Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230284

Food image recognition based on ResNet

Yiming Xiong

Beijing Normal University-Hong Kong Baptist University United International College, Faclty of Science and Technology, Zhuhai, Guangdong province, 519000, China

q030026170@mail.uic.edu.cn.

Abstract. Image classification has always been one of the basic tasks in the community, which has been widely applied in many fields, such as the food recognition. As the key technology of dining robot, the food image recognition aims to predict the category of food in the given image, which has attracted a lots of research attentions from both the academia and industry. Early efforts of food image recognition mainly rely on the manual features, whose accuracy cannot meet practical application requirements. Thanks to the rapid development of convolutional neural networks, food image recognition based on deep learning has made breakthroughs in both accuracy and speed. In this paper, we propose a food image recognition method based on the ResNet. Extensive experiments demonstrate the effectiveness of our method, which can provide some new insights for the automatic food recognition.

Keyword: food recognition, image classification, ResNet, deep learning, computer vision.

1. Introduction

Computer vision refers to the computer and system through the picture, video and other visual input, obtain the required feature information, and according to these feature information through subsequent processing operations. As the most basic task of the computer vision, the image classification extracts the feature of input images and then predict the category of specific objects with the supervision of image category label. Nowadays, the image classification has been applied in various fields, such as the identification and judgment of various scenes in automatic driving, the automatic product qualification detection in various manufacturing industries. When it comes to the catering industry, most of the relevant technologies are only applied in big data analysis of food or store recommendation. To this end, we urgently need an automated food image recognition method.

The past decade has seen a gradually increase in the usage of deep learning techniques for solving different kinds of tasks, which can be credited to the presence of huge datasets with dense annotation, such as ImageNet, MS-COCO, etc. In addition, benefiting from the powerful feature extraction and representation ability of convolutional neural network, better accuracy has been achieved on various tasks [1]. In the food recognition algorithm model, the traditional image processing and machine learning methods have many shortcomings. Most of the research is based on the food classification tasks, which is easily affected by light intensity, noise interference, location direction and other factors in the actual scene. With the development of deep learning, convolutional neural networks have been great success in image recognition, object detection and other fields. To this end, the food recognition

DOI: 10.54254/2755-2721/8/20230284

technology is gradually focused on convolutional neural networks, which can works as a basic but key technology of the robot equipped with food classification [2]. This can effectively save the labor force required by the catering industry, such as waiters and deliverymen, etc., but also optimize the restaurant's hygiene issues, such as identifying food residues and cleaning it.

In this paper, we will use a food image recognition network based on ResNet, which aims to help the food recognition robot move towards practical applications. Specifically, we naturally focus on relatively mature models in deep learning and conduct research on ResNet to find out the shortcomings. We use Food-101 datasets from Kaggle and select 10 kinds of foods to train the ResNet model through pre-processing data. To evaluate the model performance, we conduct extensive experiments to quantitatively and qualitatively analyze the model identification results. We get the accuracy, precision, recall and F1 score of our model. By comparing the above values, we obtained the advantages of the model and even supplemented how to make up for the shortcomings. We believe this work can provide some new insight for the automatic food recognition as well as food recognition robot.

2. Method

2.1. Revisiting convolutional neural network

The CNN structure mainly consists of the input layer, convolution layer, activation layer, fully conceded layer and output layer [3]. (1) Input layer and output layer are used to input and output data, respectively. (2) Convolution layer: Feature extraction and mapping are carried out through convolution kernel. Multiple convolution kernels are stacked into a filter, and a group of filters form a convolution layer. Filter slides on the input data to do convolution calculation and obtain extraction features. (3) Activation layer: Nonlinear mapping, using ReLU activation function, and the same for subsequent activation layer. (4) Pooling layer: down sample and dimension reduction. (5) Rasterization: Each pixel of all feature map of the upper layer is successively expanded and arranged in a row to fully connect with the traditional multi-layer perceptron MLP. (6) Fully Conceded layer: each neuron is connected and fitted in the tail to reduce the loss of feature information (7) Output layer: Indicates the output layer. In addition, convolution layer, activation layer and pooling layer can have multiple layers. CNN expects that with the deeper and a lot more layers, the network can be closer to the objective function through a huge number of nonlinear mappings and improved features representation. Theoretically, the more layers the neural network has, the more complex feature extraction can be carried out and better results can be achieved. However, the more layers the network has and the deeper the network will be, the more significant the problem of gradient disappearing or explosion will be, which will make the final effect fall short of the expected effect. Therefore, this paper selected ResNet to solve this problem.

