Journal of Computer and Communications, 2021, 9, 10-28
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Asian Food Image Classification Based on Deep Learning

Bing Xu, Xiaopei He, Zhijian Qu*

Department of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: 275271847@qq.com, 1305838342@qq.com, *zhijianqu@sdut.edu.cn

How to cite this paper: Xu, B., He, X.P. and Qu, Z.J. (2021) Asian Food Image Classification Based on Deep Learning. *Journal of Computer and Communications*, **9**, 10-28.
https://doi.org/10.4236/jcc.2021.93002

Received: February 7, 2021
Accepted: February 28, 2021
Published: March 3, 2021

Abstract

To improve Asian food image classification accuracy, a method that combined Convolutional Block Attention Module (CBAM) with the Mobile NetV2, VGG16, and ResNet50 was proposed for Asian food image classification. Additionally, we proposed to use a mixed data enhancement algorithm (Mixup) to have a smoother discrimination ability. The effects of introducing the attention mechanism (CBAM) and using the mixed data enhancement algorithm (Mixup) were shown respectively through experimental comparison. The combination of these two and the final test set Top-1 accuracy rate reached 87.33%. Moreover, the information emphasized by CBAM was reflected through the visualization of the heat map. The results confirmed the classification method's effectiveness and provided new ideas that improved Asian food image classification accuracy.

Keywords

Asian Food, Image Classification, Convolutional Neural Network, Attention Mechanism, Data Enhancement

1. Introduction

A good diet provides humans with nutrients needed by the body hence making it a basis for human survival. With the continuous improvement of living standards, people have gradually begun to pay attention to the nutritional balance in their daily diet. People have started using computer vision technology to classify and recognize food images that provide a newly fast and low-cost method for analyzing food composition and nutritional components. Therefore, food image classification technology has gradually become a research hotspot in the field of computer vision.

The commonly used food image classification methods include the use of traditional machine learning methods and deep learning methods. Traditional ma-

chine learning methods solve food image classification problem by extracting food image features manually and designing classifiers. For instance, Yuji Matsuda *et al.* proposed a technique that involved dividing the entire food picture into different candidate regions, analyzing each candidate region, and identifying multiple foods simultaneously. It was proved that the proposed method was effective for the recognition of multiple food images [1]. Lukas Bossard *et al.* proposed a method that includes using the random forest for food image recognition. Their method outperformed alternative classification methods, including SVM classification on Improved Fisher Vectors and existing discriminative part-mining algorithms. On the challenging mit-Indoor dataset, their method compared nicely to other s-o-a component-based classification methods [2]. On the other hand, Marc Bolaños and others proposed a food image recognition algorithm. The core of this algorithm is first to analyze the food feature map of the input food picture, and then use it to predict the kind of the input food by looking for the most similar food features. They proved that, compared to the most similar problem nowadays—object localization, was able to obtain high precision and reasonable recall levels with only a few bounding boxes. Furthermore, they also showed that it was applicable to both conventional and egocentric images [3]. Further, Shulin Yang *et al.* proposed a method of performing pairwise statistics on the feature relationship between different components of food and analyzing the statistical information to come up with a classifier for food identification. Their experiments showed that the proposed representation was significantly more accurate at identifying food than existing methods [4].

But the traditional machine learning classification technology relies on manual feature extraction and classifiers selection. A variety of factors restrict the methods of manually extracting features. It is usually difficult to accurately express the real meaning of the picture, which results in low classification accuracy. This method is based on deep learning. We automatically learn food features through deep neural networks. We can closely associate the learned features with the classifier, which solves many shortcomings caused by manual feature extraction and design classifiers. Because of the aforementioned factors, food image classification based on deep learning methods has been receiving increasing attention.

Some of the most common neural networks in food image classification tasks include MobileNet, VGG, ResNet, etc. Fu Z. *et al.* introduced a 1000-category food data set ChinFood 1000 and proposed a simple and effective baseline method that involved using the ResNet model to conduct research on the ChinFood 1000 data set. The base-line approach was evaluated on three most widely used food data sets and achieved the best performance on all of them. And this approach was also applied to the ChinFood 1000 dataset with a promising accuracy [5]. On the other hand, M. Taskiran and others used the Food 101 data set to train models such as MobileNet, VGG, and ResNet and proposed a method of comparing correlation coefficients within categories to solve the problem of in-

tra-class differences in food image classification. It was proved that GoogleNet had the highest validation accuracy value with the lowest number of epochs [6]. K. Sukvichai *et al.* used the MobileNet model as a food image classifier. They proposed a food image classification model that transplanted the MobileNet network to the Raspberry Pi, to calculate the nutritional content of food. Their network in Raspberry Pi 3 produced good prediction accuracy but slow speed. And they introduced PeachPy to speed up the network and it could run at 3.3 seconds per food image [7].

There is a variety of Asian foods, and these foods have different characteristics. Therefore, this makes it challenging to classify Asian food images manually. It then becomes essential to come up with a method for Asian food image classification. In this article, we used an algorithm that was based on deep learning with three convolutional neural networks that included MobileNetV2 [8], VGG16 [9], and ResNet50 [10] as the baseline network. Further, we used CBAM [11] (Convolutional Block Attention Module) to improve the baseline network and Mixup [12] data enhancement algorithm to expand the training set. Our method verified the performance improvement effect of the CBAM attention mechanism and the Mixup data enhancement algorithm on Asian food image classification.

2. Neural Networks

To realize the classification method of deep learning for Asian food pictures, we selected MobileNetV2, VGG16, and ResNet50 as the experimental baseline network.

2.1. MobileNetV2 Network

MobileNetV2 is a lightweight deep neural network. Its core is adding 1×1 convolution before deep separation and convolution, enabling deep separation convolution to perform feature extraction on higher-dimensional channels. Depth separable convolution integrates ordinary convolution into layer-wise convolution and pointwise convolution. Moreover, it uses Batch Normalization (BN) to prevent overfitting and ReLu function as the activation function. Additionally, to curb the problem of feature loss during training, the ReLu function of the third layer of point-to-point convolution is replaced with a linear activation function.

To efficiently play the role of depth separable convolution, MobileNetV2 builds an inverted residual block structure. This structure draws on the residual structure of ResNet, and it is used for training only when the step size is 1, and the network is deep. The block structure of MobileNetV2 is shown in **Figure 1**.

2.2. VGG16 Network

The Visual Geometry Group proposed VGG model in 2014. The most outstanding feature of the VGG models is its simplicity. Its hierarchical structure

includes convolutional layers, pooling layers, and fully connected layers. The convolutional layers are for extracting features at different locations in an image. On the other hand, the pooling layers are used to reduce the dimensionality of the features extracted by the convolution kernels. The fully connected layers are equivalent to the classifier in machine learning since they classify the extracted features. The VGG16 model consists of 13 convolutional layers and 3 fully connected layers, and its input image size is 224 × 224. The initial convolution kernel size is 3 × 3, and the pooling layers are represented by 2 × 2 max-pooling. The structure diagram of VGG16 is shown as in Figure 2.

