image and how to obtain the image representation with fine-grained characterization ability. However, the acquisition of fine-grained image requires a lot of human resources and financial resources, especially in some specific tasks, domain experts are required to participate in the process of image annotation, which brings huge challenges to FGIR.

2. Main Challenges of FGIR

Since there is little difference between different categories, how to accurately and efficiently identify different categories of grain-size images is a very important challenge. The challenge of FGIR task is mainly due to the fact that the components of objects from different subcategories are generally the same, while the components of the same subcategory are rich in diversity. These factors make it difficult for machines to accurately identify the categories of these objects, and even make it difficult for ordinary humans to identify these differences and diversity. Only through a wealth of expert knowledge can it be accurately identified. FGIR mainly has the following characteristics: (1) there are large intra-class differences and small inter-class differences among fine-grained categories. For example, in the CUB data set, the posture, background, and perspective of the same species of seagull vary greatly, but different species of seagull show a high degree of similarity, and differences only exist in some small areas, such as beak or wing. (2) Fine-grained images have complex background information. Complex background information in fine-grained data sets can not provide effective value information conducive to recognition, but will increase the difficulty of accurate recognition. For example, a bird in a bush, the background usually consists of branches and leaves, and the boundary between the bird and the background is indistinguishable. With the difference of time or Angle, the brightness will be very different, occlusion makes only part of object in image, due to different shooting equipment, image may have low resolution, fuzzy and unclear problems.

3. Methods of FGIR

FGIR is widely used in real life. Fine-grained identification of animal species can be used to identify biological information in ecosystem protection, and fine-grained identification of fruits and other commodities can be used in intelligent retail industry. However, it is a major challenge to accurately carry out FGIR because of small differences between subcategories and large differences within subcategories. Researchers deal with the problems of fine-grained images from different aspects. There are three main methods of FGIR: (1) FGIR based on location-classification subnetworks; (2) Using end-to-end feature encoding for FGIR; (3) Using external auxiliary information for FGIR. The first and second methods supervise model training by using information such as image label, boundary box and fine-grained object attribute carried by fine-grained image itself. However, due to characteristics and challenges of fine granularity, researchers gradually try to use more external but cheap information (such as network data and text description) to help fine-grained recognition, so as to further improve the accuracy, which is the third method.

3.1. Methods of FGIR Based on Location-Classification Subnetworks

To address the challenges posed by intra-class variation, researchers have focused on how to obtain discriminant characteristics of fine-grained objects. In the location-classification subnetworks, the location subnet is designed for locating key components of fine-grained objects, while the classification subnet is used for classification. The two subnets work together to finally complete fine-grained identification task. As shown in Figure 1, the existing methods can be divided into three main types: (1) fine-grained recognition based on detection or segmentation technology; (2) Fine-grained recognition based on depth filter; (3) Fine-grained recognition based on attention mechanism.

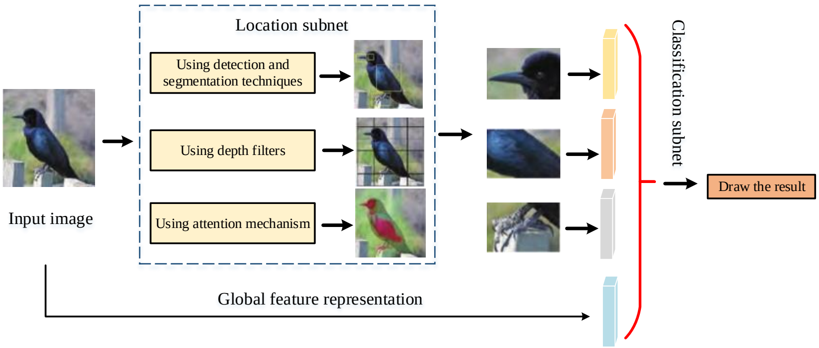

Location subnet
Using detection and segmentation techniques
Using depth filters
Using attention mechanism
Input image
Classification subnet
Draw the result
Global feature representation

Figure 1: Methods of FGIR based on localization-classification subnetworks.