2.2. ResNet

ResNet has a very significant improvement in image compared with CNN, which is due to its special residual learning block.

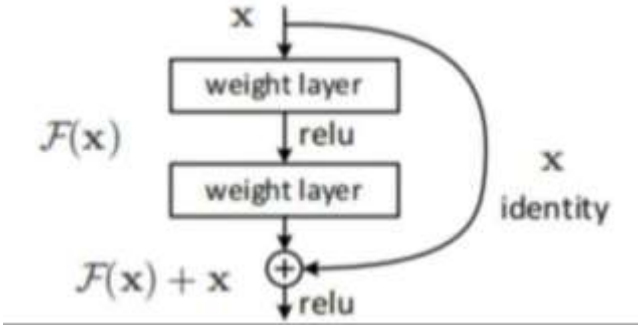


Figure 1. The framework of residual block.

DOI: 10.54254/2755-2721/8/20230284

As shown in Figure 1, residual block is used to fix the degradation problem. For a stacking residual block (composed of several layers), the features of the input are denoted as . The desired residual feature is, which is defined since residual learning is easier to learn than raw features directly [4]. When the residual equals to 0, then the accumulation layer is only identically mapped, at least the network performance will not degrade. When mapping to directly transmit the current output to input of the network of the next layer (all 1:1 original transmission, no parameters are added), it is equivalent to taking a shortcut or called skipping of this layer. This direct connection also names "skip connection" or "shortcut connection". At the same time, in the process of backward propagation, the gradient of the next layer of the network is directly transmitted to the upper layer of the network, thus solving the gradient vanish problem.

The network structure of ResNet is essentially a modification on the VGG19 network. The above shortcut connection is added between each two layers to form residual learning. ResNet makes down-simple convolution with a stride of 2, and use Global Average Pool layer to replace the Fully Connected layer. There is another important design: when the size of feature map is cut to half, the number of feature maps will be doubled, which allows the complexity of the network layer to be maintained. That is, when the convolutional layer changes (changes to different sizes), such as from conv2_x to conv_3x, the number of feature maps will change. Also, for a shortcut connection, you can add the input directly to the output if the dimension of the input and output are equal. But once the dimensions are not equal, they cannot be added directly. The ResNet authors offer e three options: (A) zero-padding shortcuts makes dimensions increased and all shortcuts are parameter-free ; (B) projection shortcuts increase dimensions, and other shortcuts are identity; (C) all shortcuts are projections [4].

ResNet has different layers, such as ResNet18, ResNet34, ResNet50, ResNet101 and so on. However, considering the portability, flexibility, etc., in the selection of models, such as ResNet101 and above need certain computing ability, but also to ensure high accuracy. Therefore, ResNet34 and ResNet50 training were adopted in this project.

2.3. ResNeXt and ResNeSt

In 2017, the ResNet production team broadened the convolution layer horizontally, increasing the number of cardinality can improve the accuracy of model classification, and it will be more effective than making the model deeper and wider. At the same time, the guarantee complexity does not change much, and the model is slightly improved [5]. ResNeSt introduced Split-Attention on the basis of ResNet in 2020, which improved the average accuracy of the paper by about 3% [6].

3. Result

3.1. Original data and pre-processing

The Food-101 dataset coming from kaggle is used for the training of recognition model, which contains image of food and organized by type of food. We only used a subset of 10 foods for training and testing. Since ResNet only does classification, not detection, it does not need to label images. The training set and verification set were randomly clipped, flipped and standardized.

3.2. Model convergence analysis

To verify the effectiveness of model training, we chose the ResNet34, ResNet50 and ResNet34-CBAM as the backbone, and further visualize the loss of model training and its corresponding accuracy changes, whose results are shown in Figure 2. It can be observed that all three networks can obtain a accuracy of 90% with the increase of training procedure, while all the losses gradually decrease from 0.8 to 0.2. All the results demonstrate that our method which can recognize various food accurately is effective.

