

Combining Multi-Feature Regions for Fine-Grained Image Recognition

Sun Fayou

Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
Email: 314565679@qq.com

Hea Choon Ngo

Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
Email: heachoon@utem.edu.my

Yong Wee Sek

Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
Email: ywsek@utem.edu.my

Received: 04 October 2021; Accepted: 12 December 2021; Published: 08 February 2022

Abstract: Fine-grained visual classification (FGVC) is a challenging task due to the subtle discriminative features. Recently, RA-CNN selects a single feature region of the image, and recursively learns the discriminative features. However, RA-CNN abandons most of feature regions, which is not only inefficient but also ineffective. To address above issues, we design a novel fine-grained visual recognition model MRA-CNN, which associates multi-feature regions. To improve the feature representation, attention blocks are integrated into the backbone to reinforce significant features; To improve the classification accuracy, we design the feature scale dependent (FSD) algorithm to select the optimal outputs as the classifier inputs; To avoid missing features, we adopt the k-means algorithm to select multiple feature regions. We demonstrate the value of MRA-CNN by expensive experiments on three popular fine-grained benchmarks: CUB-200-2011, Cars196 and Aircrafts100 where we achieve state-of-the-art performance. Our codes can be found at <https://github.com/dlearning/MRA-CNN.git>.

Index Terms: MRA-CNN, reinforce significant features, feature scale dependent, multi-feature regions.

1. Introduction

With the process of the computer vision technology, the accurate objects classification is the center of attention [1,4]. Fine-grained visual categorization (FGVC) distinguishes objects in the same category (e.g., black footed albatross, laysan albatross, etc). Currently, FGVC has important value in species identification [13], the category of foliar diseases [12] and so on.

FGVC methods consist of 1) supervised learning methods with lots of part annotations, 2) Weakly Supervised Learning with only image-level labels.

Early works in FGVC rely on the manually annotated bounding box/ part annotations. A typical method is region proposal networks (RPN) to propose discriminative regions. Wei et al. [9] proposed Mask-CNN, which is first to apply CNN to FGVC. Zhang et al. [10] adopted PBR-CNN for FGVC, which learns object detection and part localizations. Due to part annotations, this method achieves better classification results. Branson et al. [11] propose pose normalized CNN, which utilizes pose alignment over part-level image patches, then locates the discriminative parts. Human-annotated methods demand huge cost, which limit the availability in practical use.

To overcome labor intensive part annotation issues, another method is weakly supervised regions proposals. These methods usually select discriminative parts and obtain high-level feature representations. Zhang et al. [5] adopted the method of target block detection and classification. In this method, the target detection is composed of foreground and background, then foreground and key regions are used for classification learning. Lin et al. [6] used bilinear CNN to achieve fine-grained classification by associating multiple feature channels. Fu et al. [7] proposed ra-cnn model, which adopts iteratively cropping and zooming the unique feature region. Zheng et al. [8] proposed ma-cnn model, which utilizes multiple feature regions and achieves better results in image fine-grained classification. Woo et al. [22] proposed convolutional block attention module (CBAM) block with spatial attention and channel attention to reinforce features

representations. Although ra-cnn and ma-cnn have been applied in ships detection and medical image classification, their performance needs to be improved.

To address above problems, we propose a novel FGVC model MRA-CNN. First, we integrate CBAM into backbone to reinforce high-level feature representations. Second, we design a novel algorithm FSD, which selects the best classification features. Finally, we utilize k-means to select multiple feature regions for enhancing global feature attention. In this paper, the main contents are as follows:

- (1). To obtain better feature representation, our model utilizes the attention mechanism.
- (2). To select optimal outputs as inputs of the classification network, we design the feature scale dependent (FSD) algorithm.
- (3). To utilize multiple feature regions, we use k-means to select key regions.

2. Method

2.1. RA-CNN

RA-CNN consists of three scale sub-networks, and each sub-network is composed of VGG19 and attention proposal sub-network (APN). The framework of RA-CNN consists of 1) VGG19 extract feature, 2) APN obtains the feature region, 3) located image region needs to be crop and zoom in, and 4) associating three sub-networks classification results for object recognition. The framework of RA-CNN is shown in Fig. 1.

