2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340637

Distillation-based Chinese Food Ingredient Recognition and Nutrition Estimation System

Nan Zhang[1], Zhuer Le[1], Shiyin Jiang[2], Hong Cheng[3], Ling Wang*[2]

Abstract— **The foods' ingredients and nutrition are of great significance for human health so that people can meet their fitness needs or avoid consuming allergenic and post-operative contraindicated foods. However, the diversity of recipes and the randomness of combinations in Chinese cuisine make great challenges for Chinese food identification. To address the above issues, we built a new lightweight end-to-end food query and nutrition recognition system, which is based on knowledge distillation and deep learning methods. Firstly, well-performed DenseNet-121 is used to recognize the categories of food. At the same time, ResNet-50 is used as the Net-T, and pre-trained VGG-16 is used as the Net-S in the knowledge distillation framework, which is used to recognize the ingredients of the food. Finally, ingredient nutrition is obtained by querying the ingredient table. Experiments illustrate the good performance of the proposed method, with** 91.65% **Accuracy of food classification and** 92.01% **Accuracy of ingredients recognition.**

I. Introduction

The ingredients and nutrition of food play great significance in human health, especially for people with special food requirements, such as allergens, post-operative contraindicated foods, etc. Under these circumstances, it will greatly reduce the workload of medical care personnel if the food ingredients and nutrition can be automatically recognized from doctor-patient communication e-information. Therefore, an image-based intelligent ingredients recognition and nutrition estimation system is needed to be developed for users to utilize at any time.

The challenge of food recognition mainly comes from the extraction of features such as shape, color, texture, etc. The current food recognition is mostly based on deep learning recognition algorithms. Convolutional Neural Network (CNN) is applied to the task of food recognition in the multimedia community [1]. VGG-16 [2], ResNet-50 [3] and EfficientNet-B2 [4] are applied to automated food image classification. However, most of these works simply adopt CNN to extract visual features for food recognition.

Food ingredient recognition is generally a more challenging problem than food categorization. The size, shape, and color of an ingredient can exhibit large visual differences due to diverse ways of cooking and cutting, in addition to changes in viewpoints and lighting conditions. Multi-task and region-wise deep learning method [5], ingredient-guided cascaded multi-attention network [6] are proposed for food ingredient recognition. Vision Transformer (ViT) [7] is used for image recognition and obtains better results than CNNs by its self-attention mechanism. Most of these methods have conducted a series of studies for ingredient recognition, but the accuracy rate is still poor.

In recent years, distilled knowledge technique [8] provides good performance on recognition by migrating the results of large models (teacher models) and small models (student models). In this study, we combine knowledge distillation with CNNs to further improve ingredient recognition performance. And the nutritional composition of each ingredient is estimated by querying the nutrition table, which is based on the UK Food Standards Agency Nutrition and Analysis System Dietary Index (FSA-NPS DI) [9].

The framework of our proposed method is shown in Fig. 1. To start with, we choose DenseNet-121 as the model for food category recognition. As for ingredient recognition, on the other hand, we use the large model (ResNet-50) to train its ability to identify ingredients. Next, we condense the knowledge learned from the comparatively larger model into the smaller model (VGG-16) by knowledge distillation. A lightweight model of food categories and ingredient recognition is constructed by using distillation, which maintains the better performance of large models. Finally, we build food composition nutrition tables by querying FSA-NPS DI. The main contribution of this study is summarized as follows:

- A knowledge distillation-based framework is used to improve ingredient recognition performance on a a lightweight small models.
- A recognition and estimation system is constructed to provide information for any users who want to know information about what kind of ingredients and how healthy they consume.

II. Method

In this section, we introduce the proposed method in detail, including the description of the dataset, the knowledge distillation-based food recognition method, and the nutrition estimation method.

