Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241523

Research on deep learning method for fine-grained image classification

Zhenjiang Yu

School of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, 300457, China

yzj1023777@mail.tust.edu.cn

Abstract. The public has long been interested in computer vision research projects on image classification. Over the past 20 years, fine-grained image classification (FGIC) has advanced quickly because of the ongoing development of deep neural network models. The FDIC is based on the traditional image classification and further identifies the subtle differences between subclasses within the same category. Deep learning-based image categorization techniques are separated into two groups in this article: FGIC based on intensely supervised learning and weakly supervised learning. Briefly, it introduces the algorithms included in each category. Additionally, this article lists the performance of several methods on the well-known CUB-200 dataset and gives typical fine-grained picture datasets. By comparing several algorithms' outputs, it is determined that weakly supervised learning has the advantages of lower cost and higher accuracy than intensely supervised learning. Finally, the paper proposes a summary and a discussion of fine-grained images' potential future development prospects.

Keywords: Image classification, fine-grained, intensely supervised learning, weakly supervised learning, datasets.

1. Introduction

In computer vision, image classification is a common task: classifying input images into predefined categories. Specifically, the computer vision system must learn to recognize and distinguish different image categories, such as birds, dogs, cars, etc. The above classification is performed for coarse-grained image classification to identify other species with noticeable characteristic differences. However, as users increase, users will want to know more specific object or scene information in the image. Because coarse-grained image classification emphasizes visual characteristics and looks more than other factors, subtle features are often ignored. Therefore, users will face enormous challenges when using coarse-grained image classification technology to classify objects of similar appearance. For example, it can be not easy to distinguish different bird species and their subcategories in bird classification.

This challenging task drives the research and development of fine-grained image classification (FGIC). It is a more detailed and accurate response to the image classification task to satisfy the classification requirements in different fields. These include:

(1) Medical image analysis: FGIC can be used in medical image analysis to identify different types of lesions or tissue structures. This helps to improve the accuracy of medical diagnosis, facilitating personalized treatment and disease prediction [1].

DOI: 10.54254/2755-2721/45/20241523

(2) Industrial quality inspection: In the industrial field, FGIC can detect product defects and distinguish different types of parts to improve the production line's quality control and management efficiency [2].

Although FGIC has broad development prospects, it faces many difficulties in its construction and application. (1) Difficulties in data acquisition and labeling: The construction of fine-grained datasets requires many samples with accurate annotations. In some fine-grained classification tasks, accurate annotation requires corresponding professional knowledge and much time. For example, the critical points for annotating birds are the positions of wings and beaks. This dramatically increases the difficulty of data set construction. (2) Similarity between categories: Fine-grained images are affected by posture, perspective, lighting, occlusion, and background interference during the collection process, which will lead to data showing significant differences between classes and minor differences within classes.

Depending on whether additional artificial label information is used, this paper divides fine-grained images based on deep learning into two types: FGIC based on intensely supervised learning and weakly supervised learning. In this paper, two image classification methods are integrated, and the performance analysis results with the CUB-200 dataset are analyzed [3]. The paper aims to contribute a modest contribution to image classification.

2. FGIC based on intensely supervised learning

In intensely supervised learning, category labels and additional annotation information are used for training. These include: (1) Annotation information is usually in the form of bounding boxes or labeled boxes indicating discriminative feature areas in the image; (2) Mark specific vital points or critical areas of the object. For example, in bird classification, the positions of wings and beaks of different species of birds. The following are several classic models for intensely supervised FGIC.

2.1. Part-based Region with Convolutional Neural Network (Part R-CNN)

The Part R-CNN classification algorithm based on local areas accurately locates the target's position through additional annotated bounding boxes [4]. Then, the overall features and local features are extracted through this CNN. The local and global characteristics are then combined. Ultimately, the object's category is ascertained by classifying the fused characteristics using the support vector machine (SVM) model.

The algorithm improves local area detection and feature extraction through this CNN, improving classification results' accuracy. However, its region generation method produces many irrelevant regions, which causes the spending of the algorithm to decrease. Moreover, manual annotation information must be provided during training, so the practical application is minimal.

2.2. Pose Normalized CNN (PN-CNN)

The PN-CNN algorithm first uses the deformable part model algorithm to obtain predefined key position points [5]. Next, key points are used to align the pose of the image to reduce the impact of pose changes on performance. Then, use the CNN to extract different levels of features of local information. Finally, the elements of different parts are connected, and SVM is used for model training and classification. This algorithm further considers the interference caused by bird posture and reduces the negative impact caused by intra-class variance. However, it does not solve the shortcomings, such as manual annotation of information required for training the model and slow detection speed.