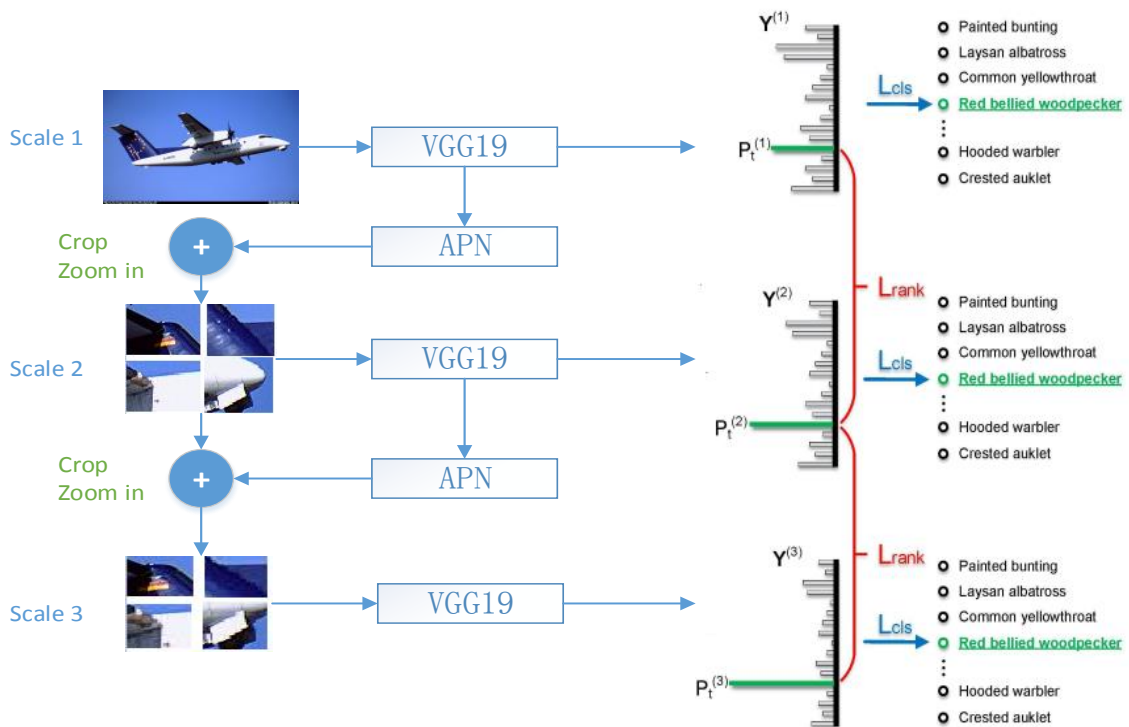


Fig.1. The framework of recurrent attention convolutional neural network (RA-CNN)

From Fig.1, it can be observed that vgg19 neglects the feature fusion; The input of each scale classification network comes from a fixed convolution layer; APN network only cuts and zooms in the image of a single located image region, which makes it impossible to utilize the global information of the image.

2.2. MRA-CNN

We utilize ra-cnn as the basic framework to design the MRA-CNN model. The workflow of the model consists of data preprocessing, backbone network (MRA-CNN) and classification network. The workflow of mra-cnn is shown in Fig 2.

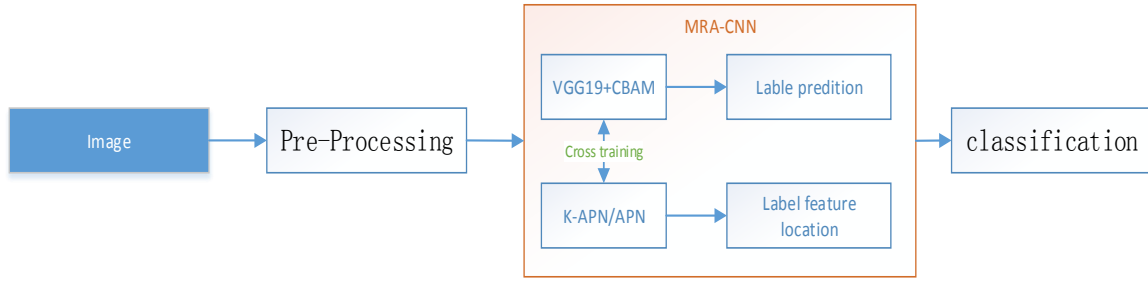


Fig.2. The workflow of our model

MRA-CNN also has three scale layers. We adjust the backbone from vgg19 to vgg19 ,CBAM and FSD. The APN network of first scale adopts k-apn network. The framework of MRA-CNN is shown in Fig. 3.