3.1.1. FGIR Based on Detection or Segmentation Techniques

FGIR based on detection or segmentation technology refers to the use of detection or segmentation technology to locate the key areas corresponding to fine-grained images, such as birds can locate the head, tail and wings of birds. According to the local information such as boundary frame or segmentation mask, more recognizable middle-level feature representation can be obtained, which can be used to further improve the learning ability of classification subnet, so as to improve accuracy of final recognition.

Earlier work in this type of approach used a lot of additional object place annotations to locate key parts of fine-grained objects. Zhang et al. [2] first proposed a component-level bounding box annotation, and then trained a region-convolutional neural network (R-CNN) model as a key regional detector. Semantic segmentation has a more accurate local positioning effect, because it replaces rough boundary box annotations and is done at a more granular pixel level. However, training with traditional detector or segmentation models requires intensive component-level labeling, which severely affects the scalability of FGIR. Therefore, method of accurately locating fine-grained parts using only image-level labels [3-6] has been proposed one after another, and has gradually become a hot topic. Since only image-level labels are used in this method, it is called "weak supervision" FGIR method. In addition, some methods attempt to obtain a more powerful and generalized fine-grained representation by learning the interrelationships between component-level features, and combine local features for learning by implementing different feature fusion strategies, such as short-short memory network [4], knowledge distillation [5] or graph [6]. The results show that this method has higher recognition accuracy than the previous independent local feature learning.

3.1.2. FGIR Based on Depth Filter

In deep convolutional neural network (DCNN), the depth filter refers to learning weights at the convolutional layer. Researchers have found that the intermediate CNN output can connect the semantic part of the common object, so people try to use the filter output as a component detector. One of the main advantages of relying on them for FGIR is that no part-level markup is required. In order to facilitate detection and classification learning, a unified end-to-end training fine-grained model [7-9] has been developed.

3.1.3. FGIR Based on Attention Mechanism

Although previous fine-grained local classification method has shown strong classification performance, its main disadvantage is that the parts of the object need to have supervisory information. In many real-world scenarios, it may be difficult to define certain parts of an object, such as unstructured objects like food or flowers. A more natural solution to finding local locations than the previous approach is to use the attention mechanism as a submodule. This makes CNN focus on defining areas of fine-grained objects, so attention mechanism is a promising direction.

Zheng et al. [10] took the lead in using attention mechanism to improve the accuracy of fine-grained object recognition. The accuracy of FGIR can be improved by using attribution-guided attention mechanism to extract image features. Combining channel attention and conducted metric learning can strengthen the correlation between different participating regions. Using attention mechanism to acquire local and global features, and using hash for classification, can promote the efficiency of image retrieval and recognition. It should be pointed out that although attention mechanism can improve FGIR accuracy, there is a higher risk of overfitting small-scale data..

3.2. Methods of FGIR Based on End-to-end Feature Encoding

Because differences between subcategories are usually small, capturing global semantic information with the full connection layer limits representation ability of model, thus affecting the final FGIR. Methods of FGIR based on end-to-end feature encoding is shown in Figure 2. At present, there are two proposed methods: (1) Fine-grained recognition based on high-level feature encoding; (2) Fine-grained recognition based on novel loss function.

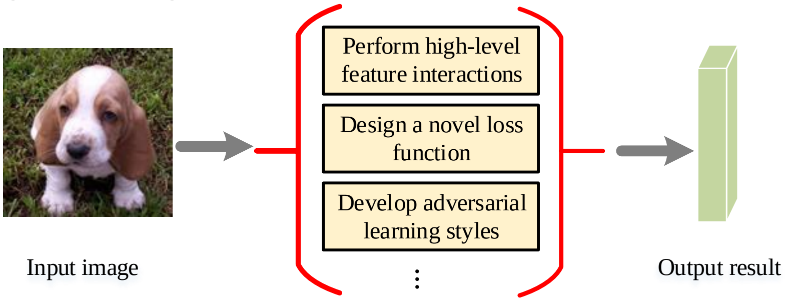

Perform high-level feature interactions
Design a novel loss function
Develop adversarial learning styles
⋮
Input image
Output result

Figure 2: Methods of FGIR based on end-to-end feature encoding.

