Food Image Recognition with Convolutional Neural Networks

Weishan Zhang[1] Dehai Zhao[1] Wenjuan Gong[1] Zhongwei Li[1] Qinghua Lu[1] Su Yang[2]
[1]Department of Software Engineering, China University of Petroleum
No. 66 Changjiang West Road, Qingdao, China. 266580
[2]College of Computer Science and Technology, Fudan University, Shanghai, China. 200433
{zhangws,wenjuangong,zhongweili}@upc.edu.cn {1085884664}@qq.com {suyang}@fudan.edu.cn

***Abstract*—In this paper, we propose a food image recognition system with convolutional neural networks(CNN), which has been applied to image recognition successfully in the literature. A CNN which consists of five layers has been built and two group of controlled trials have been processed on it.Two datasets are prepared: one is UEC-FOOD100 dataset which is an open 100-class food image dataset including about 15000 images and the other is a fruit dataset that established by ourselves including over 40000 images. We have achieved the best accuracy of 80.8% on the fruit dataset and 60.9% on the multi-food dataset. In addition, we validate the method on two groups of controlled trials and discover the effect of color under various conditions that the color feature is not always helpful for improving the accuracy by comparing the results of two group of controlled trials. As future work, we will combine image segmentation with image recognition to get a better performance.**

***Keywords*-food image recognition; convolutional neural networks; multi-layer neural network**

I. Introduction

With the rapid development of our society, more attention has been paid to the quality of life, especially the food we eat. But classifying food manually is not applicable to this fast-tempo society anymore. An automatic food classification system with increased accuracy, improved speed and reduced production cost is urgently needed. In recent years, computer vision systems have been used vastly in food recognition methods. Generally, there are mainly two methods: one is conventional method including image preprocessing, feature extraction, feature selection and classification. For example, Marios M. Anthimopoulos et al.[2] propose a method based on the bag-of-features (BoF) model by computing dense local features and finally classify the food images with a linear support vector machine(SVM) classifier, and the system achieves classification accuracy of the order of 78%. The other applies deep learning, which is a popular method recently in the image recognition field. Alex Krizhevsky et al.[3] build a deep conventional neural network(DCNN) and train it through ImageNet LSVRC-2010. The network gets an accuracy of 84.7%.

However, by studying recent research, we find most work have used conventional method rather than the potential deep learning method because it is a proven technique. Now we present our work based on CNN and divide it into the following steps: preparing dataset, building networks, training and testing the networks. To find the influence factors of the performance of the networks, we design two group of controlled trials and compare the results between them. At the same time, four different forms of dataset has been prepared by transforming the original images.

The outline of the paper is as follows: section 2 gives the food recognition method based on CNN. In section 3, we introduce the experiments that we did and discuss the results that we achieved. Section 4 presents some related work. Conclusion and future work end the paper.

II. CNN based Approach for Food Recognition

We propose a multi-food image recognition system with CNN which can handle 100 kinds of food. As can be seen in Figure 1, this is a simple structure of the system. Given an input food image, first, the CNN can accomplish all the recognition steps including feature extraction, shift and distortion invariance and classification, and then, gives the label as the output.

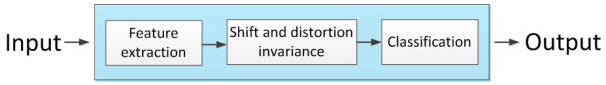

Figure 1. A simple structure of food image recognition system

A. Basic Introduction to CNN

Since Krizhevsky won the ImageNet Large-Scale Visual Recognition Challenge(ILSVRC) 2012 using Deep Convolutional Neural Network (DCNN)[3], CNN became a hot topic in the image recognition field. The characteristic of shared weights makes it be more similar to biological neural network, reduces the complexity of the network model and the amount of the weights, especially when the input is a multi-dimensional image. As a special designed multilayer perceptron for detecting 2-D shape, the network has a high invariance of translation, scale, incline and other transformations. Comparing with conventional methods, CNN

978-1-4673-7211-4/15 $31.00 © 2015 IEEE
DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.139

avoids the process of complicated feature extraction and data reconstruction because the image can be input into the network directly.