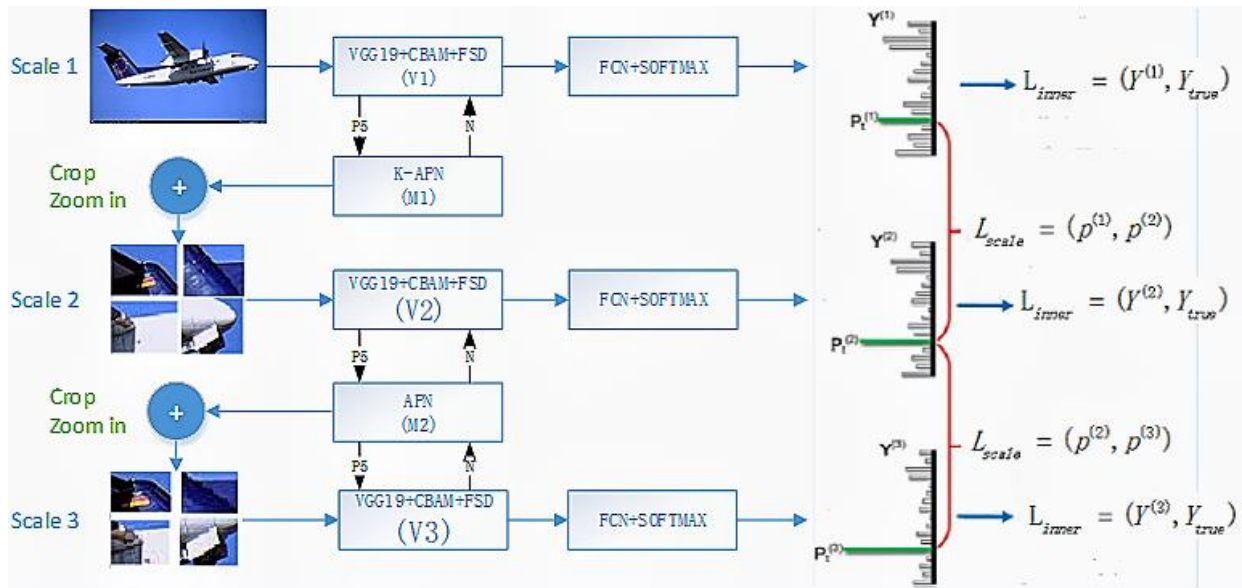


Fig.3. The framework of MRA-CNN. $p_t^{(i)}$ represents prediction probability. $Y^{(s)}$ represents prediction category. L_{inner} represents classification loss. L_{scale} represents inter-scale classification loss.

The workflow of MRA-CNN is as follows:

Step1: The image is input to scale1, then the features are extracted via V_1 .

Step2: The output P_5 of the 5th pool layer in V_1 is input k-apn to obtain the average size N of the k feature regions.

Step3: FSD algorithm associates V_1 outputs according to the size of N . Then the output feature is input to the classification network (FCN + softmax) to obtain the prediction label $Y^{(1)}$ of scale1.

Step4: The k-means algorithm is used to cluster multiple input features in $K\text{-APN}(M_1)$ to generate k feature regions.

Step5: According to the k feature regions of M_1 , k image patches are cropped and zoomed in as the inputs of the scale2.

Step6: After that, the k images patches are input to V_2 . We also utilize FSD to obtain feature representation, and associate three classification results to achieve the prediction label $Y^{(2)}$ of scale2.

Step7: Detailed k images patches are input to scale3, which repeats the processes of step6. This step outputs prediction label $Y^{(3)}$.

Step8: Finally, this model associates $Y^{(1)}, Y^{(2)}, Y^{(3)}$ to achieve classification.

2.3. Pre-Processing

As popular fine-grained benchmark datasets have different image sizes, we first adjust image size to $224 * 224$ pixels, then the pixel range is normalized from $[0, 255]$ to $[-1, 1]$. The pre-processed data can accelerate the model convergence and improve the robustness of the network. The normalization formula is:

$$X^* = \frac{2 \times (X - \min)}{\max - \min} - 1 = \frac{2X}{255} - 1 \quad (1)$$

where $\min=0$, $\max=255$, X represents the original pixel, X^* is the normalized pixel.

2.4. Backbone Network

VGG19 realizes the integration of network depth and performance, so it is used as the backbone network in r-cnn. However, VGG19 lacks the attention mechanisms. Inspired by bilinear CNN[6] and CBAM[22], we can reinforce feature representation via attention. Furthermore, if the size of feature map obtained by APN is small, fewer discriminative features availability for the classification network. To improve the classification ability of the network, the FSD algorithm is designed in the backbone to integrate the outputs of the convolution layer as the input of the classification network. The backbone is shown in Fig 4.

