A Hierarchical deep model for food classification from photographs

Heekyung Yang[1], Sungyong Kang[2], Chanung Park[2], JeongWook Lee[2], Kyungmin Yu[2] and Kyungha Min[2]

[1] Dept. of Computer Science, Graduate School, Sangmyung Univ.,
Hongji-dong, Jongro-gu, Seoul, Korea
[e-mail: yhk775206@naver.com]
[2] Dept. of Computer Science, Sangmyung Univ.,
Hongji-dong, Jongro-gu, Seoul, Korea
[e-mail: minkh@smu.ac.kr]
*Corresponding author: Kyungha Min

Received May 7, 2019; revised October 9, 2019; accepted October 27, 2019; published April 30, 2020

Abstract

Recognizing food from photographs presents many applications for machine learning, computer vision and dietetics, etc. Recent progress of deep learning techniques accelerates the recognition of food in a great scale. We build a hierarchical structure composed of deep CNN to recognize and classify food from photographs. We build a dataset for Korean food of 18 classes, which are further categorized in 4 major classes. Our hierarchical recognizer classifies foods into four major classes in the first step. Each food in the major classes is further classified into the exact class in the second step. We employ DenseNet structure for the baseline of our recognizer. The hierarchical structure provides higher accuracy and F1 score than those from the single-structured recognizer.

Keywords: CNN, DenseNet, classification, food, dietetics

http://doi.org/10.3837/tiis.2020.04.016 **ISSN : 1976-7277**

1. Introduction

Recognizing objects and classifying them into predefined classes is one of the most interesting research topics in machine learning. Many classical techniques such as support vector machine (SVM) attack this problem. Recently, the deep learning techniques based on convolutional neural network (CNN) accelerate the progress of the recognition techniques. Nowadays, the domain of recognition is extended to various practical areas such as food and dietetic items.

A technique that recognizes and classifies foods from their photographs has been a long research issue in pattern recognition and machine learning society. The challenging problem in developing a food recognition technique comes from the fact that food is a cultural product of a country. Therefore, foods from different countries are very different and this explains why the existing food recognition techniques concentrate on food of a single country such as Japan or China.

In this paper, we focus on Korean food, which has a long tradition and it has various kinds of recipes. To our knowledge, a deep learning-based recognition technique for Korean food has not been studied yet. The most distinguishing point of Korean food is that we can classify them into four major categories: rice-based meal, soup or stew, main dish and side dish. The main dish is a food, which is cooked for each meal and the side dish is a food that lasts several days or weeks. As illustrated in **Fig. 1**, an ordinary Korean meal is composed of a rice, a soup or stew, one or two main dishes and several side dishes. In this paper, we build 18 Korean food classes as illustrated in **Fig. 2**.

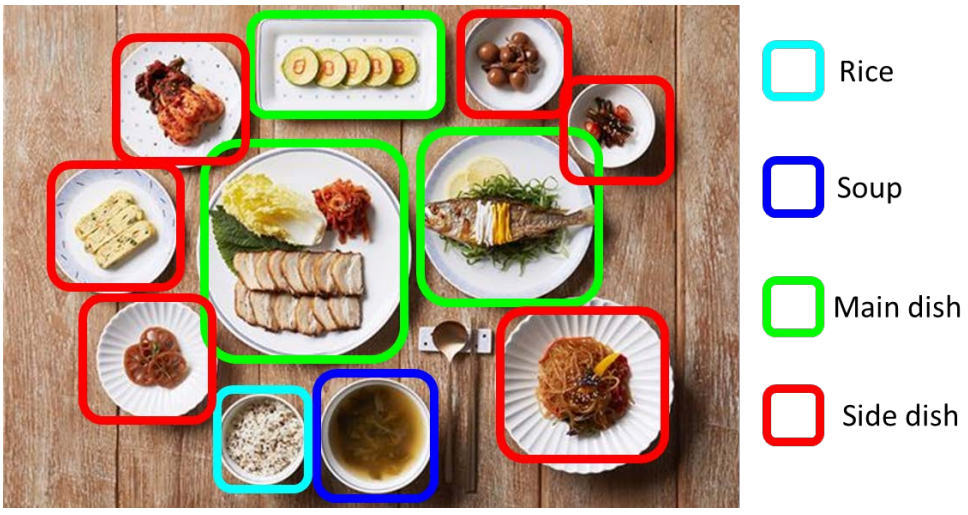

Fig. 1. Conventional Korean food table composed of foods of four categories: rice, soup, main dish and side dish.

We design a classifier for Korean food by employing a deep convolutional neural network. Our classifier is designed as a double-layered architecture to accommodate the hierarchical structure of Korean food. The first layer of our classifier has one network that classifies

Korean food into four major categories: rice, soup, main dish and side dish. The second layer of our classifier has four networks that classifies Korean food in each major category. The convolutional neural network we employ is the DenseNet whose last layer is substituted with a fully convolutional layer instead of fully connected layer.

A fundamental requirement for our hierarchical approach is a very high accuracy of the first-layer classifier. In the case of Korean food, the foods in different categories are easily distinguishable. Our double-layered structure is motivated from this characteristic of Korean food. In our research, we compare two classifiers. One classifier is a widely-used single-layered classifier that classifier our target 18 foods simultaneously. The other classifier is our double-layered classifier that classifies our target foods in two stages. With our experiment, we show the excellence of our double-layered classifier.

