Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230916

Performacnes comparison of CNN-based models for fine-grained image classification

Yuxin Lu[1, 3, †], Yang Pan[2, †]

[1]School of International Education, Guangxi University of Science and Technology, Liuzhou, Guangxi, 545006, China
[2]School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, Hubei. 430205, China

[3]15152506759@stu.sqxy.edu.cn
[†]These authors contributed equally.

Abstract. Fine-grained visual classification is regarded as a more refined classification that can identify specific types of objects. It has been widely used in commodity sales, vehicle recognition and person recognition, etc., and has played a great value in many fields. For example, it requires an algorithm to identify different species of birds or dogs to facilitate more practical applications. This task is difficult since the objects has similar appearance and there exist obvious intra-class variance and limited inter-class differences, where different kinds of birds could share very similar appearance. Deep learning techniques has been applied to image recognition, natural language processing, and many other fields. Several approaches tackling the fine-grained classification problem are proposed. To further demonstrate the different designs of these solutions, in this paper, fine-grained identification methods are compared and analyzed, among which WS-DAN achieves better results and it is preferred to be an effective method, which is expected to be more widely used in this field.

Keywords: deep learning, image recognition, fine-grained, neural network.

1. Introduction

Fine-grained visual classification means the recognition that is finer than ordinary image recognition, not only to identify the type of animal but also to identify more specific breeds. It is more challenging compared to traditional tasks due to subtle intra-class object variations in nature [1,2]. The problem is to further identify the subcategories, such as different breeds of dog. Compared to conventional classification task, targets have shuttle granularity differences. As just mentioned, fine-grained tasks differ merely in local details. Therefore, seeking a method that could identify foreground targets and extract informative local information has become one of the most important tasks to be tackled in fine-grained image classification algorithms. The so-called supervision is simply the label. Existing models, could be categorized into two types based on different supervisions, including supervised model and weakly supervised model [3,4].

The fine-grained setting has been broadly utlized for commodity recognition in smart retail scenarios, vehicles in public security scenarios, person re-identification, vehicle type recognition, and detection

DOI: 10.54254/2755-2721/17/20230916

and identification of dangerous goods. Biodiversity monitoring, and many other fields have shown great practical value, especially in the industrial applications of smart new economy and industrial Internet.

Deep learning model has been researched for a long time. There two main messages from its early endeavours: First, artificial neural networks with many hidden layers have outstanding capacity for feature extraction, and its corresponding features have a more intrinsic description of the data. It is beneficial for a visualization analysis of the data. The second is that training deep models is difficult, which could be partially relieved by a "layer-by-layer initialization" strategy. Nowadays, well-known high-tech companies are focusing themselves on big data techniques. It is because the power of the big data is revealed in recent years. The conventional fields could be revolutionized and achieve novel state-of-the-art with limited resources, which indicates more precise estimation could be made with finite resources. Deep learning is broadly used, the main reason is that in these fields, a large amount of sample data can be collected, and features are not easy to extract directly. In this paper, several approaches tacking the fine-grained classification problems are compared.

2. Method

2.1. Overview

In this paper, five methods are selected to elaborate and analyze, and they each carry out image recognition and classification from different angles. These methods are WS-DAN, RA-CNN, MA-CNN, and B-CNN.

2.2. WS-DAN

The WS-DAN aims at solving fine-grained tasks leveraging the attention mechanism [5]. In this paper, the author pointed out that the key task in tackling fine-grained problem lies in extracting informative local descriptors of the object parts, rather than the image-level embeddings for the entire object. Because the clues to distinguish different classes lie in the shuttle differences of object parts. To achieve this, this work proposes several techniques. Firstly, Weakly Supervised Data Augmentation (WSDAN) is leveraged to show features of the object part, leveraging the generated attention map. By doing so, the feature of local parts could be achieved and further leveraged for tackling the fine-grained problem. Moreover, attention regularization loss and bilinear attention pooling are leveraged to learn weights in the model in a weakly supervised manner, which could potentially improve the classification accuracy validated by a series of experiments. After achieving the location of the object and corresponding parts by leveraging the attention mechanism, the attention is also leveraged to lead the data augmentation process. Under the guidance, the training data could be efficiently augmented, which could be concluded into two mechanisms including attention cropping and attention dropping. Finally, the attention map is also leveraged to identify the location of the entire target and rescale it to increase accuracy.