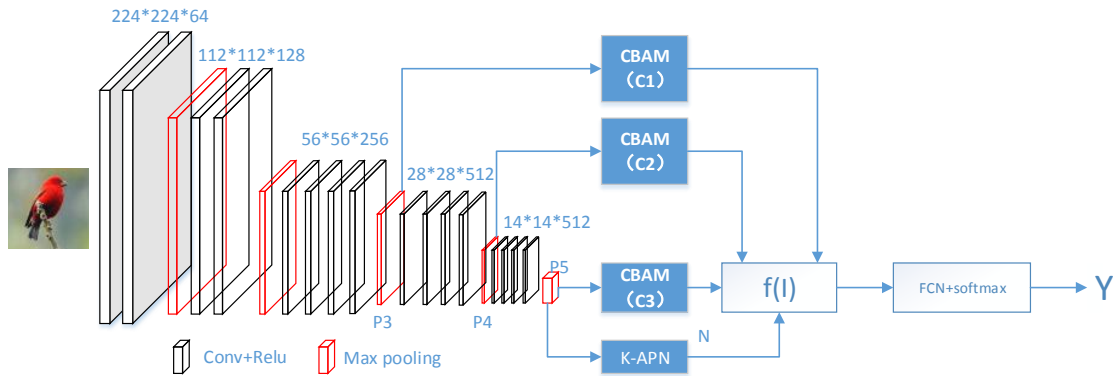


Fig. 4. The backbone of MRA-CNN. I represents the input image; P3, P4 and P5 represent max- pooling; $F(I)$ represents the FSD function. C1, C2 and C3 represent three CBAM blocks.

VGG19 extract the image features. P5 is the input of K-APN, which utilize k-means algorithm to extract k groups of feature regions, then selects the highest score in each group as the part region. Next, we calculate the average size N of k part regions via FSD algorithm (i.e., $f(I)$). The $f(I)$ integrates the features (i.e., C1, C2, C3) to achieve the best classification. The formula of $F(I)$ is as follows:

$$f(I) = \begin{cases} C_1 + C_2 + C_3 & 0 < N \leq 64 \\ C_2 + C_3 & 64 < N \leq 128 \\ C_3 & 128 < N \end{cases} \quad (2)$$

When $f(I)$ associates CBAM modules (i.e., C1, C2, C3), the interpolation algorithm is used to implement feature up-sampling to achieve the consistency of feature size. Meanwhile, classification network adopts fully convolutional network (FCN), and the formula is as follows:

$$Y = h(g(f(I))) \quad (3)$$

where $g(*)$ represents FCN convolution network and $h(*)$ represents softmax function.

Finally, we integrate the three scale prediction values to achieve object classification.

2.5. K-APN

RA-CNN gets a single feature area via APN network. Then, APN crops and zooms in the corresponding image area as the input of the next scale. In this way, it is easy to miss the subtle regions of the image. Inspired by MA-CNN, we utilize multiple feature regions for target location to reinforce the robustness of target recognition, so we propose K-APN network.

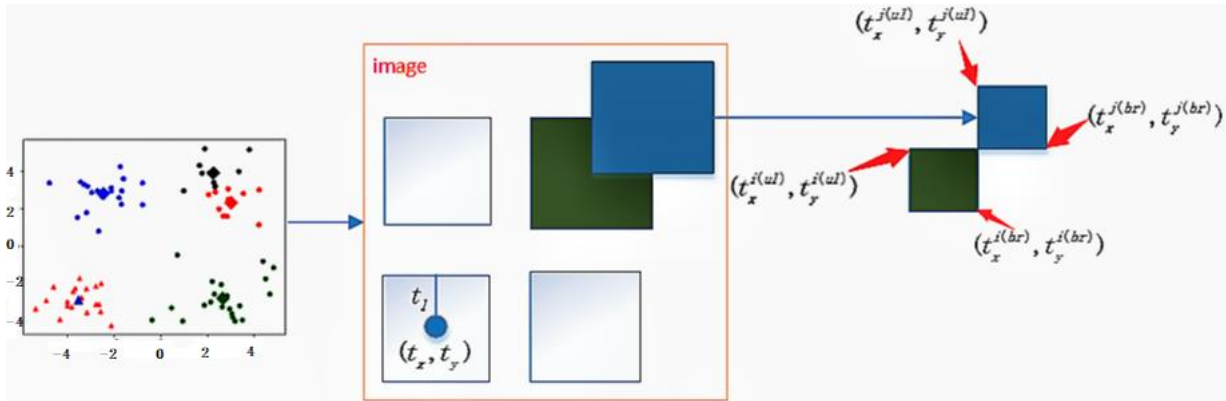
Due to the advantages of multi feature regions in image classification, this paper uses k-means algorithm to generate multi feature regions. As shown in Fig 5.



Fig. 5. The process framework of K-APN

We use k-means to obtain k feature regions, which are input into APN network to generate k parameter vectors $(t_{x_j}, t_{y_j}, t_{l_j} | 0 < j \leq K)$. After cropping and zooming in, corresponding image patches are input into the next scale.