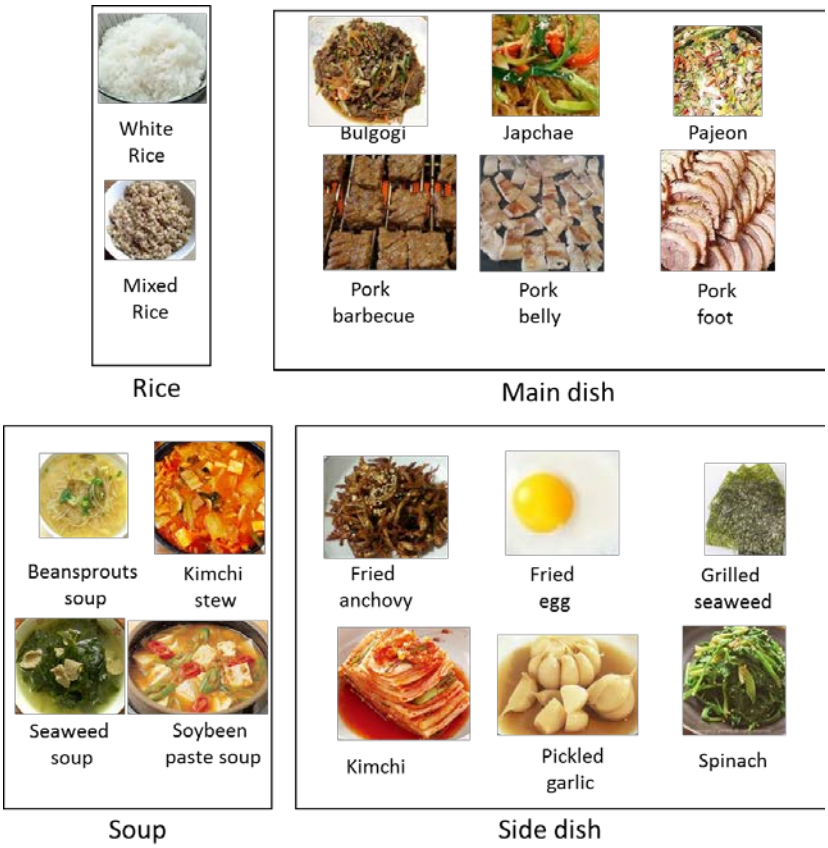

Fig. 2. Our 18 target Korean foods to classify.

The contribution of this paper is as follows:
(1) We propose a hierarchical structure composed of CNN models that classifies food images in two-way categories. Even though our approach employs existing CNN models, the hierarchical structure of the models for food recognition is a novel approach.
(2) Many existing CNN models classify general targets. In this paper, we focus on food images. We compare many widely used CNN models and select one that shows best performance in food image classification.

(3) Among many local foods, we present a scheme for classifying Korean food, which has not been classified yet.

This paper is organized as follows. In Section 2, we survey the related works about food recognition techniques. In Section 3, we suggest the structure of our classifier and explain our plan of experiments in Section 4. In Section 5, we describe our results and analyze the results. In Section 6, we conclude our work and propose a future research plan.

2. Related Work

2.1 Conventional schemes

The early-stage food recognition research is based on pattern recognition techniques using various features extracted from images. Chen et al. [1] constructed dataset on fast food images and classified them using color histogram and bags of SIFT features. Joutou and Yanai [2] applied a multiple kernel learning approach to merge various features such as color, texture and SIFT. This scheme, however, depends on the camera angle for the images very heavily. Wu and Yang [3] presented a food recognition scheme from a video clip. This scheme also presents calorie calculation for the recognized food. Yang et al. [4] developed a food recognition scheme that applies statistical techniques on local features. Instead of using food image, this scheme analyzes the ingredients of the food to recognize the image. Even though this scheme, which uses SIFT, color histogram and bag of features, shows excellent performance on many food images, it fails to recognize food of similar ingredients such as bread, donut and bagel. Bosch et al. [5] identifies food by combining various features in global and local scale. The applied decision fusion-based rules to improve the accuracy by 7%. Chen et al. [6] presented a food identification scheme that estimates the quantity of the food simultaneously. They implemented their scheme on various mobile platforms such as Android and iOS. Matsuda et al. [7] proposed a scheme that recognizes multiple foods from an image. They build a series of candidate positions for food from a photograph and apply feature-fusion-based scheme for the recognition. This scheme that recognizes multiple food simultaneously fails to present high accuracy. Bossard et al. [8] presented a food recognition scheme from an image using random forests. They show high accuracy for MIT-indoor dataset.

2.2 CNN-based schemes

Kagaya et al. [9] presented an image-based food detection and recognition framework using a convolutional neural network (CNN), which shows higher accuracy than the conventional SVM-based frameworks. They aim to classify 10 most-frequent food items on Food Log (FL). Kawano and Yanai [10] employed the features extracted from deep CNN to increase the accuracy of food recognition. They report their top-1 accuracy on UEC-FOOD100 dataset about 72.26%. The parameters of their network are too many to be used in various platforms such as mobile devices. Kawano and Yanagi [11] presented a scheme that automatically expands food image dataset. They leverage the existing knowledge on food of one culture using a generic classifier and domain adaptation. Kagaya and Aizawa [12] presented a framework that distinguishes food and non-food image from thee datasets including Instagram, Food-101, and Caltech-256. They build their framework using a CNN technique and show their accuracy about 95%. Yanai and Kawano [13] tuned a pre-trained deep convolutional

network for food image recognition. They prove that the finely tuned deep CNN is effective on small-scaled food image dataset. They also come across the practical limitation of using their deep CNN on various platforms. Pouladzadeh et al. [14] developed a framework that detects the position of a food from an image using a graph cut scheme and deep neural network. They claim that the ingredients of a food plays a key role in food detection.

2.3 Recent works

Recently, Akbari Fard et al. [15] presented a CNN-based classification scheme for food and vegetable. They collected images from ImageNet and classified 43 classes with 75% top-5 accuracy and 45% top-1 accuracy. Tatsuma and Aono [16] presented a food recognition scheme using a covariance measured on the feature maps extracted by a convolutional network. They record 58.65% accuracy on ETHZ Food-101 dataset. Hassannejad et al. [17] employed a 54-layered deep convolutional network for food image recognition. They apply their scheme on various food image datasets including ETH Food-101, UEC FOOD 100 and UEC FOOD 256. They show more than 92% top-5 accuracy on the datasets. They conclude that the depth of the network is a key factor that influence the accuracy. Jain and Khanna [18] presented a classification scheme on INDIAN Krishina Kamod rice using a neural network. They employ geometric features such as minor axis length, major axis length and eccentricity and show 97% accuracy. Ragusa et al. [19] compared various classification schemes about food image and non-food image. They report that the combined model of fine-tuned Alexnet and binary SVM shows best accuracy as 99.59% and that VGGNet shows best accuracy in classifying different images. Singla et al. [20] developed a food/non-food image classification scheme using a pre-trained GoogLeNet model. This model shows 99.2% accuracy for classifying food and non-food image. They claim their limitation about classifying images of multiple foods. Farooq and Sazonov [21] presented a scheme that recognizes food types by using the feature maps extracted from a pre-trained AlexNet. They show 94.01% accuracy in classifying 61 classes of food types.

