Food Image Classification with Deep Features

Abdulkadir ŞENGÜR[1]
Technology Faculty
Electrical and Electronics Eng.
Firat University
Elazig, Turkey
ksengur@firat.edu.tr

Yaman AKBULUT[2]
Informatics Dept.
Firat University
Elazig, Turkey
yamanakbulut@firat.edu.tr

Ümit BUDAK[3]
Engineering Faculty
Electrical and Electronics Eng.
Bitlis Eren University
Bitlis, Turkey
ubudak@beu.edu.tr

Abstract—**In this paper, deep feature extraction, feature concatenation and support vector machine (SVM) classifier are used for efficient classification of food images. Classification of foods according to their images becomes a popular research task for various reasons such as food image retrieval and image based self-dietary assessment. For deep feature extraction, pre-trained AlexNet and VGG16 models are considered. The features of size 4096 are extracted from fc6 and fc7 layers and concatenated with various combinations to determine best deep feature sequence for food image classification. The concatenated features are then classified with SVM. Three publicly available datasets namely FOOD-5K, FOOD-11 and FOOD-101 are used in evaluation of the proposed method and the accuracy metric is considered for performance evaluation. The experimental results show an accuracy of 99.00% for FOOD-5K dataset and 88.08% and 62.44% for FOOD-11 and FOOD-101 datasets, respectively. We further carried out experiments with fine-tuning of a pre-trained CNN model on FOOD-101 dataset and obtained 79.86% accuracy score. The obtained results are also compared with some other methods and it is seen that our performance is better than the other methods on FOOD-11 and FOOD-101 datasets.**

Index Terms—**Convolutional neural networks (CNN), pre-trained CNN models, deep feature extraction, SVM, food image classification.**

I. Introduction

Food, the source of energy supply for the human race, is nowadays quite a hot topic with images of food being constantly shared in social media environments. Especially, with the development of smartphones and mobile technology, people have become used to sharing images of their meals and snacks at a moment's notice through various social media platforms. Thus, millions of food images have become part of our digital world on a daily basis, which has borne the need for the development of food image retrieval applications. Food image retrieval is beneficial for health monitoring applications for people concerned with their calorific food intake. Moreover, food image retrieval can be seen as a multimedia technique for self-dietary assessment [1].

In the past decade, food image classification has excited many researchers from the computer vision and machine learning communities. Several food image datasets exist that consist of a few thousand images and a number of food categories. However, after the explosion in the usability of mobile devices, largescale food image databases that contain more than one-hundred-thousand images and more than one hundred food categories have been constructed [2, 3].

Most of the recent food image classification researches have been carried out with the deep learning [4–11]. Ragusa et al. used various deep representation models and classification techniques for food/non-food discrimination [4]. More specifically, the deep feature extraction from pre-trained CNN models was considered. Authors used the combination of UNICT-FD889, Flickr-Food and Flickr-NonFood datasets to construct the train and test sets of images and obtained 94.86% accuracy with binary SVM classifier. Singla et al. also used the deep learning for food/non-food discrimination [5]. The pre-trained GoogLeNet model was further trained with FOOD-5K dataset to obtain the fine-tuned CNN model. The reported accuracy was 99.2% for food/non-food classification. Authors further considered the same classification approach on a FOOD-11 dataset for food category recognition and obtained 83.6% accuracy. Kagaya et al. considered the CNN approach for food/non-food categorization [6]. Authors used the FOOD-101, Caltech-256 and self-prepared dataset for showing the efficiency of their proposal. The reported accuracies were 96%, 95% and 99% for each investigated dataset, respectively. Aguilar et al. investigated the use of GoogleNet, principal component analysis (PCA) and support vector machines (SVM) in food/non-food discrimination [7]. FCD (FOOD-101 and Caltech-256) and RagusaDS food datasets were considered in experimentations, and 94.97% and 99.01% accuracies were recorded for RagusaDS and FCD datasets, respectively. Yanai et al. fine-tuned a pre-trained DCNN model for food recognition [8]. UEC-FOOD100 and UEC-FOOD256 datasets were used by the authors. The obtained top-1 rates for UEC-FOOD 100 and 256 were 78.77% and 88.97%, respectively. Bossard et al. [9] also trained a DCNN model (AlexNet) with the FOOD-101 dataset, which contained one million food images of 101 food categories. They also trained DCNN from scratch and reported 56.40% accuracy. Mezgec et al. used a new DCNN model called NutriNet for food and drink image classification [10]. The proposed model was tuned on a recognition dataset which contains more than 225 thousand images of size 512×512. The reported classification accuracy was 86.72%. Liu et al. used a deep CNN model for food image classification [11]. Authors used two publicly available food image datasets namely UEC-256 and Food-101.