2.3. Mask-CNN

The Mask-CNN method solves the head, torso, and background classification issues using the fully convolutional networks (FCN) model [6]. The regional feature extraction network maps the minimum rectangular areas of different parts into feature vectors, and the pooling operation obtains the comprehensive feature vector of the entire image. Finally, classification is performed through the fully connected layer. The advantage of this method is that the features are selected first, improving the

DOI: 10.54254/2755-2721/45/20241523

network's computational efficiency. However, this method is inaccurate enough to position complex background images and manual annotation information is still required.

3. FGIC based on weakly supervised learning

None of the above methods can avoid using additional manual annotation information. Weakly supervised learning only uses the category labels and object annotation boxes of images during model training. It may accomplish classification accuracy on par with highly supervised classification models, even without further annotated critical point data. Because this method uses less annotation information, it reduces the burden and cost of annotation and has good application prospects for fine-grained classification tasks that are difficult to annotate.

3.1. Two-level (TL) attention

Only category labels are needed by the TL attention method to accomplish the classification objective; no extra annotation information is needed [7]. The algorithm first finds components with solid features, removes irrelevant information, and uses an SVM classifier for accurate classification. When category labels are the only information available, this algorithm can identify local areas, albeit local area accuracy is not very high.

3.2. FCN Attention

The fully convolutional attention localization network model, or FCN Attention model, is grounded on reinforcement learning and can make intelligent attention area selections for multi-task guiding [8]. It uses the local positioning and classification modules to realize positioning and feature extraction of multiple object components, improving the efficiency of the network and multi-component positioning capabilities.

3.3. Diversified visual attention

The diversified visual attention model determines the object category by generating attention regions at different time steps and uses a specific loss function to determine the features of multiple locations [9]. This model usually consists of four parts: attention region generation, feature extraction, diversified visual attention, and classification.

Multi-Attention CNN (MA-CNN): According to the MA-CNN paradigm, region-based fine-grained feature learning and regional placement can support one another [10].

Recurrent Attention CNN (RA-CNN): The RA-CNN model aims to enhance the attention and regional characteristics of the discriminant area by recursive learning [11]. Through the structure of three scale sub-networks, it finally integrates multiple scale networks to improve performance. The CNN is an end-to-end optimization that does not need information like bounding boxes to classify images.

3.4. Bilinear CNN

The bilinear CNN model is mainly used for feature fusion [12]. It extracts image features by building two parallel CNNs. It makes an external product of the extracted features to obtain bilinear vectors to compensate for the limitations of a single CNN losing too much information. It finally realizes the purpose of cooperation between the two CNNs.

4. Related image datasets

With the continuous development of FGIC, many researchers have emerged worldwide to build image datasets. Table 1 is the basic information of common datasets.

DOI: 10.54254/2755-2721/45/20241523

Table 1. Standard datasets and their basic information.

Dataset name	Year	Meta-class	Images	Categories
CUB200-2011 [3]	2011	Birds	11788	200
Stanford Dogs [13]	2011	Dogs	20580	120
FGVC Aircraft [14]	2013	Aircrafts	10000	100
Deep Fashion [15]	2016	Clothes	800000	1050
Veg200 [16]	2017	Vegetable	91117	200
RPC [17]	2019	Retail products	83739	200

Table 1 introduces the basic information of standard datasets, including dataset name, generation time, category, number of pictures, and subcategories. Due to the limited length of the article, except for the datasets in the above table, other datasets will not be introduced here. However, if the algorithm can achieve good performance on one of the datasets, it usually achieves good results on most of the datasets.

5. Experimental analysis and performance comparison

Since the CUB200-2011 bird dataset is the most classic, the above algorithm is used to evaluate this dataset, and the analysis results are shown in the table below [3].

Table 2. Evaluation results on the CUB200-2011 dataset [3]

Methods	Backbone	Accuracy
Part Based R-CNN [4]	Alex-Net	73.90%
Pose Normalized CNN [5]	Alex-Net	75.00%
Mask-CNN [6]	Alex-Net	78.60%
Two-level Attention [7]	VGG-16	77.90%
MA-CNN [10]	VGG-19	86.50%
RA-CNN [11]	VGG-19	85.30%
Bilinear CNN [12]	VGGD+VGGM	84.10%

The deep learning methods mentioned above are evaluated on the CUB200-2011 dataset, and the average accuracy is recorded in Table 2 [3]. At the same time, some methods are omitted because they do not report results on this dataset or have poor results. As can be seen from Table 2, FGIC based on weak supervision has achieved high accuracy. The reasons may be: (1) Data availability: In FGIC tasks, there is usually a large amount of image data. Weakly supervised methods entirely use these image data, allowing the model to learn from different areas in the image and associate these areas with image labels, improving image recognition accuracy. (2) Attention mechanism: Weakly supervised methods usually use an attention mechanism, allowing the model to focus on the areas or features that are most important for the classification task. (3) Label noise processing: Labels in fine-grained datasets may contain noise. Weakly supervised methods usually use techniques such as label smoothing and unsupervised denoising to deal with these problems, thus improving the stability of the model.