A. *Dataset*

In this study, the Chinese food set, VireoFood-251 [5], is used to verify the performance of our proposed food ingredient recognition and nutrition estimation method. There are

* Corresponding author

This study is supported by Sichuan Science and Technology Project (No.2022YFQ0008, No.2023YFS0044)

[1] Nan Zhang and Zhuer Le are with the School of Management and Economics, University of Electronic Science and Technology of China, 611731 Chengdu, China

[2] Shiyin Jiang and Ling Wang are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, 611731 Chengdu, China eewangling@uestc.edu.cn

[3]Hong Cheng is with the School of Automation Engineering, University of Electronic Science and Technology of China, 611731 Chengdu, China

979-8-3503-2447-1/23/$31.00 ©2023 IEEE

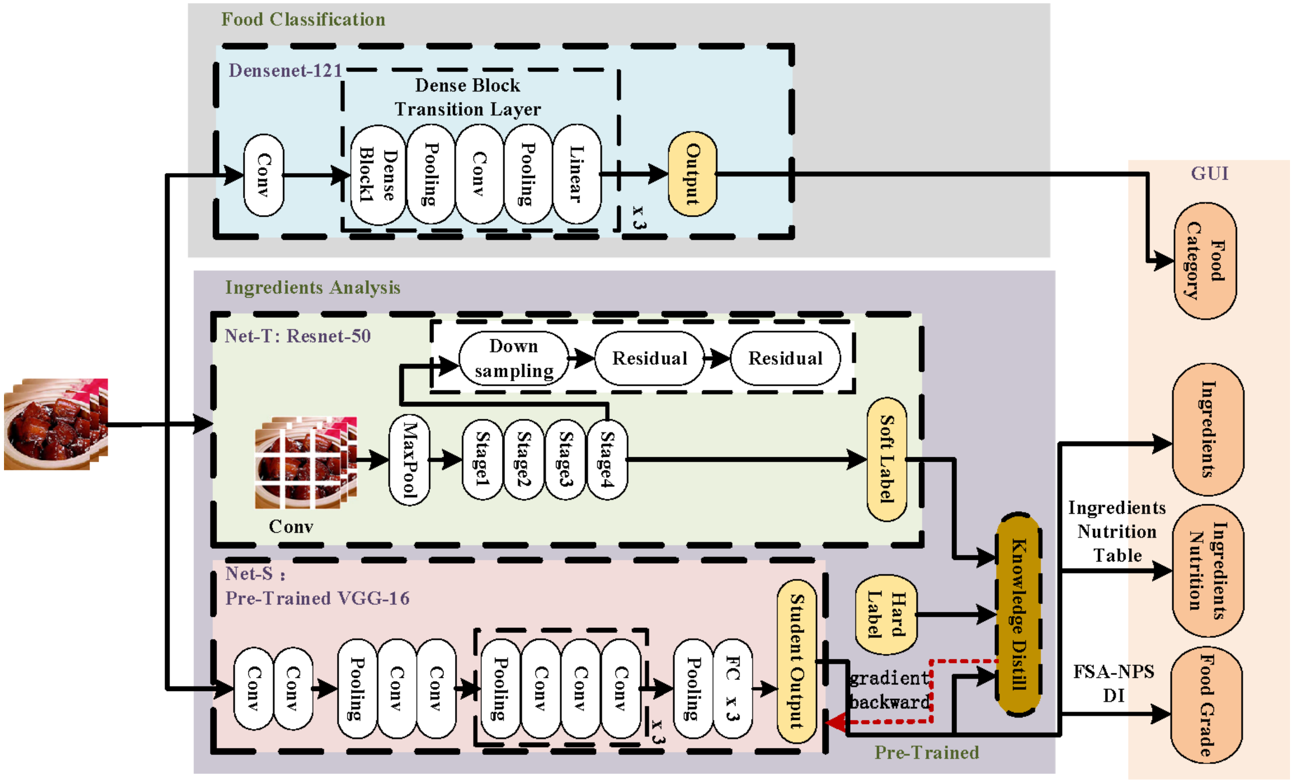


Fig. 1. Framework of the food category, ingredient recognition, and nutrition estimation system.

169, 673 food images from 251 categories that are annotated to 406 ingredients. The 251 categories cover 8 major groups of foods, including "Vegetables", "Soup", "Bean products", "Egg", "Meat", "Seafood", "Fish" and "Staple". The ingredients range from popular items such as "shredded pork" and "shredded pepper" to rare items such as "codonopsis pilosula" and "radix astragali". There are 3 ingredients in each dish on average.

