Studies advanced in weakly supervised fine-grained image classification based on deep learning

Zhuoxi Chen

Chongqing Jiaotong University, Chongqing, 400074, China

632005010109@mails.cqjtu.edu.cn

Abstract. Recently, the concept of fine-grained classification has arouse much attention., which has aroused heated disscussion of academia and industry. The extraction of picture characteristics in early efforts on fine-grained image classification relied on dense annotations, but acquiring these annotations was time-consuming and labor-intensive. Lately, weakly supervised fine-grained image classification has gradually emerged, which can mainly be separated between techniques using the attention mechanism and techniques using various neural networks. In this paper, focusing on the above two types of frameworks, we first introduce representative algorithms, including their innovation, basic processes, advantages and disadvantages. We then quantitatively compare the results of different algorithms on mainstream datasets, where the attention based methods can achieve excellent accuracy. We finally discuss the existing challenges and future development of the fine-grained classification task, which we believe can provide some new insight for this task.

Keywords: fine-grained classification, weakly supervised, attention mechanism, convolution neural networks.

1. Introduction

Fine-Grained Categorization Recognition is a popular research topic in computer vision, whose purpose is to divide the subcategories belonging to the meta category more finely. As a derivative technology of conventional image classification, The sub-categories of the sub-categories can be further broken down by the fine-grained image classification. At present, fine-grained classification has been applied in many life scenarios. Whether it is the recognition of daily retail purchases or the application of automatic driving, Fine-grained image classification must serve as the foundation for future research and development. However, the classification of fine-grained images still presents several difficult problems. For example, some categories have great difficulty in distinguishing them, so that they cannot be classify, or too large data makes the classification efficiency too low, and so on. Therefore, It would be very important for both academia and industry if low-cost, fine-grained picture recognition could be accomplished with computer vision technology.

Since the development of fine-grained image classification, the classification system has experienced and evolved for a long time. In the early stage, the study of fine-grained image classification focused on extracting features from local discriminative regions. In the composition of feature-to-feature localization, due to the complicated features and the limited ability to express features in encoding and decoding, the fine-grained image classification tasks in this period showed low efficiency and accuracy.

Thanks to the rise of deep learning and the application of neural networks to artificial intelligence, Techniques that utilize deep learning for precise classification tasks have become increasingly sophisticated.. The features obtained from the neural network have a stronger generalization ability than artificial features, this significantly improves the effectiveness of fine-grained picture categorization. To generate a keypoint section image, some works rely on the strong supervision and utilize various subjective annotation data, such as bounding boxes to obtain target information and its position, so as to classify images. However, since the process of strong supervision will lead to the imbalance of classification accuracy and efficiency due to labeling and rigid decoding process, researchers have proposed fine-grained image classification based on weak supervision. Data labeling in weakly supervised learning is incomplete, that is, in the training set, only a portion of the data are labeled, and the majority of the data are unlabeled, or the supervised learning of data is indirect. In short, weakly supervised learning covers a wide range, and the use of weak supervision can enhance the generalization capacity of classification make it more widely used so as to increase its application. In this article, the author will discuss and summarize the weakly supervised fine-grained categorization of images, divided into methods based on different attention mechanisms and methods based on different neural network structures.

Focusing on the above two frameworks, this paper provides a detailed exploration of the classification system, followed by an introduction to the experimental process. Based on the data collected by excellent methods and frameworks in recent years, the author will compare and analyze them and summarize the reasons. At the end of the article, the future developments of this task are discussed.

2. Fine-grained image classification based on weak supervision

2.1. Methods based on different attention mechanisms

The attention mechanism refers to the ability of the computer to focus on the part when there is a shortage of processing power, there are more crucial tasks that need to be completed in order to address the issue of huge information resource allocation. In the neural network In the learning of learning, the computer processing overload is often caused by the large amount of information stored in the model. Therefore, the introduction of the attention mechanism directs the computer's attention to the important parts of the model and minimize the focus on other facts. Eliminate useless information. It can be solved, and can improve the efficiency and accuracy of task processing.

2.1.1 Self-attention. When performing semantic translation and encoding, we sometimes expect to focus resources on the entire task for accuracy, but at this time. it is not feasible due to objective reasons such as hardware equipment, so we hope to focus on key parts of the information In fact, when performing classification tasks, the key region is often not independent, it is often associated with other discriminative regions in the surrounding environment, so when dealing with this discriminative region, it will also focus more on the strong correlation with it. This is what we commonly called self-attention mechanism [1]. The self-attention mechanism, which is an improvement over the attention mechanism, is better at recognizing internal associations of information or characteristics and depends less on external sources. This is aimed at the problem that the correlation cannot be established in the neural network. Identifying the correlation among fields is the core function of the self-attention mechanism.