3. Structure of our classifier

Our classifier for Korean food has a two-layered structure. The first layer has a CNN-based module that classifies Korean foods into four major categories: rice, soup, main dish and side dish. The second layer has four CNN-based modules, each of which classifies the foods in the four major categories. The overall structure of our classifier is suggested and compared with a conventional classifier in **Fig. 3**.

We design he CNN-based model of our classifier based on a DenseNet with a fully convolutional layer. Fig. 4 illustrates the structure of DenseNet with a fully convolutional layer. According to [28], a deep CNN with fully convolutional layer shows better performance than a deep CNN with fully connected layer. Global Average Pooling (GAP), which lies between conv layer and fully convolutional layer, produced smoother contours than global max pooling (GMP). Another benefit of the fully convolutional layer is a class activation map (CAM) that visualizes the clue of classification with different colors.

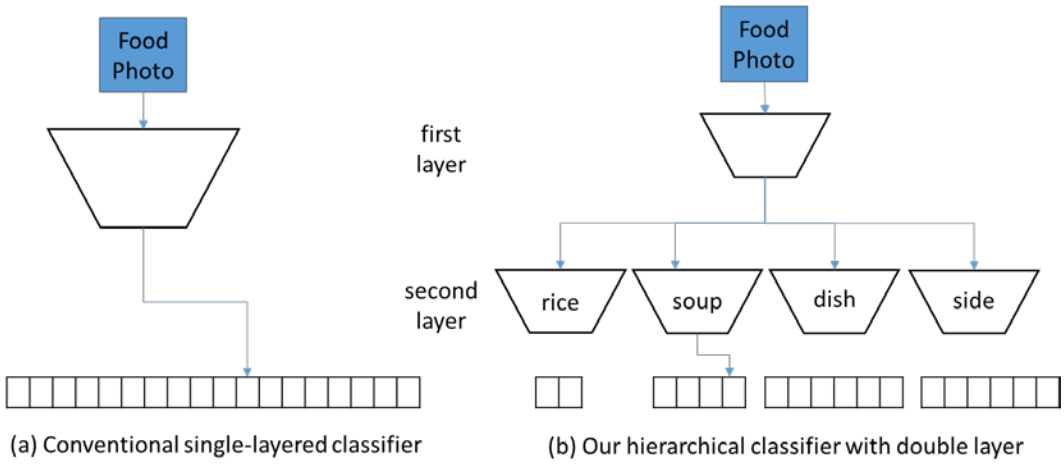

Fig. 3. Our classifier with double layer is compared with the conventional classifier with single layer. The trapezoid is a classifying module. We test several existing models including AlexNet, VGGNet, ResNet and DenseNet and select DenseNet, which shows highest accuracy.

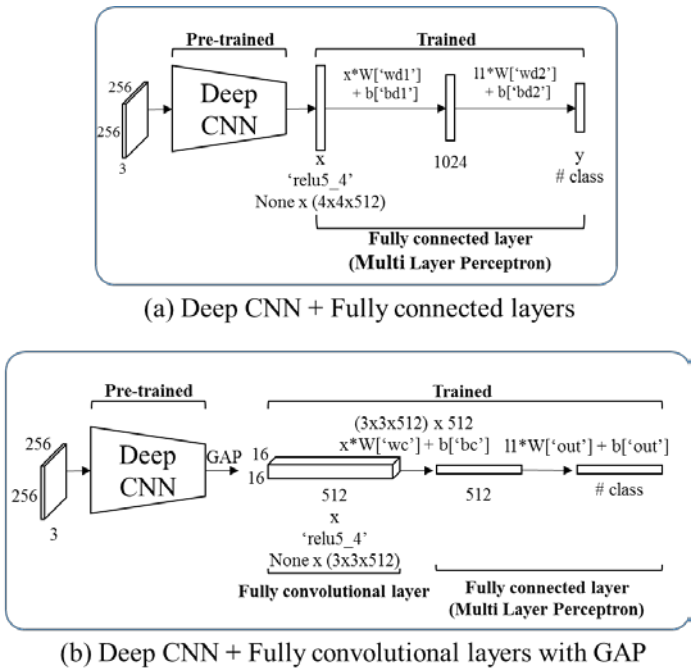

Fig. 4. Comparison of deep CNN with fully connected layers (a) and with fully convolutional layers with GAP (b).

4. Dataset collection

We collect the photographs of our target 18 Korean foods from various websites using data crawling techniques. We discard photographs that contain other objects such as tablewares to avoid training error. We collect 6452 photographs for 18 food classes. Among the dataset, we

prepare 70% of them for training, 15% for validation and 15% for test. We suggest the details of the dataset in **Table 1**.

Table 1. The dataset for our 18 Korean food classes.