As the literature about DCNN based food recognition is examined, two dominant DCNN approaches draw the attention. The first one is extracting activation features from a pre-trained DCNN model and then classifies the obtained features with a proper classifier such as SVM. The second approach is fine-tuning a pre-trained DCNN with the new dataset according to the problem at hand. Thus, previously trained DCNN model can be further adapted for a new classification task. In this paper, we opt to use a methodology, which can be settled into the first type category of DCNN approaches, for food image recognition. Instead of using just one pre-trained CNN model, we consider multiple pre-trained CNN models for activation feature extraction. To this end, AlexNet and Vgg16 models and their fc6 and fc7 layers are taken into consideration. Features from different CNN models and layers are concatenated and conveyed through an SVM classifier. The liblinear library and homogenous kernel mapping are used in SVM training. Three food image datasets namely FOOD-101, FOOD-11 and FOOD-5K are used in the experiments. While FOOD-101 (101 categories) and FOOD-11 (11 categories) datasets cover food category classification, FOOD-5K (2 categories) dataset covers the food/non-food categorization.

The organization of the paper is as follows. The next section introduces the methodology. Especially, the CNN feature extraction and SVM classification stages of the proposed method are described. Section 3 describes the datasets, the experimental works, and the obtained results. We finally conclude the work in Section 4.

II. Proposed Method

An illustrative representation of our approach is shown in Fig. 1. As seen in Fig. 1, the input food images are fed into the pre-trained CNN models (AlexNet and VGG16). The obtained feature vectors are then fused to form the final feature vector. Finally, an SVM classifier is used to determine the class label of the input images.

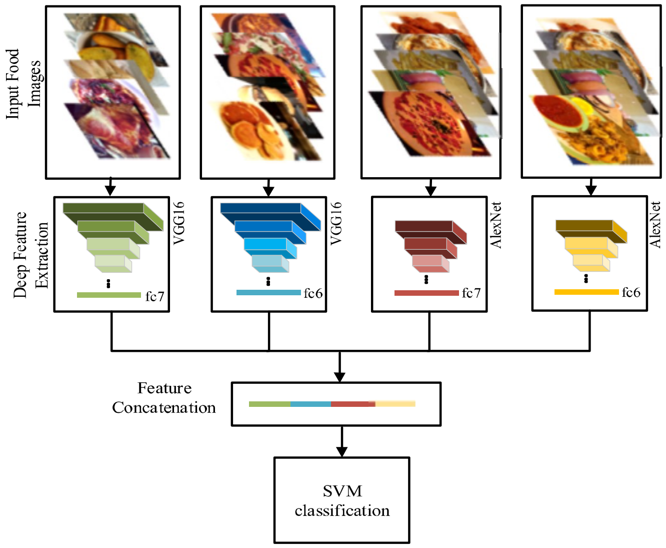

Fig. 1. The illustration of the proposed approach. The input food images are used as input for pre-trained CNN models (AlexNet and VGG16) and extract the activations of fully connected layers as deep feature vectors. These feature vectors are then concatenated. Finally, SVM is used for classification.

A. Deep Features

Two well-known pre-trained CNN models namely AlexNet and VGG16 are considered. Both deep CNN models were trained on ImageNet challenge [12, 13]. AlexNet, which covers totally 25-layers, has 5 convolutions, 3 maximum pooling, 2 dropout, 3 fully connected, 7 ReLu, 2 normalization, 1 softmax, 1 input and 1 classification layers. Fig. 2 shows the illustration of the AlexNet model.

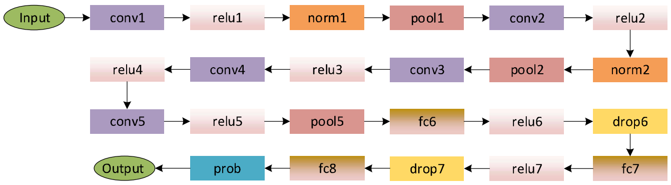

Fig. 2. The illustration of the AlexNet model.

VGG16 model was developed by Simonyan et al. [13]. Similar to AlexNet, the VGG16 model covers totally 41 layers. There are 13 convolutions, 5 maximum pooling, 2 dropout, 3 fully connected, 15 ReLu, 1 softmax, 1 input and 1 classification layers. The deep features are extracted from the fc6 and fc7 layers. The fc6 and fc7 layers produced 4096 dimensional feature vectors. Fig. 3 shows the deep VGG16 model.