2.1.2 Multi-head attention. As an advanced version of the self-attention mechanism, the mechanism for mult-head attention has been developed further. the multi-head attention mechanism [2] makes up for the lack of information capture ability of the self-attention mechanism. Compared with many mainstream methods, the efficiency of self-attention mechanism for information capture is still too low. For example, When it comes to the question of small amount of data task, since the self-attention mechanism is designed to capture important information and ignore useless information, its effect is not good. Later researchers introduced a multi-head attention mechanism. The multi-head attention mechanism uses several sets of different linear projections independently. Learned to transform the query key and value, and then sends several sets of transformed queries to the attention pool, and splices the output together in order to pass it again through the connection of layer learnsing, and finally produces and outputs. In general, the multi-head attention mechanism is an integrated system based on the single-head attention mechanism. It inherits the advantages of the attention mechanism and can have better generalization performance and flexibility.

2.1.3 Two level attention. The two-level attention [4] mainly focuses on the features of two different levels, which identified the local level as well as the object level.. It can also be considered as the label box's information as well as the local box's location in the role of supervision. In order to perform the task of fine-grained picture classification, this method strives not to rely on extra labels or information and instead solely uses category labels. Three different attention kinds are combined in two-level attention, which is the generation of potential image patches from the bottom up, the selection of relevant image patches at the object level from the top down to create particular objects, and the localization of discriminative components at the component level from the bottom up. To recover foreground objects and portions with strong characteristics, a particular DCNN is trained by merging these kinds of attention processes. It is simple to generalize the model and it does not need bounding boxes or component annotations. From the perspective of the task work process, the preprocessing stage is first to extract candidate images from the original image to eliminate distractions which is from background information. For the local image model, its purpose is to select the key part of the area from the complicated area. For the object set model, the output of a softmax layer is obtained after the convolutional network, and the average value is calculated for all regions as the final softmax layer. Finally, the object-level model and the feature-level model are combined as the final output after training among the various models.

2.1.4 FCN attention. FCN attention [3] is based on a fully convolutional attention localization network, which can adaptively select multi-task-driven attention regions. The FCN model mainly includes continuous convolutional layers, deconvolutional layers, and treaty structures. A sequential convolutional structure is proposed to detect the item's properties. It is additionally effective and could simultaneously identfy numerous object pieces and obtain information from different attention zones because it is based on the FCN architecture. In which various components may have various predetermined sizes. The network has a classification module and a local location module.

2.1.5 Multiple granularity CNN. The Multiple granularity CNN [5] structure is also a method that integrates the attention mechanism and the convolutional neural network. Due to the lack of text information in some specific situations, a multi-channel convolutional neural network and multi-head attention mechanism is proposed. The image classification model is input into a multi-channel convolutional neural network through different new feature combinations, and an attention mechanism is integrated to learn image features more fully.

2.2. Methods based on the different neural network structure

Neural networks can provide a simpler way to solve problems. The various input layers, hidden layers, and output layers produce a variety of neural network framework models to simulate biological nervous systems in different ways. In terms of fine-grained image classification, different neural network architectures can play a certain role in the efficiency and focus of classification.

2.2.1 CNNs and its related network structures. The basic components of the convolutional neural network, or CNN, that allow the feature extraction function are the convolutional layer and the pool sampling layer of the hidden layer. Convolutional layer is a multi-layer supervised learning neural network. By applying the algorithm of gradient descent to lower the loss function, changing the weight parameters in the network layer by layer, and repeatedly continuously training the network, the network model improves the accuracy of the network. When applied to fine-grained classification tasks and derived from various types, three examples will be given below.

2.2.2 RACNN. The Recurrent Attention Convolutional Neural Network network [6] recursively learns region-based characterization and discriminant regional attention in a way that reinforces each other though., and proposes a pairwise ordering loss function in a relatively new technology to optimize the attention proposal network, which enables APN to shift attention to fine-grained regions.

2.2.3 MAMC. Multi-Attention Multi-Class [7] contains multiple CNNs, each of which classifies at a given granularity and merges the features extracted from each granularity to produce the final result. The labels of subclasses under this model contain hierarchical information, and with these hierarchical information, It is possible to train a number of CNN models with various granularities, covering various regions of interest and all relevant discriminating features.

2.2.4 Subset feature learning [9]. Large-scale datasets are used to train in general-purpose CNNs, transfer learning, and clusters the same types of classes in fine-grained datasets into several subclasses.

3. Experiment and performance comparison

3.1. Common dataset

In the experimental analysis section, we first introduce the data set used in this experiment.

(1) CUB200-2011. This dataset is a fine-grained dataset proposed by California Institute of Technology in 2010, It includes 200 bird subclasses and 11788 bird photos in total. Each image includes information about the image class labeling, the bounding box of the bird in the image, the key part information about the bird, and the attribute information about the bird. There are 5994 pictures in the training dataset and 5794 in the test set.

(2) Sandford Cars. This dataset is a fine-grained classification task dataset for automobiles, which features a total of 16185 images of various car models. There are 8041 test sets and 8144 training sets, respectively.

