Highly Accurate Food/Non-Food Image Classification based on a Deep Convolutional Neural Network

Hokuto Kagaya[1], Kiyoharu Aizawa[1, 2]

[1]The University of Tokyo, Graduate School of Interdisciplinary Information Studies, Tokyo, Japan
[2]The University of Tokyo, Dept. Information and Communication Engineering, Tokyo, Japan
`{kagaya, aizawa}@hal.t.u-tokyo.ac.jp`

Abstract. "Food" is an emerging topic of interest for multimedia and computer vision community. In this paper, we investigate food/non-food classification of images. We show that CNN, which is the state of the art technique for general object classification, can perform accurately for this problem. For the experiments, we used three different datasets of images: (1) images we collected from Instagram, (2) Food-101 and Caltech-256 dataset (3) dataset we used in [4]. We investigated the combinations of training and testing using the all three of them. As a result, we achieved high accuracy 96, 95 and 99% in the three dataset respectively.

Keywords: Food/Non-Food Classification, Convolutional Neural Network, Deep Learning

1 Introduction

"Food" is an emerging topic of interest for multimedia and computer vision community. It is a very important issue for healthcare. We have been developing a novel food recording tool, "FoodLog" [1, 2], which helps users easily record their everyday meals by the assistance of image retrieval.

However, analysis of food images is in general very challenging. For example, recognition of a food item in an image is still difficult because intra-class variance is high and inter-class variance is low. Moreover, the number of food classes is not well determined yet.

In this paper, we investigate a problem of food/non-food classification. Given an image, we want to find if the image contains food or not. It is a binary classification of images. One of the applications of such classification is a pre-processing for food image recognition.

In recent years, the number of photographs uploaded to social networking services (SNS) has been explosively increasing. Photo/video sharing services such as Instagram, Flickr and Pinterest are very popular. When users upload a photo to them, they add keywords or hash tags explaining the image content. However, in reality, the hash tags they use are not very reliable. Fig. 1 shows a search result by "#food" of Insta-

gram. Although "food" is specified in the search, there are more non-food images in the results. In order to show food images, we need to automatically filter the results by the binary food/non-food classification.

In our study, we arranged three datasets for the experiments: Instagram, Food-101[5] and the one we used in [4]. We adopted convolutional neural network (CNN) [3] as feature extractor and classifier, which is the state-of-the-art technique.

Fig. 1. Top two pages of search results by query "*#food*" in Instagram.

2 Related Work

This study is an extension of our previous study on food image recognition. In our previous paper [4], we have addressed the effectiveness of CNN for food image recognition and detection (classification). We had also presented food/non-food classification by SVM with hand-crafted features for personal food logging applications [1].

Analysis of food images has attracted much attention of multimedia and computer vision communities. Bossard et al., [5] built a publicly available food image dataset, Food-101, and they examined the efficiency of their method to mine discriminative parts using Random Forests. We utilize this dataset in the experiment. Li et al. [6] presented food recognition using a small dataset: their study was a part of their Technology Assisted Dietary Assessment project. From the point of view of the datasets, Pittsburgh Fast-food Image Dataset [7] is one of the earliest datasets. It consists of American fast-food images. Kawano et al., [8] introduced a UEC Food-256 Dataset, which is a dataset constructed by crowdsourcing. These food recognition related works did not handle food/non-food classification.

As mentioned earlier, food/non-food classification will be beneficial to searching food images in current photo sharing SNS, as well as to pre-processing of food recognition.

3 CNN-Based Approach to Food/Non-Food Classification

The CNN offers a state-of-the-art technique for general image recognition. It is a multilayer neural network, whose neurons take small patches of the previous layer as input. It is robust against small shifts and rotations. A CNN mainly comprises convolution layers and pooling layers.

The biggest advantage of CNN is to be able to learn high-level efficient features from data. We expect that CNN will extract important features of "food" images. We utilize the Network in Network model [9] as CNN architecture, because it is fast for training, performs better than the AlexNet [14] and it is memory efficient. We call our model "CNN-NIN" for the rest of the paper. It has four convolution layers with two mlpconv layers (see the detail in [9]). Additionally, to Copy mid-level image representations in CNN is very efficient for computation time and accuracy [13]. Therefore, we employ the model pre-trained by ImageNet that is publicly avail-able, and after copying the model parameters, we fine-tune them for our datasets. We revise the dimension of outputs from 1000 to 2 for food/non-food classification.

We also train CNN from scratch and compare the results with the fine-tuned model. In our research, we use caffe [10] as the CNN library, which is a standard GPU implementation in C++ and python.

4   Datasets

We introduce three dataset in this section. The datasets we collected from Instagram and Food-101/Caltech-256 are described later. The dataset used in [4] is made of 1,234 food images and 1,980 non-food images collected from social media. We utilize it for a comparative study in Section 5.3.

Fig. 2. 16 samples from Instagram Food/Non-Food Dataset. Left one is positive, right one is negative.