K-Means adopts euclidean distance. As the clustering regions are mostly irregular shapes, there are some overlapping regions in the image. To address this problem, it is necessary to optimize the coordinates of the K feature regions to avoid overlapping. As shown in Fig. 6.


 Fig 6. K-APN feature map optimization process. The center coordinate of the square is (t_x, t_y) , t_l is half the length of the edge.

Supposing that the square parameter of the n -th feature region is $\{(t_x^{n(ul)}, t_y^{n(ul)}), (t_x^{n(br)}, t_y^{n(br)}), t_l^n\}$, where (ul) is the upper left corner, (br) is the lower right corner, t_l^n is the radius. The overlapping area of the feature area i and j is:

$$W = |t_x^{i(br)} - t_x^{j(ul)}| \quad (4)$$

$$H = |t_y^{j(br)} - t_y^{i(ul)}| \quad (5)$$

$$S = W \times H \quad (6)$$

where W and H represent the width and height of overlapping feature areas.

The steps of the feature region optimal algorithm are as follows:

Step 1: we calculate the overlapping region S of between feature region R and other regions by formula (6). If $S > 0$, R has overlapping regions.

Step 2: We identify the region with the largest overlapping area as P . According to the formulas (4) and (5), if $W \geq H$, the width of R and P will be reduced by $W/2$, otherwise the height will be reduced by $H/2$, respectively. Concurrently, there is no overlapping area between R and P .

Step 3: After that, we check if the revised R has overlapping areas. If there are overlapping areas, Repeating step 2.

Step 4: Selecting next feature region and repeating steps 1 to 3.

If K-APN network is used for each scale of MRA-CNN network, the feature regions of each scale will increase exponentially and increase the amount of calculation. Thus, K-APN is only used in the first scale.

2.6. Loss function

If MRA-CNN adopts intra-loss function and inter-loss function, it can not verify the advantages of K-APN. Thus, the loss function of MRA-CNN is:

$$Loss = \sum_{i=1}^3 L_{inner}(Y^{(s)}, Y_{true}) + \sum_{i=1}^2 L_{scale}(P_i^{(s)}, P_i^{(s+1)}) + L_{channel} \quad (7)$$

$$L_{channel} = \sum_{i=1}^N (\arg \min_j \|x_i - c_j\|_2^2) \quad (8)$$

where L_{inner} is the intra-loss function of APN. L_{scale} is the inter-loss function of APN. $L_{channel}$ is the clustering loss function of K-APN where N represents the number of elements, X represents elements, and C_j represents the center of class J.

As scale2 have K features, $Y^{(2)}$ is the mean value of the prediction probability of K features. For example, the probability p_i^2 of class i in $Y^{(2)}$ is calculated as:

$$P_i^2 = \frac{1}{K} (\sum_{j=1}^k P_{(i,j)}^2) \quad (9)$$

Scale3 is the same as scale2.

As $L_{channel}$ is added to the loss function, the rate of convergence can be accelerated. By using multiply feature regions, the accuracy of image recognition is finally improved.

3. Experiments

We conducted experiments on three widely used fine-grained datasets (i.e., cub-200-2011[2], Stanford car[3], fgvc aircraft[4]). The fgvc-aircraft dataset has 10,0200 samples and is divided into 102 categories. Each model of aircraft contains different subclasses, such as Airbus A300-200 has A300-200T, A300-200F and other models, as shown in Fig. 7.



Fig.7. A300-200 aircraft photos

The ratio of training data, test data and validation data is 6:2:2. The hardware and software configurations used in the experiment are listed in Table 1.

Table 1. The main configuration

OS	Win10 pro 64bit
GPU	Titan xp
Cuda	7.0
Tensorflow	1.13.1
Keras	2.1
Anaconda	4.8.3

3.1. Experiment 1: K value

In this paper, accuracy is used as the evaluation index and the formula is:

$$\text{accuracy} = (\text{true positive} + \text{false positive}) / \text{all data} \quad (10)$$

Using Elbow method to select the optimal k-value for K-Means clustering, as shown in Fig. 8.