2.3. ResNet50 Network

The ResNet model was proposed by Kaiming and colleagues in 2015. The ResNet model solves the problems of gradient explosion and network convergence, which are slowly caused by the network's excessive depth through the identity mapping of the residual block. It is more practical in deep neural networks than the VGG models. The ResNet residual block is shown in Figure 3.

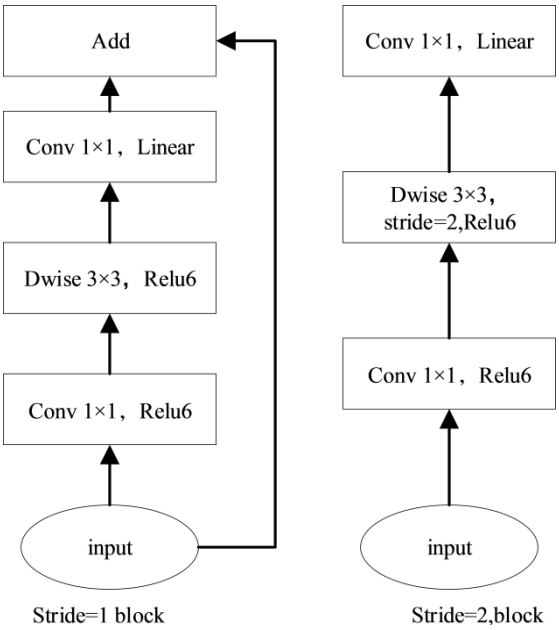

Figure 1. Block structure of MobileNetV2 [8].

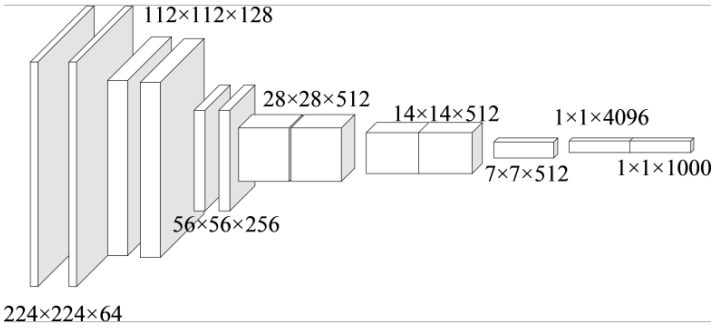

Figure 2. Structure of VGG16 [9].

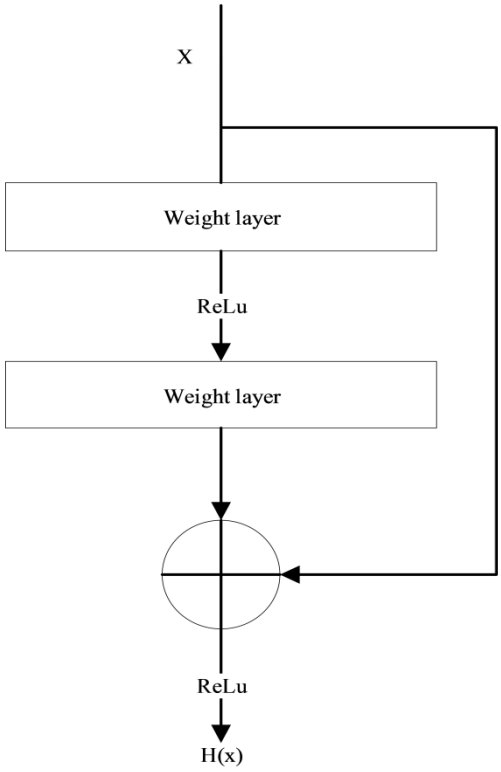

Figure 3. Block structure of ResNet50 [10].

Assuming that the output of the residual network is $H(x)$ and the output of the intermediate convolutional layer is $F(x)$, the general models make $F(x)$ approximate to $H(x)$. But the residual structure uses $H(x)-x$ to learn $F(x)$. Due to the existence of identity mapping, the residual network simplifies the learning goal of the models, thus reducing the learning difficulty, and improving the classification effect.

3. Methodology

3.1. CBAM Attention Mechanism

The attention mechanism in human vision is a signal processing mechanism. People scan the global image with the naked eye to get the target of attention, and then put more attention on the target to obtain more detailed information to reduce global information's attention. Through the attention mechanism, the speed and accuracy of human visual information processing are improved. The attention mechanism in deep learning draws on the attention mechanism in human vision. It adjusts and adapts to the learned features by changing the weight, which improves the accuracy of image classification.

We introduced the convolution block attention mechanism (CBAM) to improve the convolution models. CBAM includes channel attention and spatial attention. The channel attention learns the content of the picture, the structure is shown in **Figure 4**.

In the channel attention mechanism, C feature vectors $F \in R^{H*W*C}$ (R

represents the dimension of the feature vector, H represents the height of the feature vector space direction, and W represents the width of the feature vector space direction) are compressed into C real numbers of receptive fields $F_{max} \in R^{1*1*C}$ and $F_{avg} \in R^{1*1*C}$ by average-pooling and max-pooling, and then these real numbers generate the final weight parameter $M_c \in R^{1*1*C}$ through a multilayer perceptron. To obtain the feature vector F', the next is multiplied by the weight parameter M_c to the feature vector $F \in R^{H*W*C}$. The weight parameter of the channel attention module can be expressed by Equations (1).

$$\begin{aligned} M_c &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}\right)\right) + W_1\left(W_0\left(F_{\max}\right)\right)\right) \end{aligned} \tag{1}$$

Among them, σ represents the Sigmoid activation function, W_0 represents the weight of the fully connected layers, and W_1 represents the weight of the output layers.

Spatial attention learns the location of the input image, the structure is shown in Figure 5.