B. Food Category and Ingredient Recognition

Knowledge distillation, a method of model compression, is a training method based on the "teacher-student network". The teacher network (Net-T) is the output of "knowledge" while the student network (Net-S) is the input of "knowledge". Net-T is characterized by a relatively complex model and can be integrated by several separately trained models. Net-S is a single model with a small number of parameters and a relatively simple model structure.

The main idea of Knowledge distillation is that Net-S can fit the soft label of Net-T so that the Net-S can learn some potential semantic information and summarize the experience of the Net-T. Specifically, A loss is made between Net-S and the soft labels. Since the output scales of the Net-T and the Net-S differ significantly, their outputs need to be softmax before the BCEWithLogits Loss is made. A BCEWithLogits Loss is made between the Net-S and the real hard label to understand the difference between the real data, and the two losses are combined by a weight to form the total loss. The loss function can be formulated as follows,

$$Loss = \alpha L(SM(y_s), SM(y_t)) + (1-\alpha)\ L(y_s, y), \quad (1)$$

where y is the true label, y_t is the output of the Net-T, y_s is the output of the Net-S, SM denotes softmax operation, $L(\cdot)$ is BCEWithLogits Loss function, α is a balance weighted.

In the knowledge distillation method, many deep learning methods can be used as Net-T and Net-S respectively. Take into consideration that, VGG network obtains good performance on fine-grained recognition by using small convolution kernels and deep layers, and ResNet solves the problem of model degradation in deep networks by using a residual unit. VGG-16 and ResNet-50 are used as Net-S and Net-T respectively.

VGG-16 uses several consecutive 3×3 convolution kernels and 16 network layers (13 convolutional layers and 3 fully connected layers). Dropout and L2 regularization are used for the first two fully connected layers to prevent over-fitting. The convolution kernel focuses on expanding the number of channels and the pooling kernel focuses on reducing the width and height, resulting in a deeper and wider model architecture. The output y_s can be formulated as follows,

$$y_s = SM(\mathbf{FC} \cdot \mathbf{W} + b), \quad (2)$$

where **FC** is the output of the Full Connective layer, **W** and b are the learned weights and biases.

The ResNet-50 uses identity and residual mappings to improve recognition performance, which is an improved version of VGG. The output y_t can be formulated as follows,

$$x_l = x_{l-1} + F(x_{l-1}, \mathbf{W}_{l-1}), \tag{3}$$

$$y_t = f(x_L), \tag{4}$$

where x_l is a characteristic expression of the l-th layer, $F(\cdot)$ is the activation function, generally using ReLU and the biases are omitted for simplifying notations. $f(\cdot)$ is the last layer operation which includes mean-pooling, full connectivity, and so on.

It is noted that, since the relatively simple issue of food category, the deep learning method is used directly as a model without knowledge distillation. Different from ResNet using deeper convolutional networks to improve training performance, DenseNet [10] uses feature reusing and bypassing strategies to alleviate the vanishing-gradient problem, strengthen feature propagation, and substantially reduce the number of parameters. Then DenseNet is used for food classification in this study.

C. Nutrition Estimation

The UK Food Standards Agency Nutrition and Analysis System Dietary Index (FSA-NPS DI) can adequately characterize the nutritional quality of foods. An FSA score is computed taking into account nutrient content per 100g for food and beverages. It allocates positive points for “unfavorable” content: energy (kJ), total sugar (g), saturated fatty acids (g), and sodium (mg), 0-10 points for each one. Negative points are allocated for “favorable” contents: fruits/ vegetables/nuts, fibers, and proteins, with 0-5 points for each one. The total of positive (0-40 points) and negative (0-15 points) points are computed, yielding a global score ranging from -15 for the most healthy foods $(-15, 0)$ to $+40$ for less healthy foods $(0, 40)$. From this overall score, five categories of nutritional quality are derived, defining the categories for the Nutri-Score, ranging from ‘green’ to ‘red’.

Based on FSA-NPS DI nutrient lists and scoring standards, food nutrition can be roughly estimated by a querying process. It is noted that a modification of FSA-NPS DI is used in this study to better describe the nutritional information of the dish. According to public data from FoodData Central of USDA and China’s domestic Food Safety Communication, the energy/kJ, sodium/mg, protein/g, total lipid/g, carbohydrates/g, and carbohydrates/g per 100g of each 406 ingredient are constructed. The Carbohydrates/g, Calcium/mg, Iron/mg, Total Ascorbic Acid/mg, and other common nutritional indicators are constructed into a nutritional table. Then, according to the relationship between the nutritional composition table and Nutri-Score, the nutritional quality is obtained. It is worth noting that since the system does not estimate the food volume and mass of each ingredient. It is a rough estimation of the intake of food ingredients and macro/micro-nutrients.