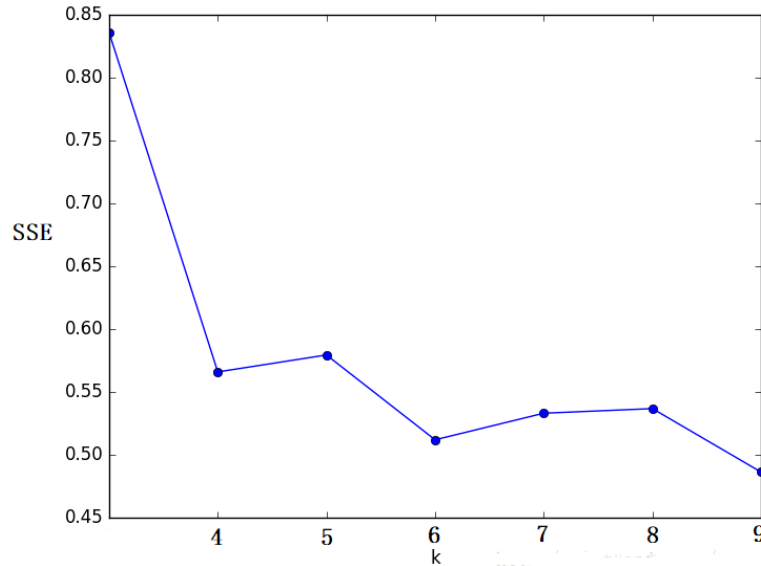


Fig.8. K-value curve

From Fig. 8, it can be observed that the optimal $k=5$. In order to verify the validity of K-cluster, a comparative experiment is designed.

Table 2. Comparison results of different K values

K value	3	4	5	6	7
accuracy	89.02%	89.16%	90.31%	89.27%	88.63%

From Table 2, we know that with the increase of K value, the classification accuracy of the model increases first and then decreases, indicating that the model can extract more feature representations from multiple feature regions, and the accuracy of the model is the best at the optimal K value.

3.2. Experiment 2: Ablations and Performance

This experiment mainly compares the accuracy of multiple fgvc models in three datasets. The experimental results are shown in Table 3.

Table 3. Comparison in terms of classification accuracy on the CUB-200-2011, Stanford-Car, and FGVC-Aircraft datasets.

Method	Accuracy/%		
	CUB-200-2011	Stanford-Car	FGVC-Aircraft
VGG19[27]	78.17	85.73	81.85
RA-CNN[7]	84.41	92.5	87.2
MA-CNN[8]	86.5	92.8	89.9
MRA-CNN (k-means)	85.3	92.6	88.3
MRA-CNN (K-means+CBAM)	87.1	93.1	90.27
MRA-CNN (K-means+CBAM+FSD)	87.4	93.3	90.42

From Table 3, it can be observed that the accuracy of MRA-CNN improve by 2.99% compared with ra-cnn in cub-200-2011. Meanwhile, the accuracy of MRA-CNN gets improvement in three datasets compared with ma-cnn model, which uses multi-feature regions.

From Fig. 9, it can be observed that the amount of calculation of mra-cnn increases slightly, and it takes about 0.238s to recognize one picture. The result verifies that our model can process the image in time.

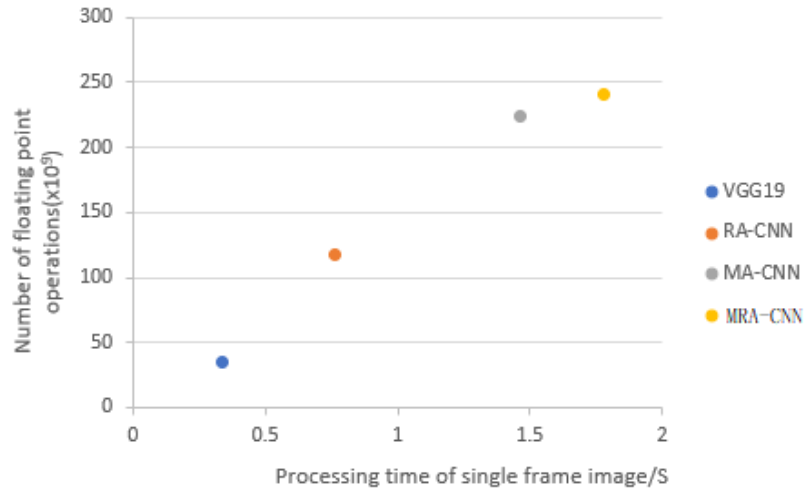


Fig.9. Algorithm performance analysis chart

3.3 Experiment 3: Visualization with Attention Map

In the k-apn part, multiple feature regions can be generated. This experiment shows the feature region outputs of scale3 in different epoch. As shown in Fig 10, the region of interest of MRA-CNN is shown when the epoch is 200, 2000 and 20000, respectively.