The spatial attention mechanism compresses the input feature map $F \in R^{H*W*C}$ into $F_{max} \in R^{H*W*1}$ and $F_{avg} \in R^{H*W*1}$ through global max-pooling and global average-pooling, and obtains a feature map with 2 channels by concat stitching, then it is compressed by a 7 × 7 convolution kernel, and the dimension is reduced to a feature map $F \in R^{H*W*1}$ with 1 channel. The weight parameter $M_s \in R^{H*W*1}$ is generated through the Sigmoid activation function. Finally, the final feature is obtained by multiplying the weight parameter with the input feature diagram F of the module. The weight parameter of the spatial attention module can be expressed by Equations (2).

$$\begin{aligned} M_s &= \sigma\left(f^{7\times7}\left(\left[AvgPool\left(F'\right); MaxPool\left(F'\right)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F'_{avg}; F'_{max}\right]\right)\right) \end{aligned} \tag{2}$$

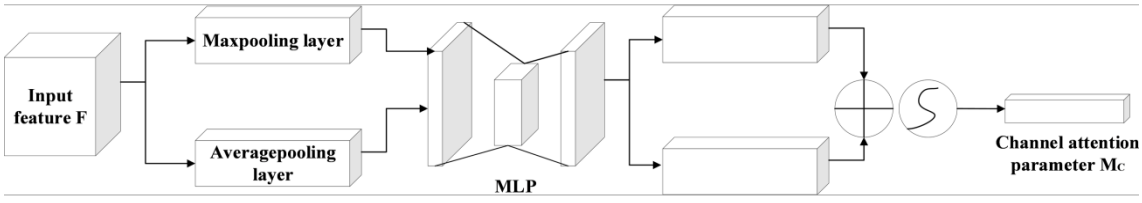

Figure 4. Channel attention structure of CBAM [11].

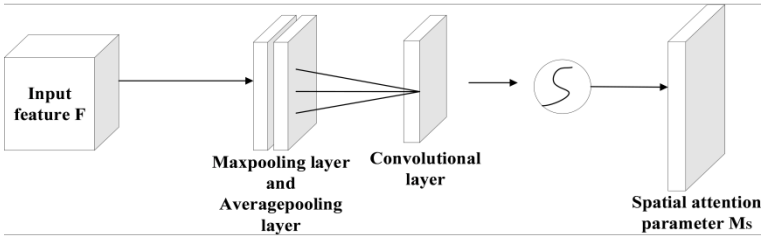

Figure 5. Spatial attention structure of CBAM [11].

The combination of spatial attention and channel attention enables the neural network model to locate quickly and focuses on the image's local key information for better adaptive effects. Moreover, the channel attention mechanism is implemented by MLP, and the pooling layers don't introduce more parameters, which significantly reduce the amount of CBAM parameters, and thus CBAM is a lightweight module.

3.2. CBAM Combination Method

Because the structure of CBAM includes channel attention and spatial attention, to maximize the effect of CBAM, we analyzed three different combinations of channel attention and spatial attention. First, the channel attention and spatial attention were connected in series. The input feature first paid attention to the feature content through the channel attention mechanism and then paid attention to the feature location through the spatial attention mechanism. We called this structure "channel before space". The structure is shown in Figure 6.

Secondly, the channel attention and spatial attention were connected in series. The input feature located the feature position through the spatial attention mechanism and then focused on the channel attention mechanism's feature content. We called this structure "space before channel". The structure is shown in Figure 7.

Finally, the channel attention and spatial attention were connected in parallel. The input features went through the channel attention mechanism and the spatial attention mechanism, respectively. It could pay attention to the feature content, and feature location, respectively, and then the structure merged the newly output features generated by these two mechanisms. We called this structure "parallel structure". The structure is shown in Figure 8.

3.3. CBAM-Based Network Model

Firstly, we introduced CBAM to MobileNetV2. Since MobileNetV2 is a lightweight network, and the inverse residual structure can simplify the model's learning goal and reduce the difficulty of training, and we added the CBAM attention mechanism after the last convolutional layer in the block with a step size of 1.

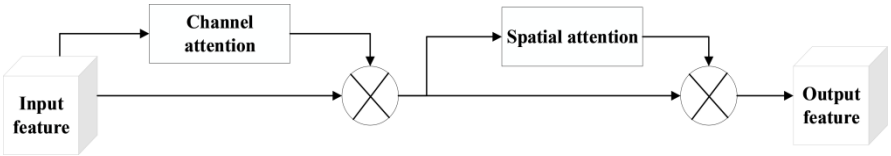

Figure 6. CBAM structure with channel before space.

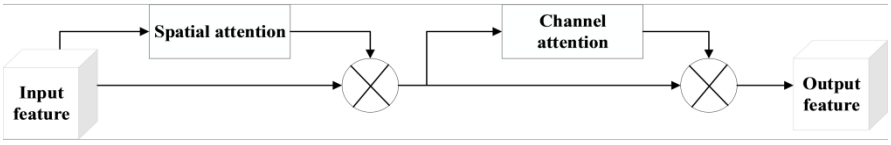

Figure 7. CBAM structure with space before channel.

The CBAM attention mechanism adjusted the convolutional layer's weight to ensure that the weighted features were transmitted farther back. The specific approach is shown in Figure 9.

Secondly, we introduced CBAM to VGG16. Since the features obtained after all convolution operations of VGG16 retain important local feature information, the convolutional layer of VGG16 was used as the backbone. We added the CBAM attention mechanism in between to enhance the original feature map's expression ability and improve the classification accuracy. The specific approach is shown in Figure 10.

Finally, we introduced CBAM to ResNet50. The features obtained by the first layer of convolution contain more local key information. Therefore, the CBAM structure was added after the first layer of ResNet50 to capture the first convolutional layer's detailed features. When the model performed identity mapping, the important features learned through the CBAM structure could be transmitted farther back to improve the classification effect. The specific approach is shown in Figure 11.

3.4. Mixup Data Enhancement

To improve the classification accuracy, we introduced the Mixup data enhancement

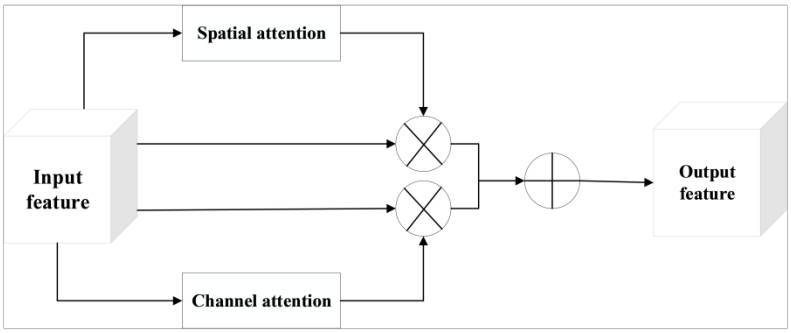

Figure 8. CBAM structure with parallel space and channel.