Fig. 3. 16 samples of images from Food-101/Caltech-256. Left one is positive, right one is negative.

Fig. 4. 16 examples of images from the dataset used in [4]. Left one is positive, right one is negative.

4.1   Instagram-#Food Dataset (IFD)

We built an Instagram Food/Non-Food Dataset (IFD, in short). We collected them from the search results of "#food" in Instagram and manually annotated images with food or non-food labels. Our annotation criteria are as follows: 1) The main content of image should be food. 2) Food in the image should be real, not illustration. The first criterion is relatively subjective. If the image captures some objects with food in the

background, we excluded them from the dataset. In the dataset, we finally have 4,230 food images and 5,428 non-food images.

4.2   Food-101/Caltech-256 Dataset (FCD)

We also built another dataset using widely available datasets. We used Food-101 [5] as food images and Caltech-256 as non-food image (This dataset is called FCD, in short). Food-101 images were originally downloaded from foodspotting.com, which is a sharing site that allows users to upload images with additional information. Among 1,000 images of each class, we chose 250 images so that the variance of a color feature (64-dim color histogram) within the class becomes high. Color feature is one of the most important features for food image recognition [4, 12]. As a result, we chose 25,250 positive food images in total.

Caltech-256[11] is a standard dataset for general image recognition and categorization. We use it for negative samples after excluding the images related to food. The categories of Caltech-256 excluded from dataset are shown in Table 1. Finally, we chose 28,322 non-food images.

Table 1. The categories excluded from Caltech-256.

Excluded categories
crab, mussels, octopus, grape, mushroom, tomato, watermelon, beer-mug, cereal-box, coffee-mug, soda-can, wine-bottle, teapot, cake, ice-cream-cone, fried-egg, hamburger, hot-dog, spaghetti, sushi, drinking-straw, frying-pan

5   Experiment

In this section, we show the experimental results on three datasets: IFD, FCD and our previous dataset. In Section 5.1, 5.2 and 5.3, we present the results of evaluation within the same dataset. In 5.3, we compare our results with previous results of conventional methods. We show the results of the evaluation across the datasets in Section 5.4. Finally, the result of comparative study between the from-scratch model and the fine-tune model is shown in Section 5.5.

Table 2. Experimental results of IFD and FCD with different train/test ratios.

Train/test Ratio		0.5	0.6	0.7	0.8	0.9
Accuracy (%)	IFD	94.5	94.7	94.5	94.8	95.1
	FCD	96.2	96.3	96.2	96.4	96.1

5.1   Instagram Food/Non-Food Dataset (IFD)

Firstly, we examined accuracy of training/testing ratio, which is the ratio of the number of training samples and that of testing. We randomly selected images from IFD, and then evaluated the accuracy.

The top row of Table 2 shows the results by different training/testing ratios, which is the averaged value over five trials. As can be seen, all values are about 96% for all ratios. The differences in the training/testing ratios are few, hence we fixed the ratio at 0.8 for the rest of the paper. Fig. 5 shows confusion matrix of the trials with train/test ratio 0.8; from the results, differences of the accuracies among classes

(food or non-food) were not observed. Table 3 shows some examples of classified images. The results that were not classified as food included food products and small food regions. Some results which were not classified as non-food included flowers (similar to food) or objects of colors similar to foods.

Table 3. Classification examples of IFD. Incorrect results may be influenced by its background, products, small food regions.

	Ground truth : food	Ground truth : non-food
Correct		
Incorrect		

Fig. 5 (left) confusion matrix in the experiments of IFD (average over 5 trials with train/test ratio 0.8). (right) confusion matrix in the experiments of FCD (average over 5 trials with train/test ratio 0.8).

5.2   Food-101/Caltech-256 Dataset (FCD)

We conducted the evaluation using Food-101/Caltech-256 Dataset in the same manner as Section 5.1. The accuracies are shown in the bottom row of Table 2, and a confusion matrix of one particular trial is shown in Fig. 5 (right), both of which show a similar trend to the results of IFD. Example images are shown in Table 4.

Table 4. Classification examples of FCD.

	Ground truth : food	Ground truth : non-food
Correct		
Incorrect		

5.3   Dataset Used in [4]

In this section, we show a comparative study of the proposed method against our two previous methods, which are the baseline method based on SVM of handcrafted features in our FoodLog system and Alex-Net CNN.

Table 5 shows the results of the comparisons. The values are averaged over ten tri-als[1] of different combinations of training and testing images using the dataset of [4]. As shown in Table 5, the proposed CNN-NIN produced the best accurate performance among the three methods.

Table 5. Comparison with the previous methods.