3.2. Quantitative comparison

The performance of various representative fine-grained image classification methods on In Table 1, various datasets are being shown. Analyzing the data in the table, For the Standford dataset, the self-attention mechanism and the multi-head attention mechanism both perform very well, which benefits from the proposal and development of the attention mechanism. The mult-head attention mechanism will perform more accurately on the Standford dataset than the self-attention mechanism. This is because the operation mode of multi attention, which essentially runs multiple attention layers and combines their common output structure. What’s more, the attention layer may contain sub-attention, so the generalization ability and accuracy of the method will be relatively improved on different datasets.

In CUB200-2011, except for the performance of the two-level attention and subset feature learning network, most methods have achieved good accuracy, including the current popular methods such as OPAM, which achieves high accuracy in image classification by dividing the data set and discriminating the target area. The FNC attention framework also achieves good accuracy, mainly due to its use of neural networks to make trade-offs for attention mechanisms, enhancing the accuracy of attention mechanisms, and therefore improving the accuracy of classification. In terms of two-level attention, the accuracy rate of 69.7% is not ideal. On the whole, two-level attention is a good solution in how to use labels to detect and divide local target areas, but how to use algorithms to classify it is very limited, so two-level attention needs to be studied in depth. For the subset feature network, the problems are the type and capacity of the dataset. Specifically, the subset learning features need to be trained on a large-scale dataset, while CUB200-2011 is not enough to support its training ability.

Table 1. Performance comparison of various methods on different datasets.

Method	**Data set**	Accuracy
self-attention [1]	*Sandford Cars*	95.6%
multi-head attention [2]	*Sandford Cars*	96.4%
two level attention [3]	*CUB200-2011*	69.7%
FCN attention [4]	*Sandford Cars*	84.3%
Multiple granularity CNN [5]	*CUB200-2011*	81.7%
RACNN [6]	*CUB200-2011*	82.4%
MAMC [7]	*CUB200-2011*	86.5%
Subset feature learning [8]	*CUB200-2011*	77.5%
PDFR [9]	*CUB200-2011*	84.5%
OPAM [10]	*CUB200-2011*	85.8%
Bilinear CNN [11]	*CUB200-2011*	85.1%

4. Discussion

Summarizing the above experimental results and data, we can conclude that the fine-grained task for classifying images with scant supervision has achieved good performance under different model frameworks. In the attention-based model, both the self-attention mechanism and the multi-head attention mechanism show high accuracy in Standford Cars, while the multi-head attention mechanism can have different combinations of attention models in different tasks, so the generalization ability will also be improved. Two-level attention, multi-granularity CNN and FNC attention framework I think there should be further improvements in subsequent experiments to obtain theoretical results. When focusing on the different results obtained by different neural network structures, CNN Multiple has a better performance, while other network structures need further research, I think we can start from a combination of the two methods, you can combine different attention mechanisms with neural network frameworks, in order to obtain better accuracy and better generalization ability in fine-grained classification tasks to adapt to more scenarios.

5. Conclusion

Due to the advent of deep learning, the effect of fine-grained classification tasks has been greatly improved, which can deal with subtle differences between different kinds in different supervision methods or discriminative regions. How to apply the results obtained from the dataset to specific scenarios in real life or more different occupations will become an important goal in the next stage of fine-grained classification tasks. Fine-grained differentiation between more semantically similar categories will use more attention mechanisms and different neural network models will be combined.

Reference

[1] Chen P, Wei W 2022. J. Fine-Grained Image Classification based on Self-attention Feature Fusion and Graph-Propagation Phys. Conf. Ser. 2246 012067

[2] Ridnik, T. et al. ML-Decoder: Scalable and Versatile Classification Head. 2021. Proc of Int Conf. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021): 32-41.

[3] Xiao T, et al. 2014. Proc of Int Conf. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014): 842-850.

[4] Liu X, et al. 2016. J. Fully Convolutional Attention Localization Networks: Efficient Attention Localization for Fine-Grained Recognition. ArXiv abs/1603.06765.

[5] Wang D, et al. 2015. Proc of Int Conf. Multiple Granularity Descriptors for Fine-Grained Categorization. IEEE International Conference on Computer Vision (ICCV) (2015): 2399-2406.

[6] Fu J, et al. 2017. Proc of Int Conf. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition.IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 4476-4484.

[7] Sun, Ming et al. 2018. J. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition. ArXiv abs/1806.05372 (2018).

[8] Ge Z, et al. 2015. Proc of Int Conf. Subset feature learning for fine-grained category classification. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2015): 46-52.

[9] Zhang X, Xiong H, Zhou W, et al. 2016. Proc of Int Conf. Picking deep filter responses for fine-grained image recognition. IEEE Conference on Computer Vision and Pattern Recognition. 2016:1134-1142.

[10] Peng Y, He X, Zhao J. 2017. J. Object-part attention model for fine-grained image classification. IEEE Transactions on Image Processing, 2017, PP(99):1-1.

[11] T. Y. Lin, A. RoyChowdhury, S. Maji. 2015. Proc of Int Conf. Bilinear CNN models for fine-grained visual recognition. In Proceedings of IEEE International Conference on Computer Vision, IEEE, Santiago, Chile, pp.1449–1457, 2015.