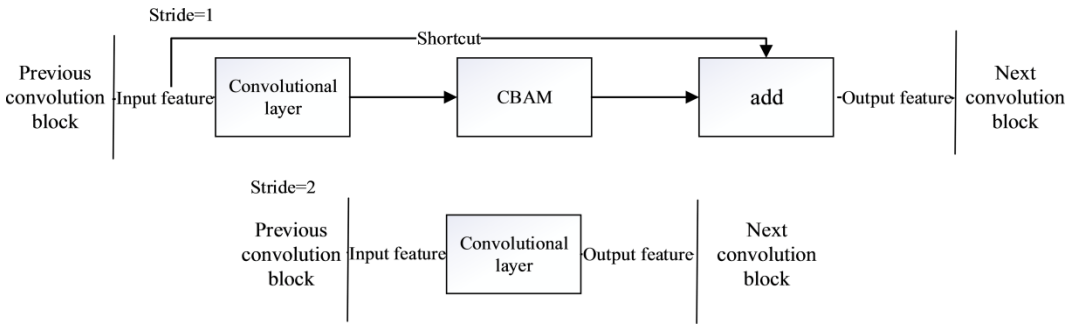

Figure 9. CBAM structure based on MobileNetV2.

Figure 10. CBAM structure based on VGG16.

algorithm. Virtual samples expand the training set to describe each sample's neighbourhood in the training data, and virtual samples are extracted from the neighbourhood of the training samples to expand the range of the training sample distribution. When the model makes decisions, the decision boundary is blurred to provide smoother predictions. Mixup can be expressed by Equations (3).

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{3}$$

Among them, (x_i, y_i) and (x_j, y_j) are two samples randomly selected from the training data, $\tilde{x}$ is the mixed virtual sample, $\tilde{y}$ is the label corresponding to the mixed sample, and $\lambda \in [0,1]$, $\lambda \sim Beta(\alpha, \alpha)$. Mixup generates virtual samples by combining linear interpolation of two training samples and linear interpolation of corresponding related labels. The blending hyperparameter α controls the intensity of interpolation between two samples. The renderings of some Asian food images enhanced with Mixup data are shown in Figure 12.

4. Experimental Analysis and Discussion

4.1. Experimental Data Set Preprocessing and Evaluation Indicators

4.1.1. Picture Segmentation

We used UECFOOD100 [1] created by Yoshiyuki Kawano from the University

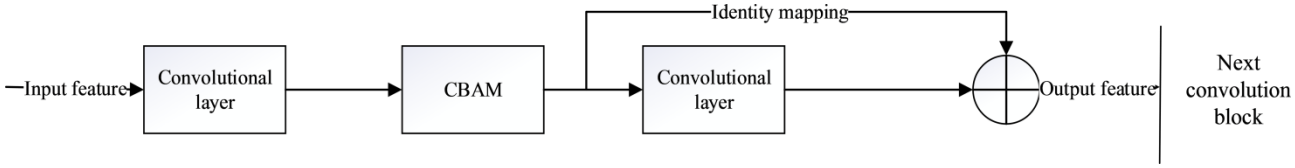

Figure 11. CBAM structure based on ResNet50.

Figure 12. Virtual samples of mixup.

of Electro-Communications, Japan. The experimental data set contains 14361 pictures of 100 types of Asian food, including rice, fried rice, curry rice, tempura rice, and udon noodles, *etc.*, as shown in Figure 13.

To avoid the interference of irrelevant features such as the background of the food image, we combined the label information documents of the food subject in the data set and wrote a python script to preprocess the image segmentation. Taking rice as an example, the food image detection frame was restored by labelling the information file, and the food subject in the detection frame was segmented, as shown in Figure 14.

Figure 13. Partial images of UECFOOD100 data set.

Figure 14. Image segmentation of Asian food dataset.

Among them, column (a) is the original picture, column (b) shows the information of the food detection frame, and column (c) shows the rendering of the content of the detection frame. It can be observed that the image segmentation method that we proposed reduces the influence of the image background and avoids the interference of irrelevant features such as the background.

4.1.2. Experimental Evaluation Index

In image classification experiments, the accuracy of Top-1 and Top-5 is often used to measure the effectiveness of the model. The accuracy of Top-1 and Top-5 can be expressed by Equations (4).

$$\begin{aligned} \text{Top-1} &= \frac{n_1}{n} \\ \text{Top-5} &= \frac{n_5}{n} \end{aligned} \tag{4}$$

Among them, n_1 is the number of graphs with the correct label corresponding to the first classification probability label in the test picture, n_5 is the number of graphs with the correct label in the test picture among the first five classification probability labels, and n is the number of tests pictures.

4.2. Experimental Process and Result Analysis

Our experiment used the Pytorch framework, and it was carried out on the NVIDIA GeForce RTX GPU. To further improve the classification accuracy, we used common data enhancement techniques, which included random cropping and random flipping to expand the training set. In the training process, the picture was scaled to 224 × 224 pixels, and the test set picture was scaled to 256 × 256 during testing, and then the centre was cropped to 224 × 224 as input.

4.2.1. The Effect of the Combination Mode of CBAM on Classification

We set the initial learning rate at 0.01, momentum at 0.9, batch size at 32, and training for 90 epochs. When training to 30 Epoch, the learning rate decayed to 0.001, and when training to 60 Epoch, the learning rate decayed to 0.0001 on the basis of 30 Epoch. In order to calculate the gradient quickly, and the models could converge at a faster speed, we selected sgd as the optimizer. The experimental results are shown in Table 1.

Table 1 shows the classification accuracy of the CBAM series structures. The classification accuracy of "Channel before Space" is higher than the classification

Table 1. Comparison of the classification effect of the modified CBAM and the original CBAM on Asian food.

Model	Channel before Space CBAM	Space before Channel CBAM	CBAM with parallel Channel and Space
VGG16	72.84%	71.92%	71.01%
MobileNetV2	77.55%	77.27%	75.86%
ResNet50	75.37%	75.16%	75.16%

accuracy of "Space before Channel" and the classification accuracy of "parallel structure".

On VGG16, the classification accuracy of "Channel before Space" is improved by 0.92% and 1.83% compared to the classification accuracy of "Space before Channel" and "parallel structure". On MobileNetV2, the classification accuracy of "Channel before Space" is improved by 0.28% and 1.69% compared to the classification accuracy of "Space before Channel" and "parallel structure". On ResNet50, the classification accuracy of "Channel before Space" is improved by 0.21% and 0.21% compared to the classification accuracy of "Space before Channel" and "parallel structure".

The line chart of Top-1 training accuracy is shown in Figure 15, and the line chart of Top-1 testing accuracy is shown in Figure 16. The x-axis in the figure represents epoch, and the y-axis represents the accuracy of Top-1 training or testing.

Because the learning rate is reduced by 0.1 when the training reaches 30 epoch and 60 epoch, it contributes to the large fluctuation of the figure's broken line. Figure 15 and Figure 16 further confirm that the classification effect of the "channel before space" is better than the other two CBAM structures. Therefore, they further confirm the "channel before space" is more suitable for Asian food.