Method	Accuracy
Baseline ([1, 4])	89.7±0.73%
CNN [4]	93.8±1.39%
CNN-NIN (this paper)	**99.1±0.81%**

5.4　Cross Dataset Evaluation

The three datasets may have statistical differences in food/non-food images. To ex-amine the effects of the difference of datasets, we evaluated accuracies of the pro-posed system using different datasets for training and testing. We conducted two evaluations as follows:

(A)　　Training: FCD　and　Testing: IFD
(B)　　Training: IFD　and　Testing: FCD

Table 6 shows accuracies, and Fig. 6 shows the confusion matrices of (A) and (B). The accuracy of food images classification was degraded in (A), and that of non-food images was degraded in (B). It is considered that IFD and FCD datasets have slight differences. The dataset should be larger for the classification to be robust.

Table 6. Result for (A) and (B).

Experiment	Accuracy
(A)	91.5%
(B)	90.6%

Fig. 6. (left) Confusion matrix of (A). (right) Confusion matrix of (B).

Table 7. Examples incorrectly classified by the model in (A) but correctly classified by the model of IFD

Ground truth : food	Ground truth : non-food

We compared the model of (A) with one of the model of IFD (section 5.1), using the best accuracy in five trials. In other words, we compared the model trained by the same dataset as when testing with the model trained by the different one. The images

---

[1]　The manner of evaluation for the existing method in [4] differs a little. See the detail in [4].

for testing are those not used in the training of this model in Section 5.1. Table 7 shows some examples. As seen in Table 7, Instagram images in which colors were modified were not correctly classified.

5.5 Fine-tune model vs. from-scratch for cross dataset evaluation

To examine the effects of differences of CNN trainings, we conducted evaluations for comparison between a fine-tune model and a from-scratch model using the datasets of (A). Fine-tune models were made as follows: We used a pre-trained model trained using ImageNet images [15]. Following the method in [13], the transferred parameters were copied from the pre-trained model and the last one or two layers were adapted. In addition, we used the pre-trained model as the initial weights and re-trained, too. We also evaluated CNN made from scratch.

Table 8 shows the result of the experiments. Fine-tune models were better than the from-scratch regarding accuracy and computational time required for training. The size of the datasets would be much larger the from-scratch could be better. Among the three fine tune ones, re-training version was the best.

Table 8. Results of comparison between from-scratch and fine-tuning

Method	Transferred Parameters	New adaption Layers	Accuracy
Fine-tune	Re-train	Last one	**91.5%**
Fine-tune	Fixed	Last one	89.6%
Fine-tune	Fixed	Last two	91.1%
From-scratch	-	-	86.4%

6 Conclusion

In this paper, we have examined the effectiveness of a CNN-based approach for food/non-food classification with three datasets. The datasets were collected from publicly available images and social media. The model for this task can be applied to pre-processing of food item recognition or filter the search result of queries related to food, meals or dishes. In the future, food/non-food classification could be applied to more complex processing of food images.

Reference

1. Kitamura, K., Yamasaki, T., & Aizawa, K. (2008, October). Food log by analyzing food images. In Proceedings of the 16th ACM international conference on Multimedia (pp. 999-1000). ACM.
2. K. Aizawa, M.Ogawa, FoodLog: Multimedia Tool for Healthcare Applications, IEEE MultiMedia, vol. 22, no. 2 pp.4-9, 2015

3. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

4. Kagaya, H., Aizawa, K., & Ogawa, M. (2014, November). Food Detection and Recognition Using Convolutional Neural Network. In Proceedings of the ACM International Conference on Multimedia (pp. 1085-1088). ACM.

5. Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–Mining Discriminative Components with Random Forests. In Computer Vision–ECCV 2014 (pp. 446-461). Springer International Publishing.

6. He, Y., Xu, C., Khanna, N., Boushey, C. J., & Delp, E. J. (2014, October). Analysis of food images: Features and classification. In Image Processing (ICIP), 2014 IEEE International Conference on (pp. 2744-2748). IEEE.

7. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., & Yang, J. (2009, November). PFID: Pittsburgh fast-food image dataset. In Image Processing (ICIP), 2009 16th IEEE International Conference on (pp. 289-292). IEEE.

8. Kawano, Y., & Yanai, K. (2014, November). FoodCam-256: A Large-scale Real-time Mobile Food RecognitionSystem employing High-Dimensional Features and Compression of Classifier Weights. In Proceedings of the ACM International Conference on Multimedia (pp. 761-762). ACM.

9. M. Lin, Q. Chen, S. Yan. Network in Network. In proceedings of International Conference on Learning Representations, 2014

10. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the ACM International Conference on Multimedia (pp. 675-678). ACM.

11. Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.

12. Bosch, M., Zhu, F., Khanna, N., Boushey, C. J., & Delp, E. J. (2011). Combining Global and Local Features for Food Identification in Dietary Assessment. IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society, 2011, 1789–1792. doi:10.1109/ICIP.2011.6115809

13. Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014, June). Learning and transferring mid-level image representations using convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 1717-1724). IEEE.

14. Krizhevsky, A., Sutskever, I. and Hinton. G. E. ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012: Neural Information Processing Systems, Lake Tahoe, Nevada

15. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.