3.2.1. FGIR Based on High-level Feature Encoding

Feature learning plays an important role in almost all visual tasks. In initial stage of deep learning, features of the full connection layer are usually used as image representation. Later, it is found that the feature mapping of the top convolutional layer contains more abundant information, which makes convolutional features widely used. Compared with the fully connected output, the encoding technique on convolutional neural networks significantly improves results of FGIR. In part, these improved coding techniques come from higher-order statistical coding of final features.

A representative high order feature interaction technique is covariance matrix based representation. By combining covariance matrix based representation with depth feature representation, it shows good accuracy in fine-grained recognition. One of the most representative methods is bilinear convolutional neural network, which represents image as two deep convolutional neural networks, and then decodes the second-order statistical code. This method has significantly improved the fine-grained recognition. However, this approach can lead to overfitting, especially on large data sets. Li et al. [11] simulated interaction of paired features using low-rank constraints and quadratic variations. Some methods also attempt to capture features of higher orders to produce stronger representations.

3.2.2. FGIR Based on Novel Loss Function

Loss function plays an important role in construction of deep network, which can directly affect classification results and model function. Therefore, the design of fine-grained loss function is also an important method of FGIR.

In FGIR, samples between classes may be visually very similar. Following this principle, using pairwise confusions optimization procedures to address overfitting and sample-specific FGIR, and subsequently reduced confidence in their prediction of oversize, thereby improving generalization ability. Humans can distinguish effectively by comparing images, and such contrast learning is common in FGIR. Attentive pairwise interactionnet (API-Net) distinguishes images through the interaction of the two pairs of attention. Designing a single loss function to locate the local area and further enhance the image level representation has gradually become a research focus. Some people proposed a module that forces networks to distinguish categories quickly, which can better distinguish fuzzy and confused fine-grained categories.

3.3. Method of FGIR Based on External Auxiliary Information

In addition to the traditional fine-grained recognition methods, another approach is to use external auxiliary information, such as network data, multi-modal data and so on. Methods of FGIR based on external auxiliary information include FGIR based on network data, FGIR based on multi-modal data and human-computer interaction.

3.3.1. FGIR Based on Network Data

Massive and well-marked images can be used as datasets to improve accuracy of FGIR. However, the tagging of massive high quality data requires a lot of cost. At the same time, the excellent results of network data in fine-grained recognition make scholars focus on how to use network data. Fine-grained image recognition based on network data can be divided into two directions. The first direction is to improve accuracy of FGIR by using the free but noisy data on the network to collect and collate the generated data set for training. This method is called webly-supervised learning. Network supervision learning methods mainly focus on eliminating the gap between network data and well marked standard data set, so as to reduce negative impact of noise data in network data set. Scholars frequently use adversarial learning techniques of deep learning and attention mechanism to solve problems generated by characteristics of network datasets. The other direction is to transfer knowledge by using well-labeled auxiliary classes as training sets.

3.3.2. FGIR Based on Multi-modal Data

With rapid growth of multimedia data, how to use multimedia data for FGIR has attracted wide attention. Multi-modal data based fine-grained recognition is to use text information or knowledge map and other multimedia data to help the model to carry out FGIR. Frequently used multimodal data includes textual descriptions (such as natural language clauses and phrases) and graphically structured knowledge bases. Multi-modal data belongs to the weakly supervised type. In addition, content in multimodal data (such as text descriptions) can be annotated without the need for domain experts, and ordinary people can use their own knowledge to give relatively accurate feedback. Among the knowledge base constructed by graphs, high-level knowledge graphs are a common resource, which contains rich professional knowledge and can provide a good auxiliary guidance for fine-grained recognition.