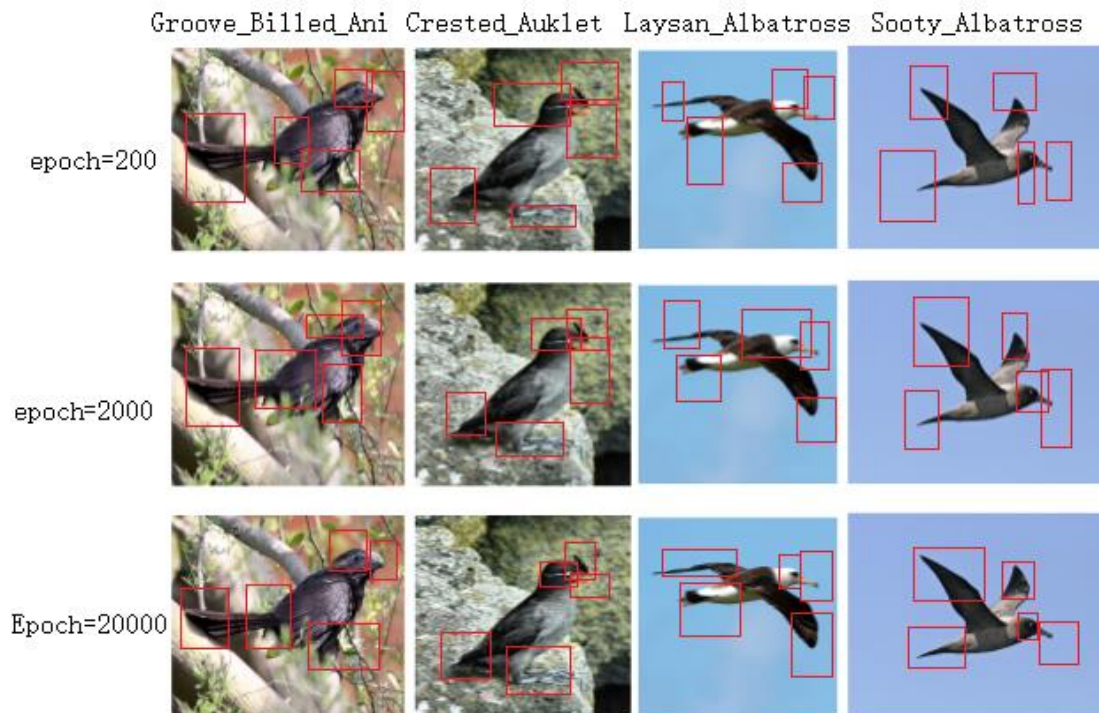


Fig.10. Transformation of feature region of interest

From Fig 10, we know that at the beginning of model training, feature regions are not obvious, which is obviously affected by background noise. When the model is trained to epoch is 20000, the regions of image are non-overlapping, and the discriminative regions of image are located.

4. Conclusion

In this paper, we propose a novel fine-grained visual classification model MRA-CNN. We adopt attention mechanism, feature scale dependent algorithm and k-means to select the optimal feature representation. Extensive experiments in three benchmark datasets demonstrate that our model is able to outperform not only ra-cnn but also ma-cnn. In the future, we will improve the proposed mra-cnn in the following directions: 1) learning region proposing, i.e., selecting better feature extraction network instead of VGG19, 2) less computational cost, 3) applying MRA-CNN to image segmentation and other fields.

With the state-of-the-art results of MRA-CNN, it can be observed that region proposal models are useful for fine-grained tasks and the MRA-CNN is a reference for future works.

Abbreviations

The table below lists several abbreviations used in this paper.

Table 4. Abbreviations

Abbreviation	Description
FGVC	Fine-grained visual classification
RA-CNN	Recurrent Attention Convolutional Neural Network
MA-CNN	Multi-attention Convolutional Neural Network
APN	Attention proposal sub-network
CBAM	Convolutional Block Attention Module
FCN	Fully Convolutional Networks
MRA-CNN	Our model
FSD	feature scale dependent

Acknowledgment

This research is supported by Universiti Teknikal Malaysia Melaka (UTeM).

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

- [1] Chang Pengfei, Duan Yunlong. "Application of Faster R-CNN Model in Aircraft Target Detection in Remote Sensing Image [J]." *Radio Engineering*, 2019, 49(10): 925-929.
- [2] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD birds-200 (2011 dataset) [R]. *Computation & Neural Systems Technical Report, CNS-TR-2011-001*, California Institute of Technology, Pasadena, CA, 2011
- [3] Krause J, Stark M, Jia D, et al. 3D object representations for fine-grained categorization [C] // *IEEE International Conference on Computer Vision Workshops*, 2013: 554-561
- [4] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft [J]. *arXiv Preprint*, 2013, arXiv: 1306.5151
- [5] Zhang N, Donahue J, Girshick R., & Darrell, T. "Part-Based R-CNNs for Fine-Grained Category Detection," In *European Conference on Computer Vision*, 2014, pp. 834-849.
- [6] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhansu Maji. "Bilinear CNN models for fine-grained visual recognition," In *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449-1457.