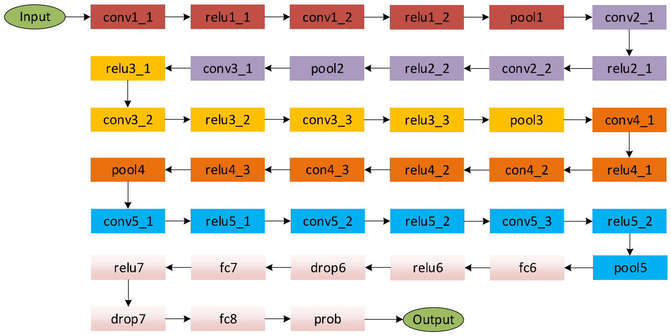

Fig. 3. The illustration of the VGG16 model.

III. Datasets and Experimental Works

A. Datasets

Three publicly available food image datasets are used in the experimental works namely; FOOD-5K, FOOD-11, and FOOD-101, respectively. Several key characteristics of the datasets are given in Table 1.

TABLE I. Several key characteristics of the datasets that were considered in experimental works

Datasets	Number of images	Number of class
FOOD-5K	5000	2
FOOD-11	16643	11
FOOD-101	101000	101

The FOOD-5K dataset, which was originally generated from FOOD-101 [9], UEC-FOOD-100 [14] and UEC-FOOD-

256 [15] datasets, covers totally 5000 images from food and non-food image categories. Half of the dataset consisted of food images (2500) and the other half consisted of non-food images (2500). Fig. 4 shows some examples of food and non-food images in Food-5K.

Fig. 4. Example images of the FOOD-5K dataset (a) Food images, (b) Non-food images.

Similar to the FOOD-5K dataset, the FOOD-11 dataset was originally generated from the FOOD-101 [9], UEC-FOOD-100 [14] and UEC-FOOD-256 [15] datasets. It consists of 16643 images grouped into 11 major types of food categories namely; Bread, Dairy products, Dessert, Fried foods, Meat, Noodles/Pasta, Rice, Seafood, Soup, Egg and Vegetable/Fruit. The dataset was mainly collected from existing food image datasets including FOOD-101 [9], UECFOOD- 100 [14] and UEC-FOOD-256 [15]. Fig. 5 shows example food images of the 11 categories.

FOOD-101 dataset was created by Bossard et al. [9]. The food images were downloaded from "www.foodspotting.com". It consists of 101,000 images grouped into 101 major types of food categories. Authors divided 75,750 images for the training set and the rest 25,250 images were used for testing set construction. Fig. 6 shows example food images of the 100 categories.

Fig. 5. Example images of the FOOD-11 dataset (a) Bread, (b) Dairy products, (c) Dessert, (d) Fried foods, (e) Meat, (f) Noodles and paste, (g) Rice, (h) Sea products, (i) Soup, (j) Egg, (k) Vegetable/Fruit.

Fig. 6. Example images of the FOOD-101 dataset.

B. Experimental Works and Results

The experiments are conducted in order to test the proposed method performance on aforementioned food image datasets. We used MATLAB for all coding on a computer having an Intel Core i7-4810 CPU and 32 GB memory. All food images are initially resized to sizes 227×227 and 224×224 for sake of convenience with AlexNet and VGG16 models, respectively. We extracted features from food images by using fc6 and fc7 [16, 17]. For a single input food image, the mentioned feature vectors are obtained in 0.64 secs. Each CNN model produces 4096-dimensional feature vector. The obtained features are concatenated accordingly. As it was illustrated in Fig.1, the concatenation of the feature vectors is achieved by combination operation. The feature vectors from AlexNet and VGG16 models are merged in a new feature vector. The SVM classifier with homogenous mapping and LIBLINEAR library with the L2-regularised L2-loss dual solver is considered because of its robustness to smaller amounts of training data [18, 19]. The SVM parameter *C* is searched in the range of [10^{-4}-10^{3}] [20, 21]. The performance of the proposed method is scored using accuracy. The classification accuracy is defined as the ratio of the number of correct predictions to the total number of predictions.

The obtained results for all examined datasets are tabulated in Tables 2, 3 and 4 respectively. As seen in Table 1, we give the accuracies of various combinations of the AlexNet and VGG16 features of fc6 and fc7 layers. The first, second, third and fourth columns of Table 2, 3 and 4 show combined feature vectors, accuracy, SVM's *C* parameter and a number of concatenated features, respectively.