III. Experiments

A. Dataset Pre-processing and Parameters Setting

Considering that there are partial data missing in VireoFood-251 dataset, we selected 124 categories of food images, which with more than 400 images of each category. We divided the dataset into training and testing datasets with a rate of $7:1$, and then randomly disrupted them for subsequent experiments. Then, there is a total of $49,600$ images, of which $43,400$ are used as the training set and $6,200$ as the testing set. The feature size $N \times C \times W \times H$ is set as $43,400 \times 3 \times 224 \times 224$.

TABLE I
Experiment parameters of food classification.

Models	Epoch	Optimizer	Loss
DenseNet-121	80	30Epochs:SGD,lr=1e-5 50Epochs:Adam,lr=1e-5	Cross Entropy Loss
ResNet-50	22	Adam,lr=1e-5	
VGG-16	22		

TABLE II
Experiment parameters of food ingredients recognition.

Models	Epoch	Optimizer	Loss
VGG-16	62	Adam,lr=1e-6	MML
ResNet-50	22		
DenseNet-121	52	31Epochs:Adam,lr=1e-5 21Epochs:Adam,lr=1e-6	
EfficientNet-B0	27	Adam,lr=1e-5	BCE
ViT-b-16	80	40Epochs:SGD,lr=1e-6 40Epochs:Adam,lr=1e-5	

MML: MultiLabelSoftMarginLoss
BCE: BCEWithLogitsLoss

Experiments are performed on VGG-16, ResNet-50, DenseNet-121, EfficientNet-B0, and ViT-b-16 to compare the performance of our proposed framework.

For food classification, the output of its fully-connected layer is all set to 124 classes, the batchsize is set to 10, and the experiment parameters are shown in Table. I

As for ingredient recognition, the output of its fully-connected layer is all set to 406 classes, the batchsize is set to 10, and the experiment parameters are shown in Table. II. $\alpha = 0.5$ for (1).

B. Food Category and Ingredient Recognition

In this experiment, the food category and ingredient are recognized respectively. Firstly, we use several deep learning methods for food category and ingredient recognition directly. Then the methods with better performance are used in knowledge distillation framework for ingredient recognition.

DenseNet-121, VGG-16, ResNet-50, EfiicientNet-B0, and Vit-b-16 are used for food and ingredient recognition directly. The experiment results are shown in Table. III and Table. IV. Table. III shows that DenseNet-121 performs best performance on the classification problem. Form Table. IV, it can be seen that VGG-16 and ResNet-50 are obtained better performance than other methods. Then in the following experiments, VGG-16 and ResNet-50 are used as Net-T respectively, and the remaining models are used as Net-S.

It is noted that the Net-T is performed on PyTorch framework and uses migration learning to import pre-trained parameters which are trained on the Vireofood-251. However,

TABLE III

EXPERIMENT RESULTS OF FOOD CLASSIFICATION.

Model	Accuracy
DenseNet-121	**91.65%**
ResNet-50	86.58%
VGG-16	82.77%

TABLE IV

EXPERIMENT RESULTS OF FOOD INGREDIENTS RECOGNITION.

Models	Precision	Recall	F1
VGG-16	**86.21%**	82.90%	**84.52%**
ResNet-50	83.57%	**84.98%**	84.22%
DenseNet-121	76.96%	75.86%	76.41%
EfficientNet-B0	78.4%	81.78%	80.05%
ViT-b-16	75.31%	66.23%	70.47%

TABLE V

EXPERIMENT RESULTS OF DISTILLATION-BASED FOOD INGREDIENTS RECOGNITION.