Category	Food	Total	Training	Validation	Test
Rice	White rice	243	170	36	37
	Mixed rice	372	260	56	56
Soup	Beansprouts soup	300	210	45	45
	Kimchi stew	200	140	30	30
	Seaweed soup	301	211	30	30
	Soybeen paste soup	265	186	40	39
Main dish	Bulgogi	349	244	52	53
	Japchae	345	242	52	51
	Pajeon	404	283	61	60
	Pork barbecue	372	260	56	56
	Pork belly	426	298	64	64
	Pork foot	232	162	35	35
Side dish	Fried anchovy	494	246	74	74
	Fried egg	261	183	39	39
	Grilled seaweed	404	283	61	60
	Kimchi	789	552	118	119
	Pickled garlic	346	242	52	50
	Spinach	349	244	52	53
Total		6452	4517	967	968

5. Experiment

5.1 Implementation

We implement our classifier on a personal computer with Pentium i7 CPU, 32 GByte main memory and nVidia TitanX GPU. Our classifier is developed using Tensorflow and CUDA. Our hierarchical classifier is designed by composing five classifiers each of which classifies objects into 2 ~ 6 classes. To build a classifier, we employ four recently popular classifiers: AlexNet [23], VGGNet [22], ResNet [25] and DenseNet [26]. We also employ the pretrained parameters from these networks. These models are trained by ImageNet datasets. After applying them to the first layer classifier and comparing their performances in both fully connected structure and fully convolutional structure, we find out that DenseNet outperforms other compared networks. Therefore, we employ DenseNet for our classifier. The comparison is suggested in **Table 2**.

An interesting characteristic of Korean food is that it can be classified into several disjoint major categories. We set the major categories as rice, soup, main dish and side dish. For our classification, we select 18 Korean based on Korean food and dietetics survey and arrange them in the four major categories.

To train our classifier, we collected the images of our target Korean foods from various internet sites. In total, we collect 6452 images, which means that the images for each food are about 358 in average. Among the collected images, we assign 70% of them for training, 15% for validation and 15% for test. The exact amounts of the collected images are suggested in Table. 1. We train our classifier for 100 epochs. We set dropout rate as 70% and learning rate as 0.0001.

Table 2. The comparison of four widely-used deep CNN structures where DenseNet shows best performance: (a) shows the results and hyper parameters of fully convolutional networks and (b) shows those of fully connected networks.

Fully convolutional								
	F1 score	accuracy(%)			Hyper parameter			
	test	validation	test	train	depth	# of nodes	dropout ratio	epochs
DenseNet	**0.97**	**97.10**	**97.63**	**98.25**	**1**	**256**	**0.7**	**80**
ResNet	0.95	95.97	95.26	97.85	1	2048	0.3	90
VGGNet	0.92	93.80	92.37	95.75	1	2048	0.5	60
AlexNet	0.92	92.97	91.55	95.57	1	2048	0.3	70

(a) The accuracies and hyper parameters of fully convolutional structure

Fully connected								
	F1 score	accuracy(%)			Hyper parameter			
	test	validation	test	train	depth	# of nodes	dropout ratio	epochs
DenseNet	**0.96**	**95.97**	**95.98**	**96.35**	**1**	**256**	**0.3**	**50**
ResNet	0.90	89.87	90.10	91.05	1	128	0.5	100
VGGNet	0.94	94.11	93.09	96.50	1	2048	0.5	80
AlexNet	0.92	94.52	92.27	96.77	1	2048	0.3	90

(b) The accuracies and hyper parameters of fully connected structure

5.2 Implementation on mobile environment

We implement our classifier on a mobile environment. The mobile device plays role of client, which presents user interface, grabs food images for classification and communicates with server. Server communicates with clients via node.js interface. It executes our classifier and sends five most possible cases with their probabilities to the client. For a reliable classification, the result of first layer is confirmed by user. Finally, information of nutrition for the recognized food is presented. **Fig. 5** presents the structure of the mobile environment and **Fig. 6** shows the screenshots of using mobile classifier.

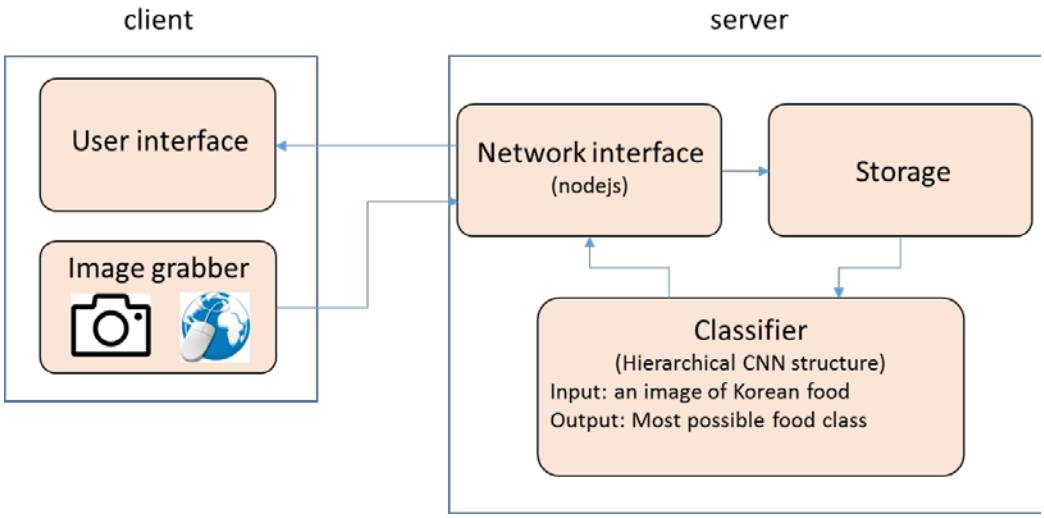

Fig. 5. The structure of mobile implementation of our classifier.

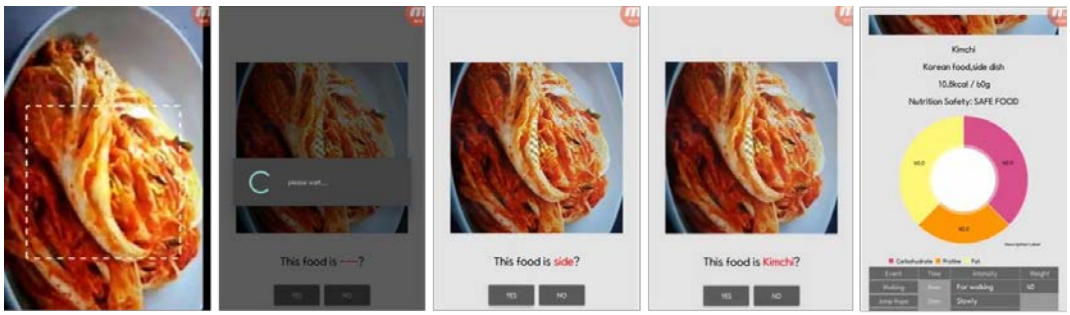

Fig. 6. The execution sequence of our mobile food classifier: (a) Taking a photo of a food, (b) Processing in server, (c) The result of our first layer, (d) The result of our second layer, (e) Nutrition information for the recognized food class.