DOI: 10.54254/2755-2721/8/20230284

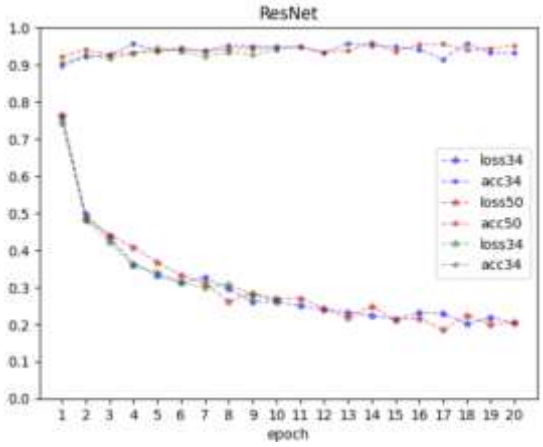
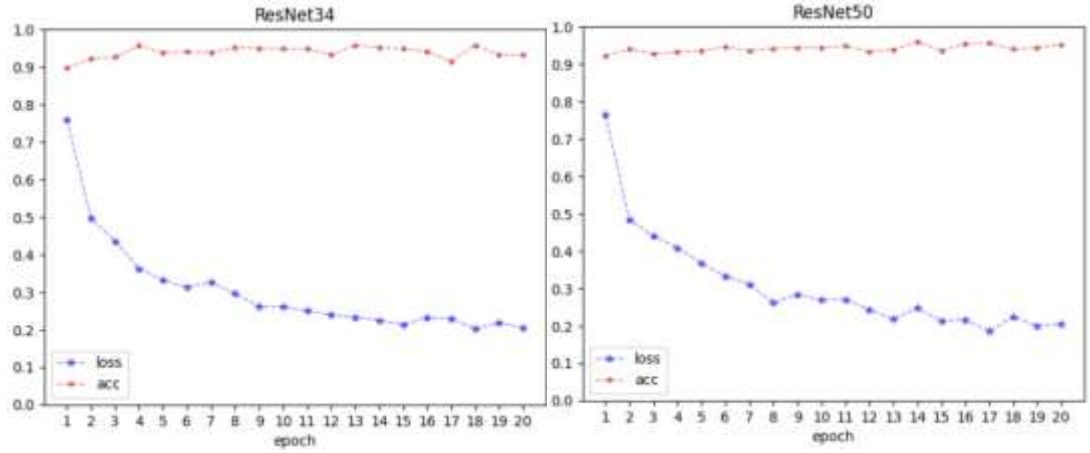
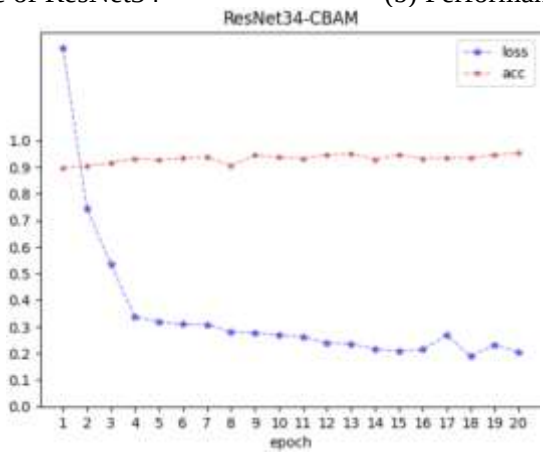


Figure 2. Comparison of three different network.

We also report the detail results of three model, which can be seen in Figure 3. For the ResNet34 and ResNet50, it can be observed that similar recognition accuracy can be achieved on the test data set, while the ResNet50 shows a better better stability. For the ResNet34-CBAM, it shows a better convergence. Specifically, the loss of ResNet34-CBAM starts at the point more than 2, while it can decrease to about 0.2 within only 20 epochs. We suppose that this is due to the help of channel attention and spatial attention in feature learning, which can combine classification losses to learn high-quality image features faster.



(a) Performance of ResNet34 (b) Performance of ResNet50



(c) Performance of ResNet34-CBAM

Figure 3. Performance of various networks for the food recognition.

DOI: 10.54254/2755-2721/8/20230284

3.3. Quantitative analysis for different categories

In order to further analyze the model recognition results, we have reported their recognition accuracy for different types of food, and the results are shown in Table 1. The values in the following table is basically higher than 0.98 and are the highest values measured reaches the possibility of 1. However, the possibility of a slice of pizza showed a significant drop in accuracy, with the pizza likely to go down and the sandwich likely to go up.

Table 1. Accuracy for different categories.

Category	0	1	2	3	4	5	6	7	8	9
Accuracy	0.992	0.999	0.982	0.985	0.992	0.994	0.993	0.987	1.0	0.995

3.4. Visualization analysis

We finally visualize the recognition result of different category of food, which can be seen in Figure 4. By calculating the probabilities of ten preset foods, and return the class which has the highest probability. And the cost of time only considering the prediction always less More than 0.2 second.

Figure 4. Visualization of result of different categories.