6. Conclusion

FGIC is a challenging task. This task requires overcoming challenges such as difficulty in dataset annotation, high category similarity, and extraction of local sensitive information. By deep learning technology, weakly supervised classification gains the ability to identify small features and improve classification performance through the combination of transfer learning and multi-modal information. FGIC is a field of potential and development prospects, providing crucial technical support for users to deeply understand and identify objects with minor differences.

DOI: 10.54254/2755-2721/45/20241523

After the research and analysis of this paper, the research hotspots in this field are summarized as follows: (1) Model interpretability: As deep learning models become more complex, the need for interpretability of model decisions also increases. Future developments will focus on how to make fine-grained classification models more interpretable to meet the rationality requirements for model decision-making in regulatory, medical, and legal fields. (2) Augmented reality and virtual reality: Augmented reality and virtual reality technologies have shown vast potential in FGIC, bringing unprecedented innovation and development to many fields. In games, FGIC can achieve a more realistic and interactive virtual experience, making the virtual world more vivid. Regarding skills training, this technology can be applied to the simulation environment to provide more specific and realistic training scenarios to help users better master various skills and knowledge. These technologies' continuous maturity and development are expected to become a pivotal factor in changing people's daily lives and work styles and bringing more innovation and convenience to society. (3) Data privacy and security: With increasing attention to data privacy and security, future development will involve processing and protecting user data in FGIC and achieving more accurate trade-offs. The development of this field will require comprehensive consideration of many factors, such as technology, law, and ethics, to achieve the best balance between data privacy and security.

References

[1] Xingru H, Shucheng H, Jun W, Shangchao Y, Yaqi W and Xin Y 2023 Lesion detection with fine-grained image categorization for myopic traction maculopathy (MTM) using optical coherence tomography *Journal Medical physics* **50** 9 5398-5409

[2] Jing Y, Jian D, Tianxiang L, Cheng H, Jianqiang L and Tielin S 2022 Tool wear monitoring in milling based on fine-grained image classification of machined surface images *Journal Sensors* **22** 21 8416

[3] Wah C, Branson S, Welinder P, Perona P and Belongie S 2011 The Caltech-UCSD Birds-200-2011 dataset

[4] Zhang N, Donahue J, Girshick R and Darrell 2014 Part-based R-CNNs for fine-grained category detection *In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland* 13 834-849 Springer International Publishing

[5] Branson S, Horn G V and Belongie S 2014 Bird species categorization using pose normalized deep convolutional nets *arXiv preprint arXiv:1406.2952*

[6] Wei XS, Xie CW and Wu J 2016 Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition *arXiv preprint arXiv:1605.06878*

[7] Xiao T, Xu Y, Yang K, Zhang J, Peng Y and 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* 842-850

[8] Liu X, Xia T, Wang J and Lin Y 2016 Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition *arXiv preprint arXiv:1603.06765* **1** 2 4

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

[10] Zheng H, Fu J, Mei T and Luo J 2017 Learning multi-attention convolutional neural network for fine-grained image recognition *In Proceedings of the IEEE international conference on computer vision* 5209-5217

[11] Fu J, Zheng H and 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* 4438-4446

[12] Lin TY, RoyChowdhury A and Maji S 2015 Bilinear CNN models for fine-grained visual recognition *In Proceedings of the IEEE international conference on computer vision* 1449-1457

DOI: 10.54254/2755-2721/45/20241523

[13] Khosla A, Jayadevaprakash N, Yao B and Li F F 2011 Novel dataset for fine-grained image categorization: Stanford dogs *In Proc. CVPR workshop on fine-grained visual categorization (FGVC)*

[14] Maji S, Kannala J, Rahtu E, Blaschko M and Vedaldi A 2013 Fine-grained visual classification of aircraft *arXiv preprint arXiv:1306.5151*

[15] Ziwei L, Ping L, Shi Q, Xiaogang W and Xiaoou T 2016 Deepfashion: Powering robust clothes recognition and retrieval with rich annotations *In Proceedings of the IEEE conference on computer vision and pattern recognition* 1096-1104

[16] Saihui H, Yushan F and Zilei W 2017 Vegfru: A domain-specific dataset for fine-grained visual categorization *In Proceedings of the IEEE International Conference on Computer Vision* 541-549

[17] Xiushen W, Quan C, Lei Y, Peng W and Lingqiao L 2022 RPC: a large-scale and fine-grained retail product checkout dataset