6. Result and analysis

6.1 Result

We classified the food photos in the dataset in two classifiers and compared the results. The first one is a single-layered classifier and the second one is our double-layered classifier. In **Table 3**, we suggest the confusion matrix of the first classifier, which classifies the input photos into 18 categories. We present two confusion matrices for the second classifier, since it classifies the input photos in two stage. The confusion matrix in **Table 4** (a) is the result of the first stage, which classifies the input photos into four major categories, and the confusion matrix in **Table 4** (b) is the result of the second stage, which classifies the input photos into 18 categories. In the second stage, the classes that belong to the major categories are only considered.

Table 3. The confusion matrix for the 18 food classes classified by the single-layered conventional classifier: Row is the predicted class and column is the true class.

	Mixed Rice	White Rice	BeanSprout Soup	Kimch Stew	Seaweed Soup	SoyBean PasteStew	Bulgogi	Japchae	Pajeon	Pork Barbecue	Pork Belly	Pork Feet	Fried Anchovy	Fried Egg	Grilled Seaweed	Kimchi	Pickled Garlic	Spinach
Mixed Rice	0.95	0	0	0	0	0	0	0	0	0	0	0	0.04	0	0.02	0	0	0
White Rice	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BeanSprout Soup	0	0	0.98	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0
Kimch Stew	0	0	0	0.97	0	0	0	0	0	0	0	0	0	0	0	0.03	0	0
Seaweed Soup	0	0.02	0.02	0	0.95	0	0	0	0	0	0	0	0	0	0	0	0	0
SoyBean PasteStew	0	0	0	0	0.01	0.97	0.02	0	0	0	0	0	0	0	0	0	0	0
Bulgogi	0	0	0.02	0	0	0	0.91	0.02	0	0	0	0	0.04	0	0	0	0	0
Japchae	0	0	0.02	0	0	0	0.02	0.85	0	0	0	0.04	0.06	0	0.02	0	0	0
Pajeon	0	0	0	0	0	0	0.05	0	0.9	0	0.02	0	0	0	0	0.03	0	0
Pork Barbecue	0	0	0	0.02	0.02	0	0.08	0.02	0	0.8	0.06	0	0	0	0	0	0	0
Pork Belly	0.03	0	0	0	0.01	0	0	0	0	0.07	0.86	0.03	0	0	0	0	0	0
Pork Feet	0	0	0	0	0	0	0	0	0	0.03	0.03	0.94	0	0	0	0	0	0
Fried Anchovy	0	0	0	0	0	0	0.01	0.06	0	0.01	0	0	0.87	0	0	0.01	0	0.03
Fried Egg	0	0.09	0	0	0	0	0	0	0	0	0	0	0	0.91	0	0	0	0
Grilled Seaweed	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0.99	0	0	0
Kimchi	0	0	0	0.02	0	0	0	0	0.02	0	0	0	0	0	0	0.96	0	0
Pickled Garlic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Spinach	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0	0	0	0.98

6.2 Comparison

We compare the results of our model with those from DenseNet, one of the state-of-the-art image classification methods. In **Table 2**, we present the accuracies of DenseNet: 0.97 for fully convolutional structure and 0.96 for fully connected structure. In **Table 5**, the accuracy of our model records 0.99. For F1 score, fully convolutional DenseNet records 0.97 and fully connected DenseNet records 0.96. Our model records 0.97 F1 score in **Table 5**. Our model shows better accuracy than DenseNet and similar F1 score with DenseNet.

6.3 Analysis

We compute the accuracy, precision, recall and f1 score for two classifiers. As illustrated in **Table 5**, our double-layered classifier shows better accuracies for 16 food classes, better precision for 13 food classes, better recall for 14 food classes and better f1 score for 16 food classes. Our double-layered classifier shows worse results only for 5 food classes. This result is visualized in **Fig. 7**.

6.4 Analysis with CAM

By substituting the last layer of the classifier from fully connected layer to fully convolutional layer (See **Fig. 4**), we can class activation map (CAM), which visualize the region that influences the classification. The red denotes higher influence and the blue denotes lower influence. In **Fig. 8**, we illustrate two very confusing classes: Kimchi stew and Kimchi. The input images in **Fig. 8** are misclassified by the single-layer classifier, but they are classified correctly by our double-layer classifier. The different CAMs of the classifiers denote the region that gives clue for the classification is different for the classifiers.

6.5 Limitation

The success of a hierarchical classifier depends on the accuracy of the first layer. The final accuracy is a multiplication of the accuracies of the first layer and the second layer. Therefore, the final accuracy may be lower than the single layer classifier even though the second layer shows better accuracy than the single layer classifier. Fortunately, the accuracy of the first layer that classifies Korean food into four major categories is close to 1.0. In other cases where the first layer does not show very high accuracy, a hierarchical structured classifier may show worse accuracy than a single layer classifier.

6.6 Direction of improvements

One of our major limitations is the deficiency of dataset. To extend our approach to other categories of food, we have to extend the dataset. Song et. al [27] presented a scheme that improves the performance of supervised learning when tagged training data is not sufficient. In many problems, a geometry-based regularization term is added to match untagged data to their similar tagged data. They proposed a semi-supervised annotation approach by learning an optimized graph from multiple cues, such as partial tags and multiple features to improve the accuracy of the matching. This approach can help extending the dataset of our approach.