This work emphasizes that although data augmentation is examined as an effective method to augment the dataset, random cropping often leads to noise in the background, which could affect the efficiency and quality of the trained models. To relieve the negative effect of it, an attention mechanism should be introduced.

2.3. RA-CNN

Observing the previous deficiencies that manually defined regions may not be the optimal solution, together with the difficulties of extracting fine-grained features, this work seeks novel solutions for fine-grained classification [6]. The authors propose that region detection and fine-grained feature extraction could be united and further reinforce the features via humans. To tackle these challenges, this work proposed a recurrent attention CNN, named RA-CNN, for fine-grained classification. Different from object detection solutions, this work does not leverage bounding boxes as the pointer of objects. It takes advantage of the recursive learning strategies and the discriminative regions' attention for extracting feature representation of local object parts. The two mechanisms could be correlated and reinforce each other. The learning process of RA-CNN consists of three parts. Firstly, a multi-scale network is

DOI: 10.54254/2755-2721/17/20230916

leveraged. It could share different parameters on different scales so that the storage consumption is largely saved. Secondly, an enlargement technique is proposed to extract fine-grained features from enlarged object parts. Compared with conventional solutions, it could learn more fine-grained features. Finally, two loss functions, including the intra-scale soft maximum and an inter-scale pairwise ranking loss, are proposed for optimizing the learning.

2.4. MA-CNN

MA-CNN points out that the previous fine-grained solutions mainly focus on bounding box-based solutions, where discriminative object parts will be identified [7]. These works attempt to localize object parts for fine-grained learning. However, these works are usually computationally expensive, and the selection of a bounding box is subjective. Therefore, the authors propose the Multi-attention Convolutional Neural Network (MA-CNN). It gets rid of the generation bounding boxes or partial annotations. The model is made up of several subnetworks. It takes as input an image and generates multiple proposals of object parts. The learning consists of several steps. Firstly, a subnetwork called channel grouping is proposed. It could gather spatially adjacent parts together and extract correlated patterns using partial attention maps. These noticed part proposals could be extracted by fixed-size cropping and used for subsequent learning. Secondly, the partial classification network is constructed. It is leveraged to classify the image using features achieved from the identified parts. These features go through pooling operations. Finally, two loss functions are leveraged to optimize the learning. They could jointly supervise the learning of multiple tasks, including channel grouping and part classification. The loss functions allow the MA-CNN to learn more descriptive features of object parts.

2.5. B-CNN

B-CNN is short for bilinear CNNs, which are designed to map an image as pooled features [8]. Similar to previous works, the model uses part features for tackling the fine-grained classification problem. Observing that object parts could be regarded as a series of orderless texture representations, this work designs several bilinear models which could be trained in an end-to-end manner, for generalizing conventional image features. This design could introduce the visualization of the learned models by approximate inversion and domain-specific fine-tuning. Extensive experiments are conducted, demonstrating that these models could produce better performances on fine-grained classification tasks.

The B-CNN could solve many drawbacks lying in previous part-based representation learning models. Firstly, previous works models texture encoders as two features' outer product. It is a simple solution however, the non-linear relations between two features are ignored. These bilinear models, however, allow the second-order pooling representation of two identical features, which makes it even surpasses the performances of CNN models, validated on a series of datasets designed for fine-grained task, even though only image-level supervisions are leveraged.

2.6. MPN-COV

The matrix power normalized covariance pooling method (MPN-COV) is designed for large-scale image classification [9]. By making improvements on first-order pooling, classic neural network architectures are greatly improved. It shows that when training with high-dimensional features with a limited number of samples, a covariance estimator is leveraged to calculate the matrix square root. In this model, meta-layer is proposed with a directed graph structure with loop embedded. The meta-layer is made up of three layers with consecutive nonlinear structure. Its execution time is much faster than previous works and the performances are much better.