CNN is a multi-layer neural network. Each layer consists of several 2-D surface and each surface has plenty of single neural cells. Like the two related visual cortex cells, two basic cell types have been identified: simple cells respond maximally to specific edge-like patterns within their receptive field and complex cells have larger receptive fields and are locally invariant to the exact position of the pattern[1]. Generally, Us is defined as the feature extraction layer which is consisted by the simple cells. The input of each cell connects with the retina below and extract the local feature. Once the local feature is extracted, the position relationship with other features is confirmed. Similarly, complex cells make up the feature mapping layer called Uc, on which all the cells have the shared weights. This advantage makes the network have a few amount of free parameters, thus reduce the heavy work of parameter adjustment. The constraints on the model enable CNN to achieve better generalization on vision problems. Additionally, in CNN, each filter is replicated across the entire visual field. These replicated units form a feature map. If we denote the k-th feature map at a given layer as h^k, whose filters are determined by the weights w^k and bias b^k, then the feature map h^k is obtained as follows:

$$h_{ij}^k = tanh((W^k * x)_{ij} + b^k)$$

B. Overall Architecture

Now we describe the overall architecture of the proposed CNN model. As shown in Figure 2, the network contains five layers, the first four are convolutional-pooling layers and the remaining one is a fully connected layer.

The input of the first layer is original images in the dataset and consists of 3 features maps (an RGB color image) of size 128x128. The kernels of the second to the fourth layer are connected to the kernel maps in the previous layer and they are sparse connectivity. CNN exploits spatially-local correlation by enforcing a local connectivity pattern between neurons of adjacent layers. In other words, the inputs of hidden units in layer m are from a subset of units in layer m-1, units that have spatially contiguous receptive fields[2]. We can illustrate this graphically as Figure 3. The output of the last fully-connected layer is fed to a 100-way softmax function which produces a distribution over the 100 class labels.

The first convolutional layer filters the 128 x128 x3 input image with 30 kernels of size 11 x 11 x 3. The second convolutional layer takes as input the output of the first convolutional layer and filters it with 60 kernels of size 6 x 6 x 30. The third , fourth and fifth convolutional layers are connected to one another without any intervening pooling layers. The third convolutional layer has 120 kernels of size 6 x 6 x 60 connected to the outputs of the second convolutional layer. The fourth convolutional layer has 120 kernels of size 4 x 4 x 120, and the fifth convolutional layer has 120 kernels of size 4 x 4 x 120.The fully-connected layer has 1000 neurons.

[1]http://deeplearning.net/tutorial/

[2]http://deeplearning.net/tutorial/lenet

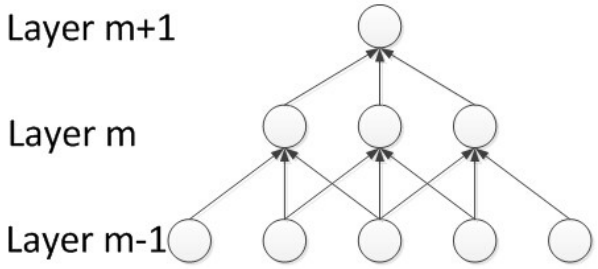

Figure 3. Units in layer m have receptive fields of width 3 in the input retina and are thus only connected to 3 adjacent neurons in the retina layer. Layer m+1 have 3 receptive fields with respect to the layer below, but their receptive field with respect to the input is larger (5)

III. Experiments on Recognition Accuracy

What we do first is enlarging the dataset because it is the easiest and the most common method to reduce overfitting on image data[4]. What's more, it's a necessary factor for the CNN to get a better learning result. There are two groups of controlled trials in our experiments: one is single fruit images vs multi-food images and the other is gray images vs RGB images. So we prepare four datasets by doing different preprocessing on the images we collect. For example, we perform flipping and blurring on the images to enlarge the dataset. In addition, it's necessary to resize the image for their ultimate, optimal use of size 128 x128. The original dataset we use is UEC-FOOD100 dataset[1] which is an open 100-class food image dataset and a fruit dataset that we establish. Figure 4 shows a part of images in the two datasets.