Another limitation is that the source of food photos can come from video. To properly capture and caption food on video, Song et al.'s scheme [29] can be employed. They proposed a generative approach that models the uncertainty in the data using latent stochastic variables. Their scheme, denoted as multimodal stochastic RNN (MS-RNN), can improve the performance of video captioning and generate multiple labels to describe a video. This scheme will tag many food scenes from food video to enrich our dataset.

Wang et al. [30] also presented a scheme that recognizes scenes from video of arbitrary sizes. They decomposed a video into spatial and temporal shots, and a sequence of shots are processed using a spatial temporal pyramid pooling (STPP) convNet with a long short-term memory or CNN-E model. Their softmax scores are combined by a late fusion. The STPP convNet extracts descriptions for each variable-size shot, and CNN-E learns a global description for the input video.

Table 4. The confusion matrices for our experiment.

	rice	soup	dish	side
rice	1	0	0	0
soup	0	0.99	0.01	0
dish	0	0	0.99	0.01
side	0	0	0	1

(a) The confusion matrix of the first-layer that classifies food photos into four major classes

	Mixed Rice	White Rice	BeanSprout Soup	Kimch Stew	Seaweed Soup	SoyBean PasteStew	Bulgogi	Japchae	Pajeon	Pork Barbecue	Pork Belly	Pork Feet	Fried Anchovy	Fried Egg	Grilled Seaweed	Kimchi	Pickled Garlic	Spinach
Mixed Rice	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
White Rice	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BeanSprout Soup	0	0	0.98	0.01	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0
Kimch Stew	0	0	0	0.99	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0
Seaweed Soup	0	0	0	0	0.99	0	0	0	0	0	0	0.01	0	0	0	0	0	0
SoyBean PasteStew	0	0	0	0.01	0.01	0.97	0.01	0	0	0	0	0	0	0	0	0	0	0
Bulgogi	0	0	0	0	0	0	0.92	0.07	0	0	0	0	0.01	0	0	0	0	0
Japchae	0	0	0	0	0	0	0.02	0.95	0	0	0	0.02	0	0	0.01	0	0	0
Pajeon	0	0	0	0	0	0	0.04	0.02	0.91	0.02	0	0	0	0	0	0	0.01	0
Pork Barbecue	0	0	0	0	0	0	0.04	0.02	0	0.85	0.06	0.02	0	0	0	0	0	0.01
Pork Belly	0	0	0	0	0	0	0.01	0	0	0.08	0.89	0.01	0.01	0	0	0	0	0
Pork Feet	0	0	0	0	0	0	0	0	0	0.03	0.01	0.95	0	0	0	0	0.01	0
Fried Anchovy	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0	0.01	0	0	0
Fried Egg	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Grilled Seaweed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Kimchi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Pickled Garlic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Spinach	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

(b) The confusion matrix of the second-layer that classifies food photos into 18 classes

Table 5. The comparison of the performances with single-layered classifier and our hierarchical double-layered classifier: For 18 classes and four metrics including accuracy, precision, recall, and f1 score, our classifier shows better result in 58 cases, same result in 9 cases and worse result in 5 cases.

	Accuracy			Precision			Recall			f1 score		
	single	double		single	double		single	double		single	double	
Mixed rice	1.00	1.00	0.00	0.97	1.00	0.03	0.95	1.00	0.05	0.96	1.00	0.04
White rice	0.99	1.00	0.01	0.90	1.00	0.10	1.00	1.00	0.00	0.95	1.00	0.05
Bean Sprout Soup	0.98	1.00	0.02	0.93	1.00	0.07	0.98	0.98	0.00	0.96	0.99	0.03
Kimchi Stew	0.98	1.00	0.02	0.96	0.98	0.02	0.97	0.99	0.02	0.97	0.99	0.02
Seaweed Soup	0.98	1.00	0.02	0.95	0.99	0.04	0.95	0.99	0.04	0.95	0.99	0.04
Soybean Paste Stew	0.98	1.00	0.02	1.00	1.00	0.00	0.97	0.97	0.00	0.98	0.98	0.00
Bulgogi	0.97	0.99	0.02	0.91	0.88	-0.03	0.91	0.92	0.01	0.87	0.90	0.03
Japchae	0.96	0.98	0.02	0.89	0.90	0.01	0.85	0.95	0.10	0.87	0.92	0.05
Pajeon	0.96	0.99	0.03	0.98	0.99	0.01	0.90	0.91	0.01	0.94	0.95	0.01
Pork Barbaque	0.95	0.98	0.03	0.88	0.87	-0.01	0.80	0.85	0.05	0.84	0.86	0.02
Pork Belly	0.96	0.98	0.02	0.89	0.93	0.04	0.86	0.89	0.03	0.87	0.91	0.04
Port Feet	0.97	0.99	0.02	0.93	0.94	0.01	0.94	0.95	0.01	0.94	0.95	0.01
Fried Anchovy	0.97	1.00	0.03	0.83	0.98	0.15	0.87	0.99	0.12	0.85	0.99	0.14
Fried Egg	0.98	1.00	0.02	1.00	1.00	0.00	0.91	1.00	0.09	0.95	1.00	0.05
Grilled Seaweed	0.99	1.00	0.01	0.97	0.98	0.01	0.99	1.00	0.01	0.98	0.99	0.01
Kimchi	0.99	1.00	0.01	0.92	1.00	0.08	0.96	1.00	0.04	0.96	1.00	0.04
Pickled Garlic	1.00	1.00	0.00	1.00	0.98	-0.02	1.00	1.00	0.00	1.00	0.99	-0.01
Spinach	1.00	0.99	-0.01	0.97	1.00	0.03	0.98	1.00	0.02	0.98	1.00	0.02
Total	**0.97**	**0.99**	0.02	**0.93**	**0.97**	0.04	**0.93**	**0.07**	0.04	**0.93**	**0.97**	0.04