3. Result

3.1. Experimental setting

To compare these methods, four datasets designed for the fine-grained task is used, including CUB-200-2011, FGVC-Aircraft, Stanford Cars, and Stanford Dog. In CUB-200-2011 different birds should be

DOI: 10.54254/2755-2721/17/20230916

categorized into 200 subcategories, including about 6000 samples for training and around 6000 samples for testing. FGVC-Aircraft dataset has 100 classes of airplanes that requires an algorithm to identify. The Stanford Cars and Dogs datasets is collected for the detailed classification of cars and dogs. There are 196 and 120 subclasses respectively [10].

3.2. Comparison of numerical results

The aforementioned results are compared using the four fine-grained datasets. The performances are demonstrated in Table 1. WS-DAN model performs good on all the four datasets.

Table 1. Result comparison.

	Cub-200-2011	FGVC-Aircraft	Standford Cars	Standford Dogs
WS-DAN	89.4%	93.0%	94.5%	92.2%
RA-CNN	86.5%	88.4%	92.5%	87.3%
MA-CNN	86.5%	89.9%	92.8%	--
B-CNN	84.1%	84.1%	--	--
MPN-COV	88.7%	91.4%	93.3%	--

4. Discussion

From the perspective of the method itself, some algorithm improvements can be made. For the WS-DAN framework, some additional attention mechanisms can be introduced because it only uses spatial attention, and the category loss of speech information is constrained by such category loss. More feature enhancement methods could be constructed during its learning, such as geometric transformations, color transformations, and other applications.

The running speed of this WS-DAN can be improved, and some lightweight optimizations can be considered for this method. For example, more refined model design, weight quantization, weight sharing, computational acceleration, network decomposition (tensor decomposition), transformation of data, and so on.

The application prospects of this method are very promising, for example, it can be applied to the agricultural field to achieve the identification and prevention of pests and diseases. For the identification and control of pests and diseases, it is important to detect plant lesions quickly in the early stage, as late detection can make prevention and control difficult and meaningless. However, when early plants suffer from pests and diseases, the characteristic areas of the disease spots are usually small and less obvious, making it difficult to detect. Therefore, it is necessary to create accurate disease spot features that can identify plant pests and diseases to determine the situation of plant diseases, in order to timely prevent and to rescuer plants.

However, in practical applications, there are still many complex influencing factors involved, such as changes in lighting in different application scenarios, differences in background, and changes in imaging height, which may affect its feature recognition performance. These all require us to make multiple adjustments and improvements to this method to optimize recognition accuracy.

5. Conclusion

In this paper, the authors compared four different deep learning solutions towards fine-grained classification setting. In this setting, method is designed to identify the subcategories of the target. For example, different breeds of the dog or different types of the flower. The four compared models are WS-DAN, RA-CNN, MA-CNN, and B-CNN. These models are tested on a series of datasets embodies fine-grained settings. The performances demonstrate the superiority of the WS-DAN, which reveals the protentional effectiveness of the attention mechanism in fine-grained visual recognition problems. In the future, more research could be conducted to dig the usability of the attentional designs in neural network.

DOI: 10.54254/2755-2721/17/20230916

References

[1] Zhao, B., Feng, J., Wu, X., & Yan, S. (2017). A survey on deep learning-based fine-grained object classification and semantic segmentation. International Journal of Automation and Computing, 14(2), 119-135.

[2] Arslan, B., Memiş, S., Sönmez, E. B., & Batur, O. Z. (2021). Fine-grained food classification methods on the UEC food-100 database. IEEE Transactions on Artificial Intelligence, 3(2), 238-243.

[3] Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., & Wang, L. (2018). Learning to navigate for fine-grained classification. In Proceedings of the European conference on computer vision (ECCV), 420-435.

[4] Dubey, A., Gupta, O., Raskar, R., & Naik, N. (2018). Maximum-entropy fine grained classification. Advances in neural information processing systems, 31.

[5] Hu, T., Qi, H., Huang, Q., & Lu, Y. (2019). See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891.

[6] 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, 4438-4446.

[7] 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, 5209-5217.

[8] 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, 1449-1457.

[9] Li, P., Xie, J., Wang, Q., & Gao, Z. (2018). Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In Proceedings of the IEEE conference on computer vision and pattern recognition, 947-955.

[10] Rong, Y., Xu, W., Akata, Z., & Kasneci, E. (2021). Human attention in fine-grained classification. arXiv preprint arXiv:2111.01628.