TABLE II. Obtained accuracies for various combinations of feature sets for FOOD-5K dataset

Concatenated Feature Sets	Accuracy %	C	Number of features
AlexNet fc6, VGG16 fc6	99.00	0.1	8192
AlexNet fc6, AlexNet fc7	97.90	1	8192
VGG16 fc6, VGG16 fc7	98.50	0.01	8192
AlexNet fc7, VGG16 fc7	98.90	1	8192
AlexNet fc6, AlexNet fc7, VGG16 fc6	98.80	10	12288
AlexNet fc6, AlexNet fc7,VGG16 fc6, VGG16fc7	98.90	0.01	16384

From Table 2, it is seen that all deep feature combinations produce reasonable accuracies and the concatenation of AlexNet fc6 and VGG16 fc6 feature sets produces the highest accuracy. The calculated highest accuracy is 99.00% where the SVM's *C* parameter is 1 and the number of features is 8192. The second highest accuracy 98.90% is obtained for AlexNet fc7 and VGG16 fc7 concatenation and AlexNet fc6, AlexNet fc7, VGG16 fc6 andVGG16fc7 concatenation.

Table 3 shows the obtained results for the FOOD-11 dataset. It is obvious that concatenation of all feature sets produces the 88.08% highest accuracy for the FOOD-11 dataset. On the other hand, the same feature set concatenation did not produce the highest accuracy for the FOOD-5K dataset. Moreover, AlexNet fc6 and VGG16 fc6 feature set concatenation also yield the second highest accuracy. The worst accuracy 80.10% is produced by AlexNet fc6 and AlexNet fc7 feature set concatenation.

TABLE III. Obtained accuracies for various combinations of feature sets for FOOD-11 dataset

Concatenated Feature Sets	Accuracy %	C	Number of features
AlexNet fc6, VGG16 fc6	89.33	0.1	8192
AlexNet fc6, AlexNet fc7	80.10	0.1	8192
VGG16 fc6, VGG16 fc7	86.14	0.01	8192
AlexNet fc7, VGG16 fc7	86.76	10	8192
AlexNet fc6, AlexNet fc7, VGG16 fc6	85.84	0.1	12288
AlexNet fc6, AlexNet fc7,VGG16 fc6, VGG16fc7	88.08	0.01	16384

Finally, Table 4 shows the concatenated feature sets achievements on FOOD-101 datasets. Concatenated AlexNet fc6 and VGG16 fc6 feature sets obtain 62.44% accuracy which is the highest among all combinations.

TABLE IV. Obtained accuracies for various combinations of feature sets for FOOD-101 dataset

Concatenated Feature Sets	Accuracy %	C	Number of features
AlexNet fc6, VGG16 fc6	62.44	1	8192
AlexNet fc6, AlexNet fc7	48.88	0.1	8192
VGG16 fc6, VGG16 fc7	57.86	1	8192
AlexNet fc7, VGG16 fc7	59.22	1	8192
AlexNet fc6, AlexNet fc7, VGG16 fc6	55.51	0.01	12288
AlexNet fc6, AlexNet fc7,VGG16 fc6, VGG16fc7	60.45	0.1	16384

The second highest accuracy is obtained with the concatenation of all deep feature sets. The calculated accuracy is 60.45%. In addition, the worst achievement is obtained with AlexNet fc6 and AlexNet fc7 feature sets concatenation.

We also applied fine tuning on FOOD-101 dataset classification. Fine tuning is known as further training of a pre-trained CNN model on a new problem at hand. Thus, faster and easier training of a deep network can be achieved. ResNet50 model was also used for fine-tuning. As a usual procedure of the fine-tuning, the last three layers of the ResNet50 model was discharged and that layers were arranged according to the FOOD-101 dataset classification problem. The number of maximum epoch was set to 10, miniBatchSize was set to 10 and the learning rate was set to 10^{-3}. The training of the network was achieved by stochastic gradient descent with momentum method. While 75% of the dataset was used for training of the network, 25% of the dataset was used for testing of the network performance. The training was completed in almost three days. The classification accuracy was calculated for evaluation of the fine tuning achievement on FOOD-101 dataset. The obtained accuracy was 79.86%.

We further compare the obtained highest accuracies for each dataset with some of the other published results in the literature. To this end, Tables 5 and 6 are presented. In Table 5, we compare the achievements of Singla et al. [5] and our method. As seen in Table 5, Singla et al.'s achievement on FOOD-5K is 0.2% better than our achievement. On the other hand, our accuracy on the FOOD-11 dataset is 5.83% higher than the Singla et al.'s accuracy.