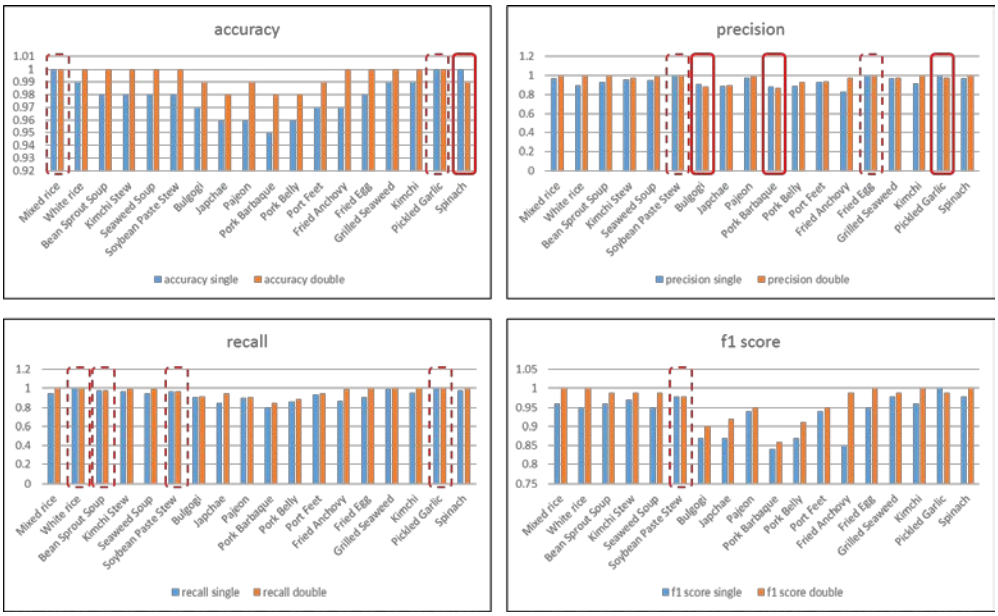

Fig. 7. The results in **Table 5** are visualized using bar graph.

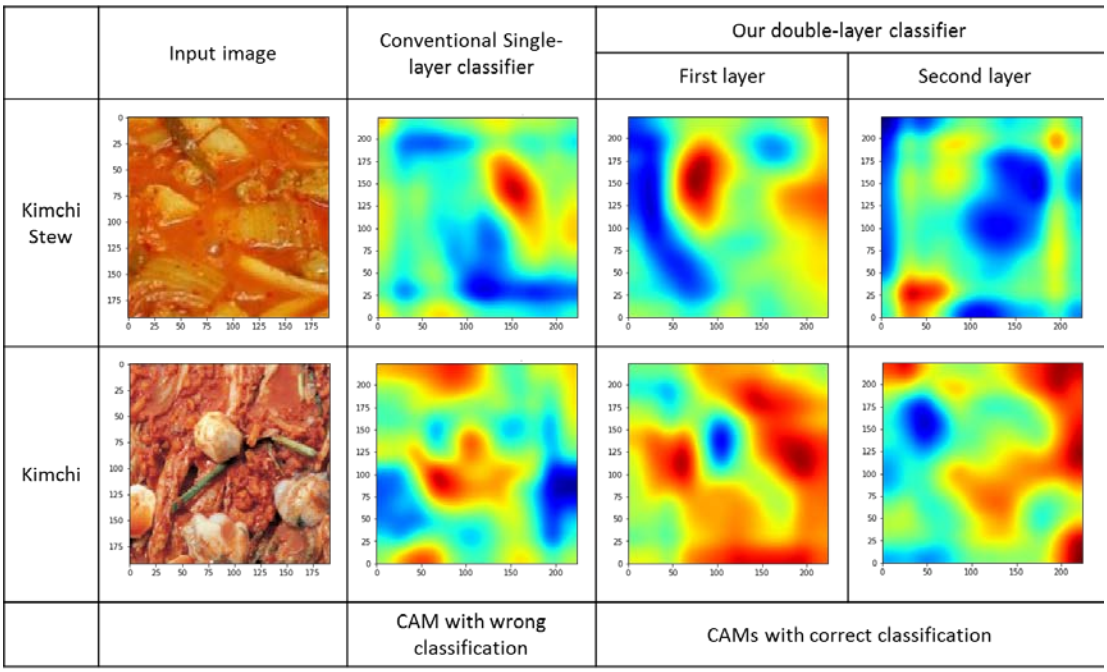

Fig. 8. CAMs for two food photos, which are misclassified by single-layer classifier, but classified correctly by our double-layer classifier: The different CAMs indicate the background regions are different.

7. Conclusion and future work

We present a hierarchical structured recognizer for Korean food of 18 classes. Our recognizer, whose baseline is constructed using DenseNet, classifies Korean food in two stages: the first stage classifies food into 4 major classes, which are rice, soup, main dish and side dish, and the second stage classifies the foods in the major class into the exact class. This hierarchical structure improves the accuracy of the recognition than the conventional single-layer structured recognizer.

As a future work, we aim to extend the Korean foods to cover many Korean food cases. Furthermore, we plan to apply our hierarchical structured recognizer into other related areas. Finally, some other areas such as dietetics can incorporate with our recognizer to build various applications very effective on many users.

Acknowledgement

This research was supported by a research grant from Sangmyung Univ. in 2019.

Reference

[1] Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., and Yang, J., "PFID: Pittsburgh fast-food image dataset," in *Proc. of IEEE International Conference on Image Processing*, pp.289-292, 2009. Article (CrossRef Link)

[2] Joutou, T. and Yanai, K., "A food image recognition system with multiple kernel learning," in *Proc. of IEEE International Conference on Image Processing*, pp.285-288, 2009. Article (CrossRef Link)

[3] Wu, W. and Yang, J., "Fast food recognition from videos of eating for calorie estimation," in *Proc. of IEEE International Conference on Multimedia and Expo*, pp.1210-1213, 2009. Article (CrossRef Link)

[4] Yang, S., Chen, M., Pomerleau, D., and Sukthankar, R., "Food recognition using statistics of pairwise local features," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.2249-2256, 2010. Article (CrossRef Link)

[5] Bosch, M., "Combining global and local features for food identification in dietary assessment," in *Proc. of IEEE International Conference on Image Processing*, pp.1789-1792, 2011. Article (CrossRef Link)