Net-T	Net-S	Precision	Recall	F1
ResNet-50	DenseNet-121	76.75%	74.62%	75.67%
	EfficientNet-B0	84.03%	86.94%	85.46%
	ViT-b-16	84.81%	84.29%	84.55%
	VGG-16	**92.01%**	**91.80%**	**91.90%**
VGG-16	DenseNet-121	77.84%	77.37%	77.63%
	EfficientNet-B0	84.75%	88.22%	86.45%
	ViT-b-16	85.06%	84.39%	84.72%
	ResNet-50	84.87%	85.73%	85.30%

the Net-T is imported via torchvision, the Net-S is imported through the structures we have constructed.

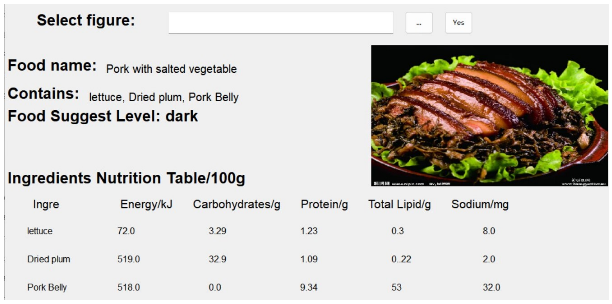
From Table. V, we can see that, the best ingredient recognition results are obtained when ResNet-50 is worked as Net-T and VGG-16 is worked as Net-S. There is a 6.6% improvement in F1 score than instead setup, and there is almost 16.2% improvement than DenseNet does. The reason is that VGG can find more fine-grained features than other methods by its smaller convolution kernels, and the ResNet can learn more general features in lower sample size dataset than others. The knowledge distillation improves the generalization ability of ResNet.

C. Nutrition Estimation

We build a Graphical User Interface (GUI) to display the ingredients and nutrition of food based on the Python library wxPython. When GUI startup, the distillation-based trained model is loaded automatically. And then, the food and ingredient classification method are performed when a food image is selected. The food nutrition is obtained by querying the nutritional composition table, and Nutri-Score is computed. The recognition results and its nutrition and score are displayed in Fig. 2. It is intuitive and convenient.

IV. CONCLUSION

In this study, we provide a new lightweight deep neural network-based food ingredient recognition and nutrition estimation system by using the knowledge distillation technique. The system identifies the type of food and its ingredients, gives the nutritional information of each ingredient, and gives a preliminary rating of the nutritional value based on FSA-NPS DI. Firstly, this nutrition estimation system provides end-to-end recognition and estimation for food querying. Secondly, the knowledge distillation technique is used to improve ingredient recognition performance while making the network model lightweight. It solves the contradiction between model size and performance. Furthermore, we construct a nutrition table of 406 common ingredients, which can be queried and provides the nutrition of each identified ingredient in real-time. In future work, we will discuss the relationship between food category and ingredients, so as to improve the ingredient recognition performance. We will conduct further experiments on a non-standardized dataset to improve the performance, which is closer to the images used by people in daily life.



Select figure:

Food name: Pork with salted vegetable

Contains: lettuce, Dried plum, Pork Belly

Food Suggest Level: dark

Ingredients Nutrition Table/100g

Ingre	Energy/kJ	Carbohydrates/g	Protein/g	Total Lipid/g	Sodium/mg
lettuce	72.0	3.29	1.23	0.3	8.0
Dried plum	519.0	32.9	1.09	0..22	2.0
Pork Belly	518.0	0.0	9.34	53	32.0

Fig. 2. Graphical User Interface of Food Ingredient Recognition System.

REFERENCES

[1] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1085–1088.

[2] S. Yadav, S. Chand *et al.*, "Automated food image classification using deep learning approach," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1. IEEE, 2021, pp. 542–545.

[3] Z. Zahisham, C. P. Lee, and K. M. Lim, "Food recognition with resnet-50," in *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. IEEE, 2020.

[4] F. S. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "Mediterranean food image recognition using deep convolutional networks," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 1740–1743.

[5] J. Chen, B. Zhu, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "A study of multi-task and region-wise deep learning for food ingredient recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 1514–1526, 2020.

[6] W. Min, L. Liu, Z. Luo, and S. Jiang, "Ingredient-guided cascaded multi-attention network for food recognition," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[9] C. Julia and S. Hercberg, "Nutri-score: Evidence of the effectiveness of the french front-of-pack nutrition label," *Ernahrungs Umschau*, vol. 64, no. 12, pp. 181–187, 2017.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.