We use stochastic gradient decent to train our model with a batch size of 100 examples, initial learning rate of 0.01 and epoch of 200. A small learning rate is very important for the model to learn and it can reduce the model's training error. The classifier we use is logistic regression because this probabilistic, linear classifier is effective in the model. We train the network with the trainset of fifty thousand images, which takes about three hours on NVIDIA GTX TITAN 12GB GPU.

A. Results and Discussion

The two groups' results are shown in table I. Comparing with the error rate of gray multi-food images of 29.1%, gray fruit images have a smaller error rate of 19.2%. And RGB images have the same relationship of 46.5% and 15.5% respectively. When adding RGB color features, fruit images

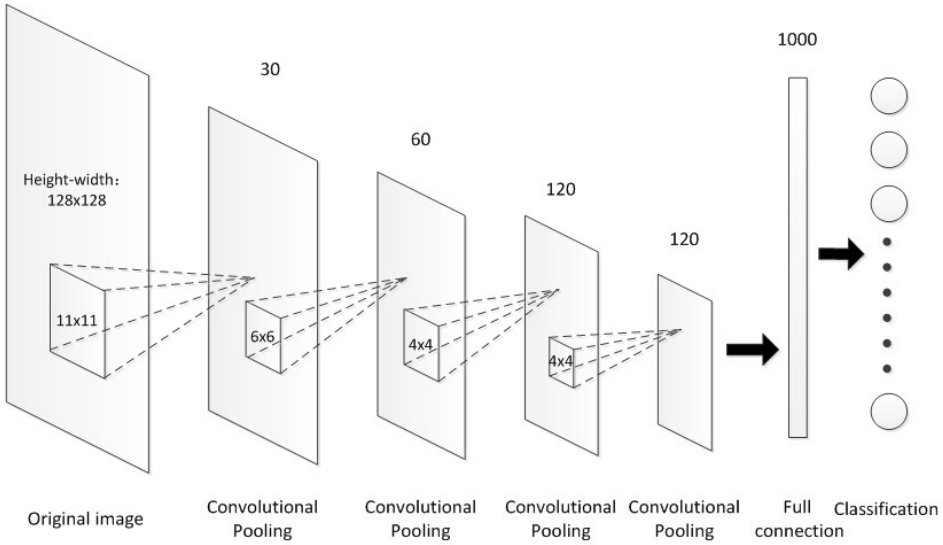

Figure 2. An illustration of the architecture of our CNN. Giving an image of size 128x128 as the input and it will give a lable as the output. The first to the fouth layer are convolutional-pooling layers and the fifth layer is fully connected layer which consists of 1000 neurons. The amount of convolutional kernel is 30, 60, 120, 120 respectively.

Figure 4. A part of images in the two datasets. The first 30 kinds are fruit images that are collected by ourselves and the last 30 kinds are multi-food images in the UEC-FOOD100 dataset.

have an obvious improvement of accuracy but the multi-food images have an unexpected outcome of smaller accuracy.

Table I
ERROR RATE OF THE TWO GROUPS OF CONTROLLED TRIALS

Dataset	Fruit images	Multi-food images
Gray	19.2%	29.1%
RGB	15.5%	46.5%

The first line shows the error rate of gray fruit images and gray multi-food images. The second line shows the error rate of RGB fruit images and RGB multi-food images.

Some reasons has been found after analyzing each dataset. RGB fruit images have a clear and single color feature, which can provide more information and effective reference for classifying. So the accuracy of the RGB fruit images can get a significant improvement. On the contrary, RGB multi-food images have not gotten an ideal result. An important point can been seen that multi-food images have an ambiguous color component and most images are yellow or white, which is almost helpless for classifying. The color information is not only an useless factor, but adding the information complexity to this dataset. And the network learns more about the colorful object such as a black bowl or a red cup but the white rice. What's more, the food materials which have been cut to pieces are so confused that even person can not detect them accurately, let alone the computer. Though the gray multi-food images are more complex and thus have a smaller accuracy than the gray fruit images, they possess certain effects. This may be due to the reason that the gray processing can chop away a part of interference information. An adding experiment is training and testing the network with a part of RGB multi-food images that are selected manually. We select some images that contain only one kind of food and are obvious to be

detected from UEC-FOOD100. Using this selective dataset, we finally achieve the best accuracy of 70.2%.