4.2.2. The Effect of CBAM on Classification

To further verify the proposed CBAM mechanism's effectiveness, we combined CBAM with MoblieNetV2, VGG16, and ResNet50, and selected "channel before space" as the CBAM structure. Our models used the ImageNet pre-training

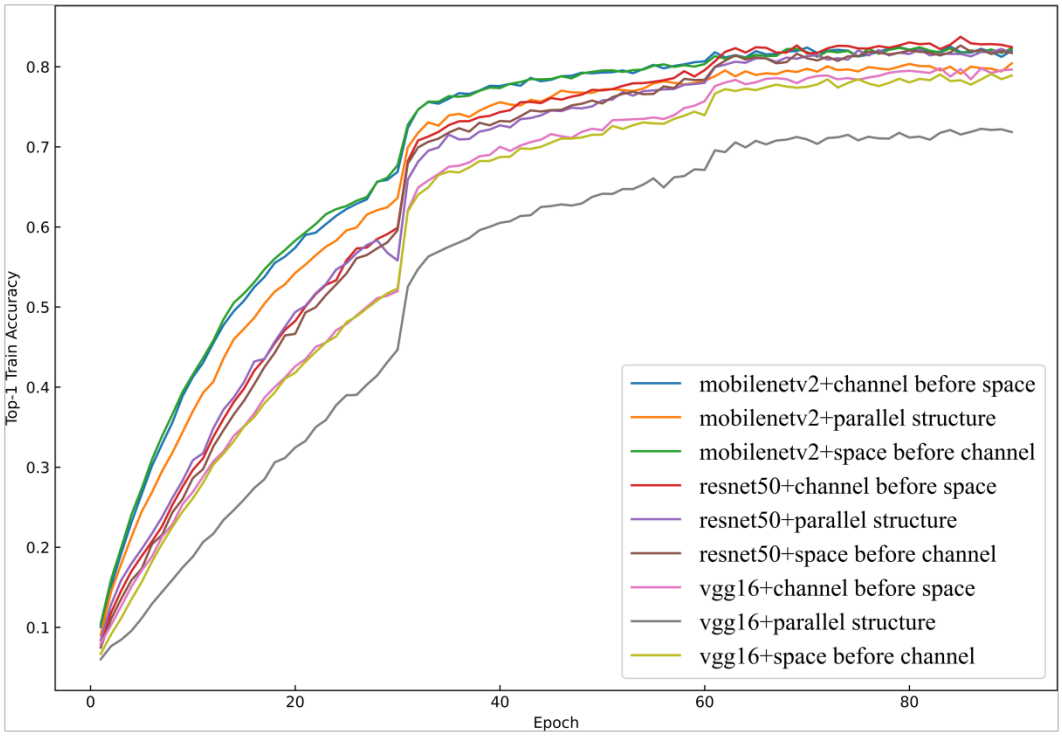

Figure 15. Top-1 train accuracy line chart of different CBAM structures.

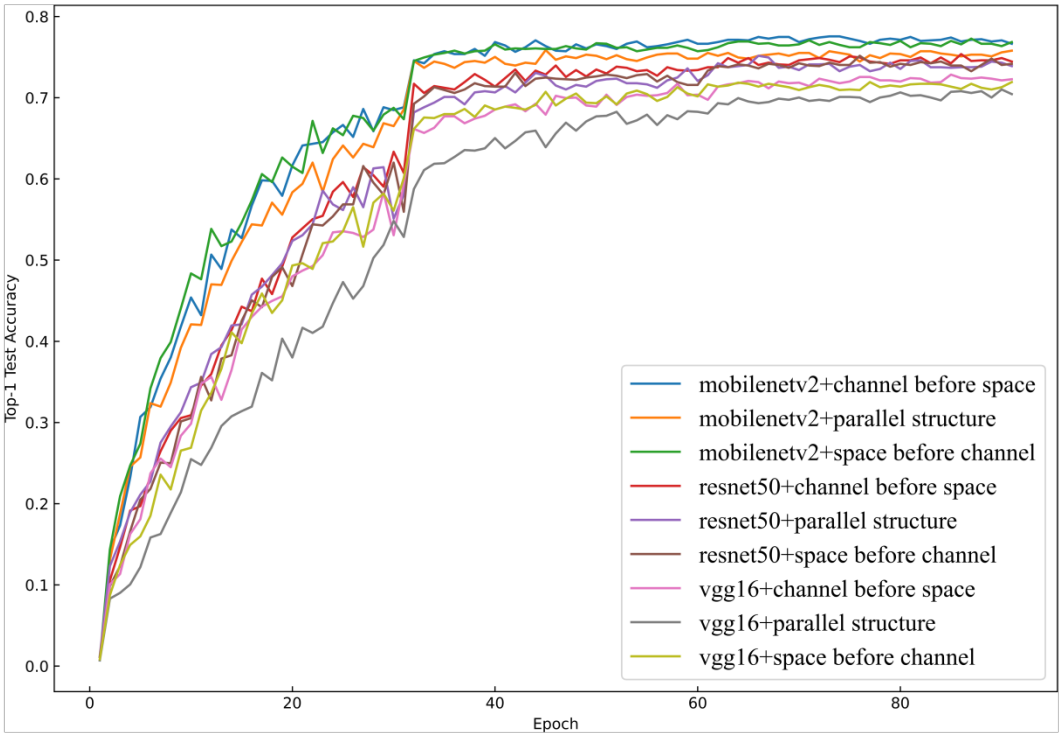

Figure 16. Top-1 val accuracy line chart of different CBAM structures.

Table 2. Comparison of classification effects of different network models on Asian food datasets.

Model	Top-1/%	Top-5/%
VGG16	82.90	96.55
VGG16 + CBAM	83.95	96.62
MobileNetV2	84.24	96.62
MobileNetV2 + CBAM	85.22	96.76
ResNet50	85.15	97.04
ResNet50 + CBAM	86.56	97.19

weights. We set initial learning rate to 0.01, the momentum to 0.9, and the batch size to 32. What's more, we set the training epochs to 160. When training to 90 Epoch, the learning rate decayed to 0.001, and when training to 120 Epoch, the learning rate decayed to 0.0001 on the basis of 90 Epoch. In order to calculate the gradient quickly, and the models could converge at a faster speed, we selected sgd as the optimizer. The results are shown in Table 2.

It can be seen from Table 2 that CBAM models have better classification performance on the Asian food data set than benchmark models. The x-axis in the figure represents epoch, and the y-axis represents Top-1 training or testing accuracy. Because the learning rate is reduced by 0.1 when the training reaches 90 epoch and 120 epoch, it contributes to the large fluctuation of the figure's broken line.