- [7] Fu J, Zheng H, Mei T. "Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3. DOI: 10.1109/CVPR.2017.476.
- [8] Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In: Int. Conf. on Computer Vision(2017).
- [9] XIUSHEN W, CHENWEI X, JIANXIN W, et al. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018, 76.
- [10] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection[C]//European Conference on Computer Vision. Springer International Publishing, 2014: 834-849.
- [11] BRANSON S, VAN HORN G, BELONGIE S, et al. Bird species categorization using pose normalized deep convolutional nets[J]. 2014.
- [12] Hu Z W, Yang H, Huang J., & Xie, Q. "Fine-grained tomato disease recognition based on attention residual mechanism," Journal of South China Agricultural University, 2019, 40(6), 124-132.
- [13] Huo Y H, Xu Z J, "Photoelectric ship target identification method based on improved RA-CNN," Journal of Shanghai Maritime University, 2019, (3), 38-43.
- [14] Russakovsky, O., Deng, J., Su, H., et al. "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, 2015, 115(3), 211-252.
- [15] Yang, F., Choi, W., Lin, Y. "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, 2016, pp. 2129-2137.
- [16] Xiong C Z, Jiang J. "Research on fine-grained classification algorithm of multi-scale regional features," Journal of Zhengzhou University (Natural Science Edition), 2019, 51(3), 55-60.
- [17] Qiao D, Liu G, Yang Z J, et al. "Ship target recognition based on transfer learning," Application Research of Computers, 2020, 37(1): 324-325+328.
- [18] Zhang, L., Gan, C., Hu, Y. "Ship detection algorithm research on high resolution optical remote sensing image," Computer Engineering and Applications, 2017, 53(9), 184-189.
- [19] Zhang, Z. Y., Jiao, S. H. "Infrared ship target detection method based on multiple feature fusion," Infrared and Laser Engineering, 2015, 44(1), 29-34.
- [20] Liu, X., Song, Y. "Classification of ship based on multi feature fusion," *Ship Science and Technology*, 2016, 38(14), pp. 88-90.
- [21] Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. "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, 2015, pp. 842-850.
- [22] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In The European Conference on Computer Vision (ECCV), September 2018.
- [23] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. TIP, 26(6):2868-2881, 2017.
- [24] Zhang, N., Donahue, J., Girshick, R., & Darrell, T. "Part-based R-CNNs for fine-grained category detection," In European Conference on Computer Vision, 2014, pp. 834-849.
- [25] Zhao, B., Wu, X., Feng, J., Peng, Q., & Yan, S. "Diversified visual attention networks for fine-grained object classification," IEEE Transactions on Multimedia, 2017, 19(6), 1245-1256.
- [26] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In ICCV, pages 5209-5217. 2017. 1, 2, 3, 6, 7.
- [27] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. abs/1409.1556.

Authors' Profiles



SUN FAYOU is currently pursuing Ph.D. degree in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka. His research interests include computer vision, generative adversarial networks and information network security.



HEA CHOON NGO is a senior lecturer at the Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). He received his Bachelor's degree in Computer Science (Software Development) from the Universiti Teknikal Malaysia Melaka (UTeM) in 2004, a Master's degree in Information Technology from University of New South Wales (UNSW, Sydney) in 2007 and a PhD in Computer Science from Universiti Sains Malaysia (USM) in 2016. His research interests involve computational intelligence, data science and analytics, planning and scheduling, optimization, health informatics and intelligent systems. He is currently a faculty member of the Faculty of Information and Communication Technology of the Universiti Teknikal Malaysia Melaka (UTeM). He is also

a member of the Computational Intelligence and Technologies Lab under the Centre for Advanced Computing Technology, UTeM.



YONG WEE SEK is a senior lecturer at the Department of Intelligent Computing and Analytics, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). He completed his PhD in Business Information System in 2017 from RMIT University Melbourne, Australia. He received his Bachelor degree of Statistics at the Unversiti Kebangsaan Malaysia (UKM) and Master degree in Information Technology at the Universiti Putra Malaysia (UPM). His research interests involve operation research, information systems, web based and multimedia learning and mathematics. He is currently a member of the Computational Intelligence and Technologies Lab under the Centre for Advanced Computing Technology, UTeM.

How to cite this paper: Sun Fayou, Hea Choon Ngo, Yong Wee Sek, " Combining Multi-Feature Regions for Fine-Grained Image Recognition", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.14, No.1, pp. 15-25, 2022.DOI: 10.5815/ijigsp.2022.01.02