[6] Chen, M., "Automatic chinese food identification and quantity estimation," in *Proc. of ACM Siggraph Asia*, pp.1-4, 2012. Article (CrossRef Link)

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

[8] Bossard, L., Guillaumin, M. and van Gool, L., "Food-101-mining discriminative components with random forests," in *Proc. of European Conference on Computer Vision*, pp.446-461, 2014. Article (CrossRef Link)

[9] Kagaya, H., Aizawa, K and Ogawa, M., "Food detection and recognition using convolutional neural network," in *Proc. of ACM International Conference on Multimedia*, pp.1085-1088, 2014. Article (CrossRef Link)

[10] Kawano, Y. and Yanai, K., "Food image recognition with deep convolutional features," in *Proc. of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp.589-593, 2014. Article (CrossRef Link)

[11] Kawano, Y. and Yanai, K., "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. of European Conference on Computer Vision Workshops*, pp.3-17, 2014. Article (CrossRef Link)

[12] Kagaya, H. and Aizawa, K., "Highly accurate food/non-food image classification based on a deep convolutional neural network," in *Proc. of International Conference on Image Analysis and Processing*, pp.350-357, 2015. Article (CrossRef Link)

[13] Yanai, K. and Kawano, Y., "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. of IEEE International Conference on Multimedia and Expo*, pp.1-6, 2015. Article (CrossRef Link)

[14] Pouladzadeh, P., Yassine, A. and Shirmohammadi, S., "Foodd: food detection dataset for calorie measurement using food images," in *Proc. of International Conference on Image Analysis and Processing*, pp.441-448, 2015. Article (CrossRef Link)

[15] Akbari Fard, M., Hadadi, H. and Tavakoli Targhi, A., "Fruits and vegetables calorie counter using convolutional neural networks," in *Proc. of ACM International Conference on Digital Health*, pp.121-122, 2016. Article (CrossRef Link)

[16] Tatsuma, A and Aono, M., "Food image recognition using covariance of convolutional layer feature maps," *IEICE TRANSACTIONS on Information and Systems*, 99(6), 1711-1715, 2016. Article (CrossRef Link)

[17] Hassannejad, H., "Food image recognition using very deep convolutional networks," in *Proc. of ACM International Workshop on Multimedia Assisted Dietary Management*, pp.41-49, 2016. Article (CrossRef Link)

[18] Jain, N. K. and Khanna, S. O. and Chetna, M., "Feed Forward Neural Network Classification for INDIAN Krishna Kamod Rice," *International Journal of Computer Applications*, 134(14), pp. 38-42, 2016. Article (CrossRef Link)
[19] Ragusa, F., "Food vs non-food classification," in *Proc. of ACM International Workshop on Multimedia Assisted Dietary Management*, pp.77-81, 2016. Article (CrossRef Link)
[20] Singla, A. and Yuan, L. and Ebrahimi, T., "Food/non-food image classification and food categorization using pre-trained GoogLeNet model," in *Proc. of ACM International Workshop on Multimedia Assisted Dietary Management*, pp.3-11, 2016. Article (CrossRef Link)
[21] Farooq, M. and Sazonov, E., "Feature extraction using deep learning for food type recognition," in *Proc. of International Conference on Bioinformatics and Biomedical Engineering*, pp.464-472, 2017. Article (CrossRef Link)
[22] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv print arXiv:1409.1556*, 2014. Article (CrossRef Link)
[23] Krizhevsky, A., Sutskever, I. and Hinton, G., Imagenet classification with deep convolutional neural networks," in *Proc. of Advances in Neural Information Processing Systems*, pp.1097-1105, 2012. Article (CrossRef Link)
[24] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1-9, 2015. Article (CrossRef Link)
[25] He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016. Article (CrossRef Link)
[26] Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K. Q., "Densely connected convolutional networks," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.4700-4708, 2016. Article (CrossRef Link)
[27] Song, J., Gao, L., Nie, F., Shen, H. T., Yan, Y., and Sebe, N., "Optimized graph learning using partial tags and multiple features for image and video annotation," *IEEE Transactions on Image Processing*, Vol. 25, No. 11, pp. 4999-5011, 2016. Article (CrossRef Link)
[28] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., "Learning deep features for discriminative localization," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2921-2929, 2016. Article (CrossRef Link)
[29] Wang, X., Gao, L., Wang, P., Sun, X., Liu, X., "Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length," *IEEE Transactions on Multimedia*, Vol. 20, No. 3, pp. 634 - 644, 2018. Article (CrossRef Link)
[30] Song, J., Guo, Y., Gao, L., Li, X., Hanjalic, A., and Shen, H. T., "From deterministic to generative: Multi-modal stochastic RNNs for video captioning," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 30, No. 10, pp. 3047 - 3058, 2019. Article (CrossRef Link)

Heekyung Yang received her BS degree, MS degree and Ph.D degree from Sangmyung University, Seoul, Korea in 2010, 2012 and 2019, respectively. She is currently a post-doctoral researcher in Sangmyung University Industry-Academy Collaboration Foundation. Her research interests are computer vision, deep learning and computer graphics.

Sungyong Kang received his BS degree in Computer Science from Sangmyung University, Seoul, Korea in 2019. He is currently preparing a start-up venture. His main research interests are deep learning and algorithm optimization.

Chanung Park received his BS degree in Computer Science from Sangmyung University, Seoul, Korea in 2019. He currently works for Samsung Electronics. His main research interests are image recognition and convolutional neural network.

JeongWook Lee received his BS degree in Computer Science from Sangmyung University, Seoul, Korea in 2018. He currently works for Samsung Data System (SDS). His main research interests are deep learning and system software.

Kyungmin Yu received his BS degree in Computer Science from Sangmyung University, Seoul, Korea in 2018. He currently works for Samsung Electronics. His main research interests are machine learning and image classification.

Kyungha Min received his MS in Computer Science from KAIST in 1992. He received his BS and Ph.D in Computer Science and Engineering from POSTECH in 1994 and 2000, respectively. His main research interests are computer graphics and image processing.