Figure 17 and Figure 18 show that the classification accuracy of models after introducing the CBAM structure is improved compared with the original models. This shows that the CBAM attention mechanism is effective in solving the

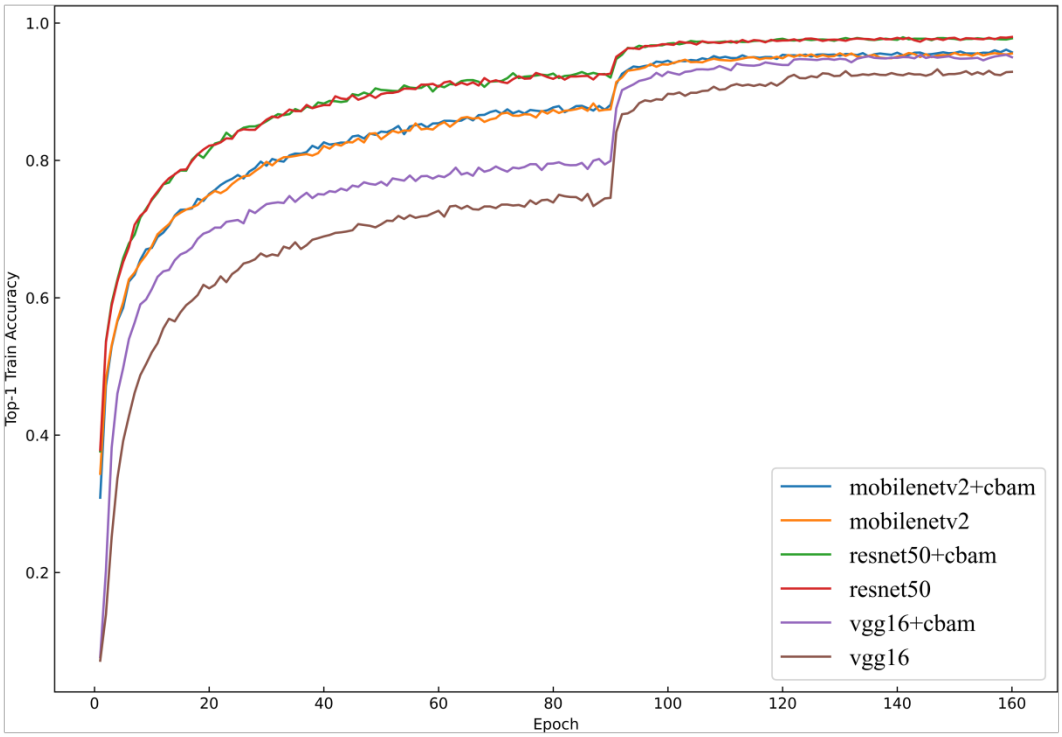

Figure 17. Top-1 train accuracy line chart of different models.

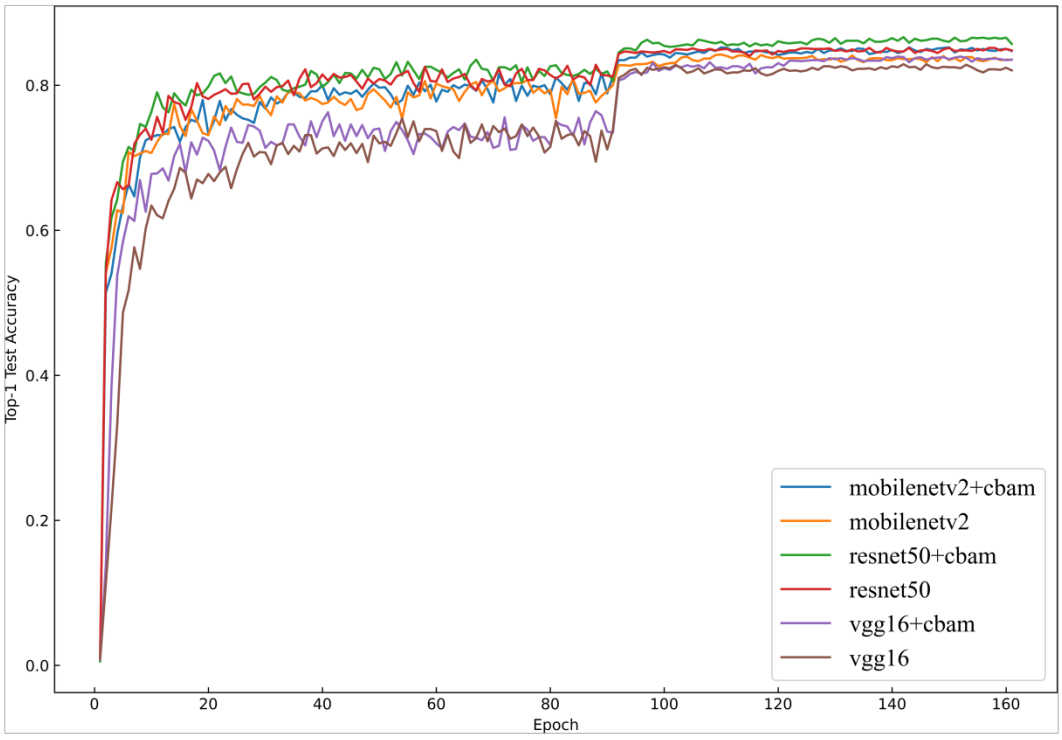

Figure 18. Top-1 val accuracy line chart of different models.

problem of Asian food image classification.

4.2.3. The Effect of Mixup on Classification

In addition, we performed Mixup data enhancement preprocessing on the Asian food data set, we set the hyperparameter to 0.4, and the experimental results are shown in Table 3.

It can be seen from Table 3 that the classification performance of our Mixup preprocessing models are better than benchmark models on the Asian food data set. The line graph of Top-1 training accuracy is shown in Figure 19. The Top-1

Table 3. Comparison of classification effects of models on Asian food after Mixup data enhancement.