3.3.3. FGIR Based on Human-computer Interaction

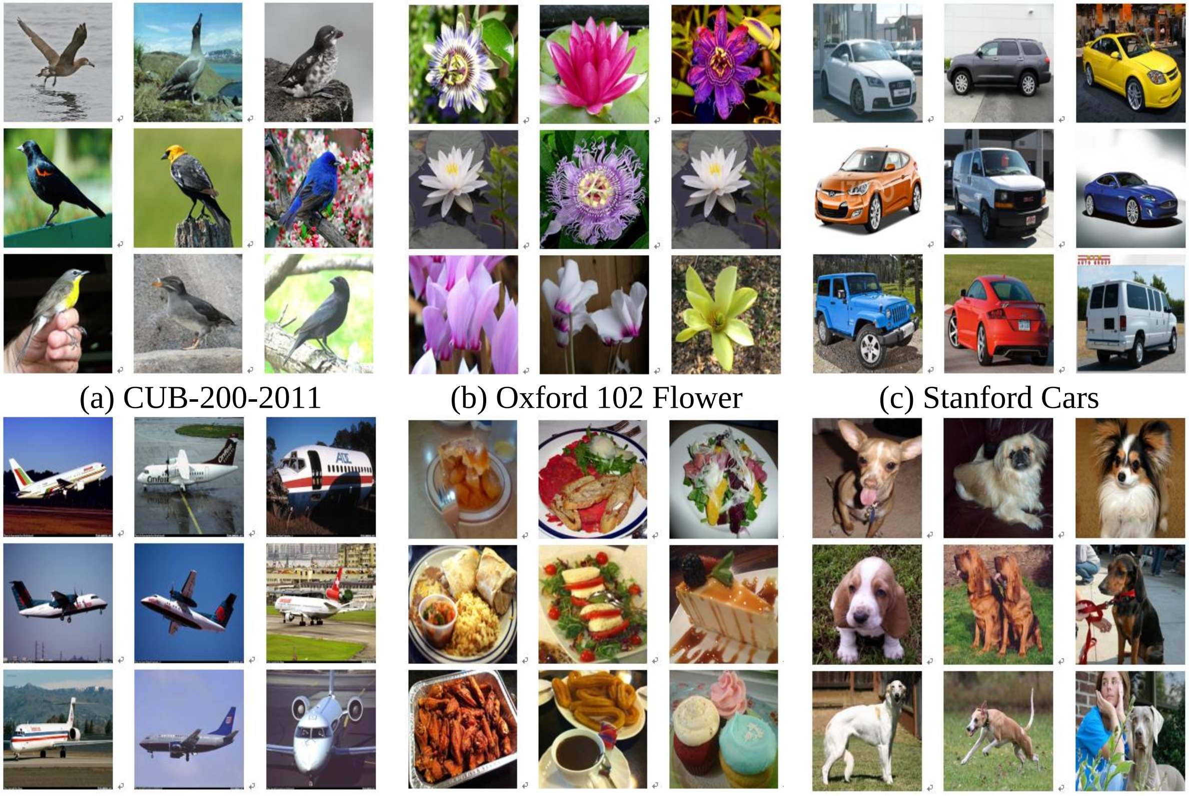
FGIR of human-computer interaction is an iterative system consisting of machine and human users, combining human intelligence guidance and machine intelligence, requiring the system to work as much as possible in the manner of human labor. In general, systems in each round seek to understand how humans perform recognition.