IV. Related work

Many research exist to help address food image recognition challenges. Parisa Pouladzadeh et al.[5] considered images' shape, color, size and texture characteristics with SVM as the classifier. A very high accuracy of 92.6% has been achieved with their own dataset including 12 kinds of fruit by this system. But there are only about 1600 images totally in the dataset, which is really a small one.

Kong and Tan[6] proposed a method using scale-invariant feature transform (SIFT) features clustered into visual words and fed to a simple Bayesian probabilistic classifier that matches the food items to a food database containing images of fast-food, homemade food, and fruits. A recognition performance of 92% was reported given that the number of references per food class in the database is larger than 50 and the number of food items to be recognized is less than six.

Yuji Matsuda et al.[1] proposed a two step method. The first step is detecting several candidate regions by fusing output of region detectors and the second step is fusing features including bag-of-features of SIFT and CSIFT with spatial pyramid (SP-BoF), histogram of oriented gradient (HoG), and Gabor texture features. As results, they have achieved the 55.8% classification rate, which is not a high accuracy.

Yoshiyuki et al.[8] presented a food image recognition method with deep conventional features and achieved 72.26% accuracy for the 100-class food dataset, UEC-FOOD100. Fanyu Kong et al.[7] introduced DietCam system utilizing a multi-view recognition method that separates every food by estimating the best perspective and recognizing them using a probabilistic method. The system achieved an average accuracy of 84% for the selective regular shape food.

Comparing the research above, our network can achieve a medial accuracy. Though some systems can achieve a high accuracy, they are based on a selective or a small dataset. Our experiment shows that a selective dataset can improve accuracy significantly.

V. Conclusion and Future Work

In this paper, we present a food image recognition method with CNN and achieve an accuracy of 80.8% for the fruit image dataset we establish ourselves and 60.9% for the UEC-FOOD100. By analyzing the results of the two group of controlled trials, we find that color information is not always helpful for improving the image accuracy. A reason why we have not gotten a high accuracy is the small dataset. In general, CNN does not work well for a small-scale dataset, while CNN works surprisingly well for a large-scale dataset[9].

For the future work, before recognition, we can process segmentation[10] on the image and it can work on CNN. However, a large-scale dataset is needed to train the CNN and even a well-trained network can not have a segmentation accuracy of 100%, which will enlarge the error rate of recognition. Therefore we can build a larger dataset by manual selection to get a better result.

Acknowledgment

The research is supported by the National Natural Science Foundation of China (Grant No. 61402533) and also Natural Science Foundation of Shandong Province (Grant No. ZR2014FM038), "Key Technologies Development Plan of Qingdao Technical Economic Development Area".

References

[1] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*. IEEE, 2012, pp. 25–30.

[2] M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. Mougiakkou, "A food recognition system for diabetic patients based on an optimized bag of features model," 2014.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] D. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[5] P. Pouladzadeh, G. Villalobos, R. Almaghrabi, and S. Shirmohammadi, "A novel svm based food recognition method for calorie measurement applications." in *ICME Workshops*, 2012, pp. 495–498.

[6] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012.

[7] ——, "Dietcam: Regular shape food recognition with a camera phone," in *Body Sensor Networks (BSN), 2011 International Conference on*. IEEE, 2011, pp. 127–132.

[8] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 2014, pp. 589–593.

[9] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[10] M. Wazumi, X.-H. Han, D. Ai, and Y.-W. Chen, "Auto-recognition of food images using spin feature for food-log system," in *Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference on*. IEEE, 2011, pp. 874–877.