Model	Top-1/%	Top-5/%
VGG16	82.90	96.55
VGG16 + Mixup	85.10	97.11
VGG16 + CBAM + Mixup	85.15	97.11
MobileNetV2	84.24	96.62
MobileNetV2 + Mixup	86.28	96.62
MobileNetV2 + CBAM + Mixup	86.28	97.11
ResNet50	85.15	97.04
ResNet50 + Mixup	87.26	97.26
ResNet50 + CBAM + Mixup	87.33	97.33

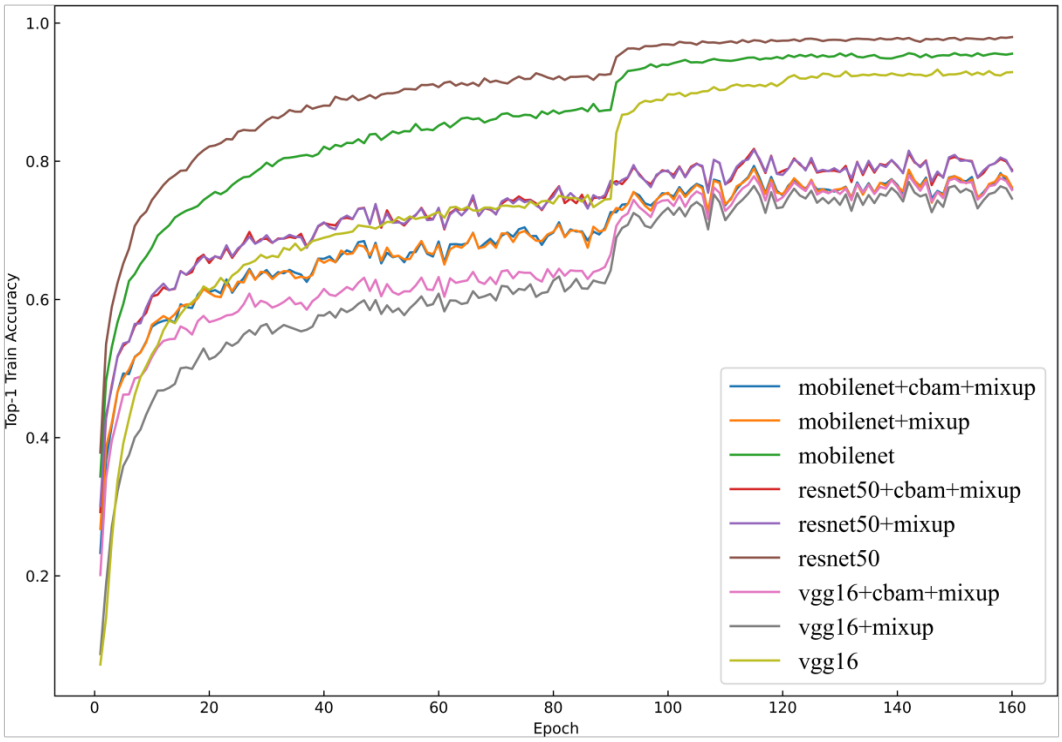

Figure 19. Top-1 train accuracy line chart of different mixup models.

test accuracy is The line chart is shown in Figure 20. The x-axis in the figure represents epoch, and the y-axis represents Top-1 training or testing accuracy. Although the virtual samples reduce training accuracy, the classification accuracy of the models trained on the virtual samples has improved on the test samples. It shows that the Mixup data enhancement pre-processing method can show good classification results on multiple models, and it effectively solves the problem of Asian food image classification.

4.2.4. Comparison of Different Asian Food Image Classification Methods

Table 4 shows the comparison results between our models and the previous research on UECFOOD100. It can be seen from Table 4 that our method are superior to other methods in classification effect. Among them, the VGG16 +

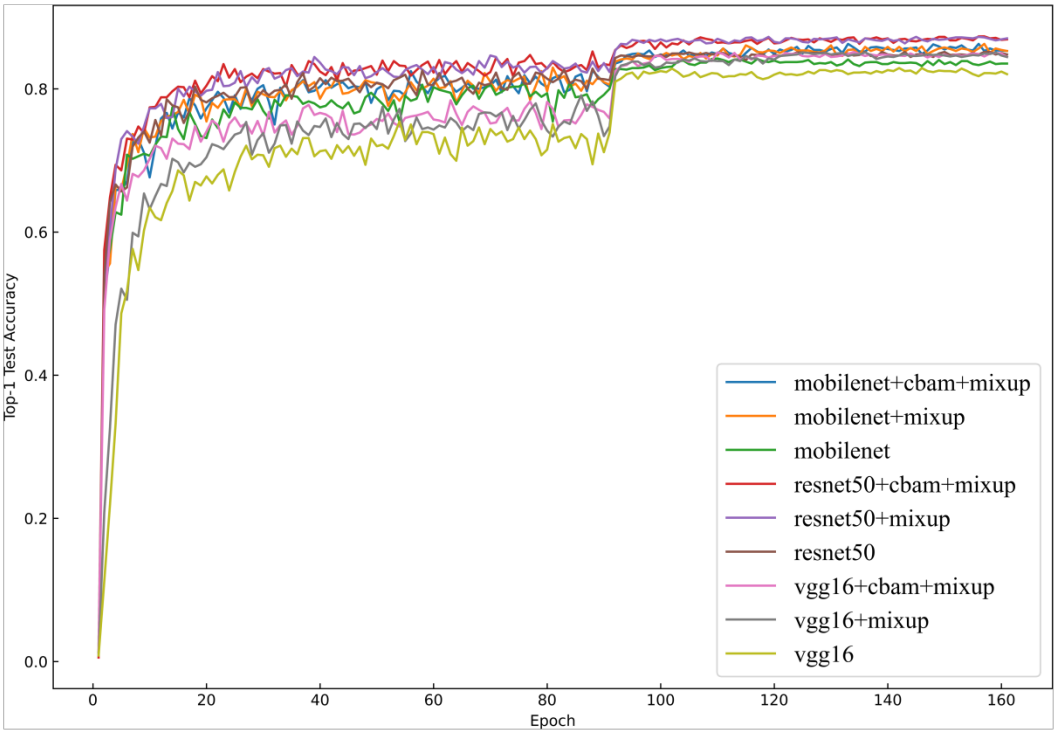

Figure 20. Top-1 val accuracy line chart of different mixup models.

Table 4. Comparison of classification effects of different methods on Asian food.

Method	Top-1/%	Top-5/%
DeepFoodCam [13]	72.26	92.00
DeepFood [14]	76.30	94.60
FV + DeepFoodCam [13]	77.35	94.85
DCNN-FOOD [15]	78.77	95.15
VGG16 + CBAM + Mixup	85.15	97.11
MobileNetV2 + CBAM + Mixup	86.28	97.11
ResNet50 + CBAM + Mixup	87.33	97.33

CBAM + Mixup model is 7.80% higher than the Top-1 of the FV + DeepFood-Cam [13] method, Top-5 is increased 2.26%, which is 8.85% higher than the Top-1 of the DeepFood [14] method, and Top-5 is increased by 2.51%, which is 6.38% higher than the Top-1 of the DCNN-FOOD [15] method, Top-5 is increased by 1.96%. This shows that our method is effective in dealing with the problem of Asian food image classification and improves the classification accuracy.

4.3. Visual Comparison

To visually display the key information that our CBAM models focus on, we introduced Grad-CAM [16] for visualization experiments. The heat map was used to display the Asian food picture information that the convolutional neural network and CBAM focus on. We took ResNet50 as an example and randomly selected 3 pictures, and the generated heat map comparison chart is shown in **Figure 21**.