4. Discussion

According to Figure 2, it can be clearly seen that after 5 iterations of ResNet34 and ResNet50 models in this scenario, the accuracy of verification set has tended to be stable and fluctuated in a small amplitude, basically maintaining around 0.93, preferably no more than 0.952. As for the training loss, the original ResNet34 and ResNet50 started to decline slowly in the middle of the training, basically around 0.22. In Figure 2, the curves of ResNet34 and ResNet50 under epoch20 are basically the same. Therefore, it can be seen that the model of ResNet34 is able to easily classify the foods in 10. However, the project did not stop there. The author made the following three attempts: 1) Add a fully connected layer after the original fully connected layer, as shown in the following two figures; 2) There was also an attempt to add a convolution layer and a softmax layer to replace the full connection layer (not show the code here). 3) Add the attention model, namely ResNet-CBMA [7].

But the first two attempts have not worked well. For the attempt 1, the original intention is to increase the effect of feature integration by adding one more full connection layer, so as to improve the feature purity and facilitate the subsequent classification. However, the improvement of the final effect is shown in Figure 3(a). There is no actual improvement, but the training speed is reduced due to the excessive calculation amount of the full connection layer (so only epoch10 is trained). For the

DOI: 10.54254/2755-2721/8/20230284

attempt 2, the author borrowed the idea of FCN, named full convolutional neural network [8], to try to replace full connection layer with less computational deconvolution. In this attempt, the author found that the speed improvement could be increased by more than two times under cpu training, but the image details could not be captured due to the final up-sampling processing, which was obvious in the classification of pizza. Single slice pizza and whole pizza could easily be divided into two kinds of food, and such effect was obviously not desirable. As for attempt 3, after adding the two layers attention model to ResNet34, the training speed should theoretically decrease, but in fact, it is slightly higher than the pure ResNet34. As shown in Figure 3(c), the initial training loss of Resnet34-CBMA was much higher than that of the original ResNet34, but in the subsequent training, the decline rate of Resnet34-CBMA was faster than that of the original ResNet34. The loss of ResNet34-CBMA was lower than 0.2 for the first time and reached 0.19. Although the change of accuracy is still small, the change of loss can indicate that the speed and effect of model training are improved after the addition of CBMA.

Compared with PCA combined with KNN, SVM and other basic classification, the accuracy is only about 58%, the results of ResNet are particularly excellent. When considering robots, there are: 1) portability; 2) Anti-interference capability; 3) Add video stream input [9]. portability means less computing needs, and ResNet34-CBMA can do great job when there is a few different kinds of food, the accuracy is not that different, the computing cost is less than 50,101, and 0.2 second of perdiction time is perfectly acceptable. Considering the anti-interference capability, although ResNet has achieved excellent results in 10 food categories, it has not been tested in real world scenarios. In the real scene, different lighting, different forms of the same food, target occlusion and so on [10]. In the training and test data, most of the whole image is the target theme, and there is no interference from other food. In addition, the video stream input obtained by the camera should be converted into the model can accept, or add video stream reception to the model. All in all, robots installed in models are still a long way from being implemented, which is far from being completely solved by an ideal high-precision model.

5. Conclusion

In this paper, we tested a food image recognition algorithm based on the ResNet, which aims to serve the application of food classification robot. Specifically, we construct three different kinds of ResNet models based on the ResNet34, ResNet50 and ResNet34-CBAM. Extensive experiments result data demonstrate the effectiveness of our proposed method, which can achieve a accuracy of more than 90%. We also compare the convergence of three models and recognition accuracy for each category of food. All the results show our method also has a good robustness for the change of scenes.

Reference

[1] Agarwal, A., Mangal, A., & Vipul. (2020). Visual Relationship Detection using Scene Graphs: A Survey. arXiv preprint arXiv:2005.08045.

[2] Ye, L. F. (2020). Research on food automatic recognition algorithm based on deep learning [Master's thesis]. Zhejiang Normal University.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[5] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).

[6] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Muller, J., Manmatha, R., Liang X. & Vasconcelos N. (2020). ResNeSt: Split-Attention Networks. arXiv preprint arXiv:2004.08955.

DOI: 10.54254/2755-2721/8/20230284

[7] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Revisiting ResNets: Improved Training and Scaling Strategies. arXiv preprint arXiv:2103.07579, 2021.
[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. arXiv preprint arXiv:1411.4038, 2014.
[9] Basler. (2021). AI-powered computer vision for industrial robots. Basler, 1(1), 1-5.
[10] Liu, Y., & Zhang, Y. (2021). Detection of paper notes - based on YOLO deep convolutional neural network for robot apple picking positioning under complex background. Journal of Physics: Conference Series, 1946(1), 012008.