4. FGIR Datasets

FGIR aims to classify a large number of subcategories with little difference under traditional semantic categories. Many fine-grained baseline data sets have emerged in recent years, covering everything from birds, dogs, vehicles, aircraft, flowers, food, fruit, retail goods, and so on. A series of public fine-grained image recognition data sets have been published in the field to uniformly evaluate the fine-grained recognition accuracy of corresponding methods and promote the development of related technologies. Table 1 lists the widely used fine-grained image recognition datasets and gives information about each dataset [1]. Figure 3 shows examples of these fine-grained images. To some extent, the establishment of these fine-grained benchmark data sets shows the realistic demand of visual intelligence in contemporary society. Fine-grained benchmark data sets not only serve as a common basis for measuring model effectiveness, but also move the field of fine-grained recognition in a more practical direction.

Table 1: Fine-grained image recognition datasets.

Datasets	Number of categories	Number of images	Training set	Test set
CUB-200-2011	200	11788	5994	5794
Oxford 102 Flower	102	8189	4712	4017
Stanford Cars	196	16185	8144	8041
FGVC-Aircraft	102	10200	6667	3333
Food-101	101	101000	75750	25250
Stanford Dogs	120	20580	1200	8580



(a) CUB-200-2011 (b) Oxford 102 Flower (c) Stanford Cars

(d) FGVC-Aircraft (e) Food-101 (f) Stanford Dogs

Figure 3: Examples of fine-grained image recognition datasets.

In addition, more and more fine-grained data sets are being proposed that are more practical and more challenging. For example, large-scale retail product checkout dataset (RPC) for fine-grained commodity perception in intelligent retail scenarios, and iNaturalist for natural species such as different animals and plants. Some concrete characteristics of real data distribution can be found from these novel and realistic data sets, such as large scale and long tail distribution. These data characteristics and distribution characteristics can show practical problems in real life from the side.

5. Conclusion

FGIR is a long-term hot field in computer vision and pattern recognition. Since deep learning requires large-scale data with high-quality labels for training, practicability and scalability are constrained. In order to solve this problem, using free data on the network to train the model has become a feasible direction. The development direction of FGIR in future has following aspects: generate large-scale fine-grained image recognition datasets, few-shot learning for FGIR, automatic model of FGIR, and FGIR in more realistic settings.

References

[1] Wei, X.S., Wu, J.X. and Cui, Q. (2019) Deep learning for fine-grained image analysis: a survey. arXiv preprint arXiv:1907.03069.

[2] Zhang, N., Donahue, J., Girshick, R. and Darrell, T. (2014) Part-based R-CNNs for fine-grained category detection. In Proceedings of the European Conference on Computer Vision. Berlin: Springer, 834–849.

[3] Peng, Y., He, X. and Zhao, J. (2017) Object-part attention model for fine-grained image classification. IEEE Transactions on Image Processing, 27, 1487–1500.

[4] Ge, W., Lin, X. and Yu, Y. (2019) Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3034–3043.

[5] Liu, C., Xie, H., Zha, Z.J., Ma, L., Yu, L. and Zhang, Y. (2020) Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In Proceedings of the AAAI Conference on Artificial Intelligence, 34, 11555–11562.

[6] Wang, Z., Wang, S., Li, H., Dou, Z. and Li, J. (2020) Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In Proceedings of the AAAI Conference on Artificial Intelligence, 34, 12289–12296.

[7] Wang, Y., Morariu, V.I. and Davis, L.S. (2018) Learning a discriminative filter bank within a cnn for fine-grained recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4148–4157.

[8] Ding, Y., Zhou, Y., Zhu, Y., Ye, Q. and Jiao, J. (2019) Selective sparse sampling for fine-grained image recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 6599–6608.

[9] Huang, Z. and Li, Y. (2020) Interpretable and accurate fine-grained recognition via region grouping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8662–8672.

[10] Zheng, H.L., Fu, J.L., Mei, T. and Luo, J.B. (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 5219–5227.

[11] Li, W., Wang, L.M., Li, W., Agustsson, E. and Van, G.L. (2017) WebVi-sion database: visual learning and understanding from web data. arXiv preprint arXiv: 1708. 02862.