TABLE V. COMPARISON OF THE PROPOSED METHOD WITH SINGLA ET AL.'S METHOD

Method	FOOD-5K	FOOD-11
Singla et al. [5]	99.20%	83.50%
Proposed method	99.00%	89.33%

The related comparison on the FOOD-101 dataset is given in Table 6. The performances of five methods, which are previously applied on the FOOD-101 dataset, are considered in comparison. From Table 6, it is seen that the proposed method obtains the highest accuracy. Our achievement is 2.46% more accurate than Liu et al.'s method [11].

TABLE VI. COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS ON FOOD-101 DATASET

Methods	FOOD-101
RFDC [9]	50.76%
CNN [9]	56.40%
BoW [9]	28.51%
IFV [9]	38.88%
MLDS [9]	42.63%
Ours (Deep Features)	62.44%
Liu et al.[11]	77.40%
Ours (Fine-tuning)	79.86%

IV. CONCLUSION

In this paper, we investigate the effect of the concatenation of deep features on food classification performance. In other words, our effort is to determine the best deep feature sets combination on the task of food/non-food and food image category classification. Three food image datasets and two pre-trained CNN models are considered in the research. The experimental results show an accuracy of 99.00% on food/non-food image classification and 88.08% and 62.44% on 11 and 101 class food category classifications. In addition, with fine-tuning, we obtained 79.86% accuracy value for FOOD 101 dataset. From the comparisons, it is seen that our performance is better than the other methods on food category classification. In addition, when overall experimental results are considered, it is seen that the AlexNet fc6 and VGG16 fc6 concatenation produce the highest accuracies. In the future works, we are planning to investigate the effect of the early layers feature concatenation on food category classification. We further plan to use the pre-trained GoogleNet to extract feature and concatenate it with AlexNet and VGG16 feature sets.

REFERENCES

[1] G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato. "Retrieval and classification of food images", Computers in Biology and Medicine, 77: 23-39, 2016.

[2] http://mmspg.ep.ch/food-image-datasets.

[3] https://www.vision.ee.ethz.ch/datasets_extra/food-101/

[4] Ragusa, F. et al.: Food vs Non-Food Classification. In: Proceedings of the 2nd International Workshop on MADiMa (2016).

[5] Singla, A., Yuan, L., & Ebrahimi, T. (2016, October). Food/non-food image classification and food categorization using pre-trained googlenet model. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management (pp. 3-11). ACM.

[6] H. Kagaya and K. Aizawa. Highly Accurate Food/Non-Food Image Classification Based on a Deep Convolutional Neural Network, pages 350-357. Springer International Publishing, Cham, 2015.

[7] Aguilar, E., Bolanos, M., & Radeva, P. (2017). Exploring food detection using cnns. arXiv preprint arXiv:1709.04800.

[8] Yanai, K., and Yoshiyuki K., "Food image recognition using deep convolutional network with pre-training and fine-tuning." Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on. IEEE, 2015.

[9] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 - mining discriminative components with random forests," in Proc. of European Conference on Computer Vision, 2014.

[10] Mezgec, S., & Koroušić Seljak, B. (2017). Nutrinet: A deep learning food and drink image recognition system for dietary assessment. Nutrients, 9(7), 657.

[11] Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., & Ma, Y. (2016, May). Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In International Conference on Smart Homes and Health Telematics (pp. 37-48). Springer, Cham.

[12] 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.

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computing Research Repository (CoRR), vol. abs/1409.1556, 2014.

[14] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In Proc. of IEEE International Conference on Multimedia and Expo (ICME), 2012.

[15] Y. Kawano and K. Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV), 2014.

[16] Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., ... & Schuller, B. Snore sound classification using image-based deep spectrum features. In Proceedings INTERSPEECH, 2017.

[17] R.-E. Fan, K.-W., Chang, C.-J., Hsieh, X.-R. Wang, and C.-J. Lin., "LIBLINEAR: A library for large linear classification," Journal of Machine Learning Research, vol. 9, pp. 1871-1874, 2008.

[18] A. Vedaldi and A. Zisserman, "Efficient Additive Kernels via Explicit Feature Maps," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2010.

[19] Yadav A R, Anand R S, Dewal M L, Gupta S (2015) 'Multiresolution local binary pattern variants based texture feature extraction techniques for efficient classification of

microscopic images of hardwood species', Applied Soft Computing, Vol. 32, pp. 101-112.

[20] Yadav, A.R., Anand, R., Dewal, M. and Gupta, S. (2015) 'Gaussian image pyramid based texture features for classification of microscopic images of hardwood species', Optik-International Journal for Light and Electron Optics, Vol. 126, No. 24, pp.5570-5578.