Among them, the column of (a) is the original picture, the column of (b) is the effect picture of the benchmark network, the column of (c) is the effect picture after adding the CBAM attention mechanism. It can be observed that after the introduction of the CBAM module, the models pay more attention to the features in the Asian food images area, thereby greatly improve the classification effect.

5. Conclusion

To improve Asian food image classification accuracy, we designed a method to

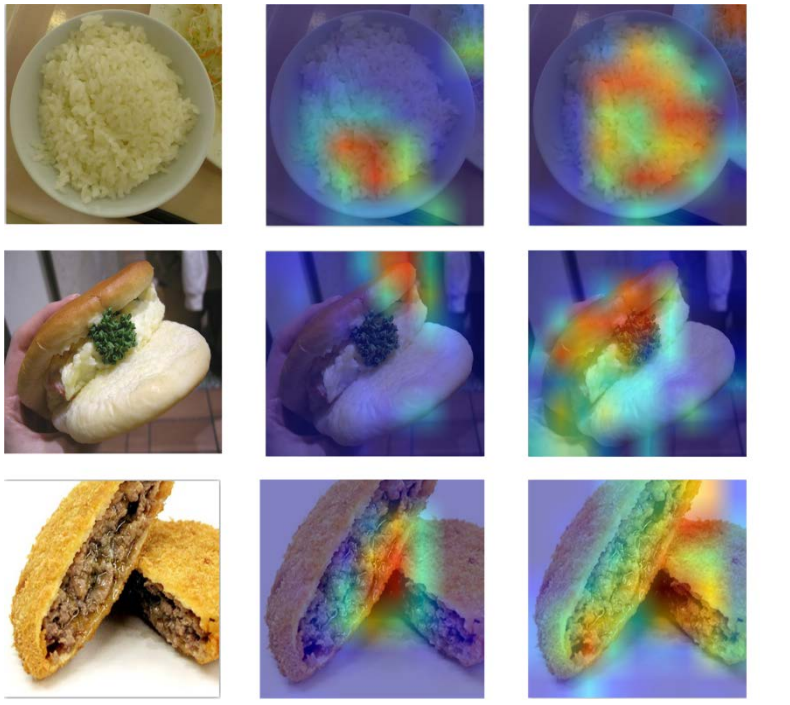

Figure 21. Visual comparison of heat map.

use three convolutional neural networks MobileNetV2, VGG16, ResNet50 as the reference network, and used the CBAM attention mechanism to improve the reference network. In addition we used the Mixup data enhancement algorithm to expand. The training set and comparative experiments on multiple models show that the method can effectively improve the accuracy of Asian food image classification, and the accuracy rate is higher than the accuracy of Asian food image classification in recent years, and we used the heat map to further verify the effectiveness of CBAM attention mechanism. Our method provides a new idea for solving the problem of Asian food image classification, which is in line with the purpose of introducing the CBAM attention mechanism and Mixup data enhancement.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Matsuda, Y., Hoashi, H. and Yanai, K. (2012) Recognition of Multiple Food Images by Detecting Candidate Regions. *Proceedings of* 2012 *IEEE International Conference on Multimedia and Expo*, Melbourne, 9-13 July 2012, 25-30. https://doi.org/10.1109/ICME.2012.157

[2] Bossard, L., Guillaumin, M. and Gool, L.V. (2014) Food-101-Mining Discriminative Components with Random Forests. In: *European Conference on Computer Vision* (*ECCV*), Springer, Cham, 446-461. https://doi.org/10.1007/978-3-319-10599-4_29

[3] Bolaños, M. and Radeva, P. (2016) Simultaneous Food Localization and Recognition. *International Conference on Pattern Recognition* (*ICPR*), Cancun, 4-8 December 2016, 1-6. https://doi.org/10.1109/ICPR.2016.7900117

[4] Yang, S., Chen, M., Pomerleau, D., *et al.* (2010) Food Recognition Using Statistics of Pairwise Local Features. *The* 23*rd IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), San Francisco, 13-18 June 2010, 2249-2256. https://doi.org/10.1109/CVPR.2010.5539907

[5] Fu, Z., Chen, D. and Li, H. (2017) Chin Food 1000: A Large Benchmark Dataset for Chinese Food Recognition. In: *International Conference on Intelligent Computing*, Springer, Cham, 273-281. https://doi.org/10.1007/978-3-319-63309-1_25

[6] Taskiran, M. and Kahraman, N. (2019) Comparison of CNN Tolerances to Intra Class Variety in Food Recognition. *IEEE International Symposium on Innovations in Intelligent Systems and Applications* (*INISTA*), Sofia, 3-5 July 2019, 1-5. https://doi.org/10.1109/INISTA.2019.8778355

[7] Sukvichai, K., Maolanon, P., Sawanyawat, K., *et al.* (2019) Food Categories Classification and Ingredients Estimation Using CNNs on Raspberry Pi 3. *International Conference of Information and Communication Technology for Embedded Systems* (*IC-ICTES*), Bangkok, 25-27 March 2019, 1-6. https://doi.org/10.1109/ICTEmSys.2019.8695967

[8] Sandler, M., Howard, A., Zhu, M., *et al.* (2018) MobileNetV2: Inverted Residuals and Linear Bottlenecks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474

[9] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556v1

[10] He, K., Zhang, X., Ren, S., *et al.* (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[11] Woo, S., Park, J., Lee, J., *et al.* (2018) CBAM: Convolutional Block Attention Module. *European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[12] Zhang, H., Cisse, M., Dauphin, Y.N., *et al.* (2017) Mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations*, Vancouver, April 2018, 1-13. https://arxiv.org/abs/1710.09412

[13] Kawano, Y. and Yanai, K. (2014) Food Image Recognition with Deep Convolutional Features. *ACM International Joint Conference on Pervasive and Ubiquitous Computing* (*Ubi-Comp*), Seattle, September 2014, 589-593. https://doi.org/10.1145/2638728.2641339

[14] Liu, C., Cao, Y., Luo, Y., *et al.* (2016) Deepfood: Deep Learning-Based Food Image Recognition for Computer-Aided Dietary Assessment. *IEEE International Conference on Smart Homes and Health Telematics*, Volume 9677, 37-48. https://doi.org/10.1007/978-3-319-39601-9_4

[15] Yanai, K. and Kawano, Y. (2015) Food Image Recognition Using Deep Convolutional Network with Pre-Training and Fine-Tuning. *IEEE International Conference on Multimedia & Expo Workshops* (*ICMEW*), Turin, 29 June-3 July 2015, 1-6. https://doi.org/10.1109/ICMEW.2015.7169816

[16] Selvaraju, R.R., Cogswell, M., Das, A., *et al.* (2017) Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/ICCV.2017.74