# Food/Non-food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model

Ashutosh Singla
ashutosh.singla@epfl.ch

Lin Yuan
lin.yuan@epfl.ch

Touradj Ebrahimi
touradj.ebrahimi@epfl.ch

Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne
Station 11, 1015 Lausanne, Switzerland

## ABSTRACT

Recent past has seen a lot of developments in the field of image-based dietary assessment. Food image classification and recognition are crucial steps for dietary assessment. In the last couple of years, advancements in the deep learning and convolutional neural networks proved to be a boon for the image classification and recognition tasks, specifically for food recognition because of the wide variety of food items. In this paper, we report experiments on food/non-food classification and food recognition using a GoogLeNet model based on deep convolutional neural network. The experiments were conducted on two image datasets created by our own, where the images were collected from existing image datasets, social media, and imaging devices such as smart phone and wearable cameras. Experimental results show a high accuracy of 99.2% on the food/non-food classification and 83.6% on the food category recognition.

## Keywords

Caffe; convolutional neural network (CNN); food/non-food classification; food recognition; deep learning; GoogLeNet

## 1. INTRODUCTION

Well-being is becoming a topic of great interest and an essential factor linked to improvements in the quality of life. Modern information technologies have brought a new dimension to this topic. It is now possible, thanks to various wearable devices (health bands, smart watches, smart clothes, etc.), to gather a wide range of information from subjects such as number of steps walked, heart rate, skin temperature, skin conductivity, transpiration, respiration, etc. and analyze this information in terms of the amount of calories spent, level of stress, duration and quality of sleep, etc. An accurate estimation of daily nutritional intake provides a useful solution for keeping healthy and to prevent diseases. However, it is not easy to assess the nutritional value of food and beverage consumed by subjects in an automatic and accurate way.

In the recent years, there has been a lot of developments in the field of dietary assessment based on multimedia techniques, for example, based on food image analysis. An automatic image-based dietary assessment system follows the basic steps of: food image detection, food item recognition, quantity or weight estimation, and finally caloric and nutritional value assessment [1]. In the last couple of years, advancements in image processing, machine learning and in particular deep learning, and convolutional neural network (CNN) proved to be a boon for the image classification and recognition tasks, including for the problem of food image recognition. Researchers have been working on different aspects of a food recognition system, but there is still a lack of good-enough solution to high-accuracy food classification and recognition, considering a wide variety of food items and highly mixed food items in many images. Therefore, it is extremely difficult to correctly recognize every food item, as many of the food items may look similar in color or shape and are not even distinguishable to human eyes, e.g., beef vs. horse meat. Moreover, in reality, a plate with highly mixed food makes the problem even more difficult to solve. Therefore, we state that it would be good enough to recognize the general type of a certain food item, based on which we can approximately estimate its dietary value, e.g., calories. It can already provide people with basic information on their daily intake.

The paper reports two sets of experiments: 1) food/non-food image classification, and 2) food category recognition. In order to train our model for classification and recognition, we created two datasets from the existing food image datasets, social media and mobile devices. A GoogLeNet model based on deep CNN was fine-tuned and trained using our image data in a deep learning framework - Caffe.

The rest of the paper is structured as follows. Section 2 introduces the related works carried out by other researchers after a brief discussion of the differences between food detection and food classification. Section 3 briefly introduces the convolutional neural network (CNN) and GoogLeNet model. Section 4 describes the food image datasets used for experiments. Then Section 5 shows the experimental results on food/non-food classification and food category recognition. Finally, we conclude the paper and discuss the future work in Section 6.



*MADiMa'16, October 16 2016, Amsterdam, Netherlands*



ISBN 978-1-4503-4520-0/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2986035.2986039

## 2. RELATED WORK

Food image detection and recognition are the active research topics in the area of computer vision. Researchers have published several approaches to solve these two problems. The first problem is to detect automatically the images that contain food items. This is an indispensable step for an automatic food analysis system. In some cases, it is enough to classify a food image, when the main objective is to annotate images that contain food for the purpose of organizing them into different categories. In multimedia dietary assessment, one should be able to also find out what food items are in an image, their locations, as well as their amount.

### 2.1 Food Image Detection

The task of detecting whether an image contains food item is a binary classification problem, namely, food/non-food classification. Given an image, a food classifier identifies an image as food or non-food. This is similar to any other image classification problem where a classifier is trained on image data using machine learning techniques. Classical approaches to image classification extract features such as interest point descriptors from scale-invariant feature transform (SIFT) [2], pool the features into a vector representation e.g., bag of words [3] and Fisher Vectors [4] and then use a clustering algorithm such as Support Vector Machine (SVM) for classification. Kitamura et al. [5] applied SVM on image features consisting of color histograms, DCT coefficients and detected image patterns in food image detection and obtained an accuracy of 88%. [6] reports an automatic detector that finds circular dining plates in chronically recorded images or videos. As an important application, the method can be used to detect food intake events automatically by identifying dining plates in chronically recorded video acquired by a wearable device.

Recently, the Convolutional Neural Network (CNN) [7] offers a state-of-the-art technique for many general image classification problems. It has been applied in food classification and resulted in a high accuracy. Kagaya et al. [8] applied CNN in food/non-food classification and achieved significant results with a high accuracy of 93.8%. Then, in the work [9], the accuracy of food detection was increased to 99.1%, using a subset of their image dataset. Compared to previous works that use conventional machine learning approaches, CNN seems to provide superior performance.

### 2.2 Food Image Recognition

Most research works in food recognition assume that only one food item is present in the image. Thus, food recognition can be solved as a multiclass classification problem. Researchers have been working on food recognition using conventional approaches based on classical image features and machine learning for many years. Joutou et al. [10] created a private Japanese food dataset with 50 classes. They proposed a Multiple Kernel Learning (MKL) method using combined features including SIFT-based bag-of-features, color histogram and Gabor Texture features. An accuracy of 61.3% on their dataset was achieved. A follow-up study by Hoashi et al. [11] achieved an accuracy rate of 62.5% using the same method on an extended dataset of 85 classes. Chen et al. [12] created the Pittsburgh food database which contained 101 classes of American fast food images taken in a controlled environment. Yang et al. [13] defined eight basic food materials and learned spatial relationships between these ingredients in a food image using pairwise features. They achieved a classification accuracy of 28.2% on 61 food categories which was a subset of Pittsburgh dataset [12]. Bettadapura et al. [14] used combined 6-feature descriptors (2 color-based and 4 SIFT-based) and SMK-MKL Sequential Minimal Optimization to train an SVM classifier. They experimented on a dataset consisting of 3750 food images of 75 categories (50 images per category) and reported an accuracy of 63.33% on their test dataset. Interestingly, they incorporated the geological information of where the food picture was taken so that they could get the information about the restaurant and then downloaded the menu online. An assumption of their work is that the food image must be one of the items in the menu. Rahmana et al. [15] presented a new method for generating scale and/or rotation invariant global texture features using the output of Gabor filter banks, which provides a good accuracy of food classification for a mobile phone based dietary assessment system. The top-5 accuracy they achieved was almost 100%. However, the experiment was conducted on a special image dataset of only 209 food images created with controlled environment. He et al. [16] investigated different features and their combinations for food image analysis and a classification approach based on k-nearest neighbors and vocabulary trees. The experimental results indicate that a combination of three features, Dominant Color Descriptor (DCD), Multi-scale Dense SIFT (MDSIFT) and Scalable Color Descriptor (SCD), provides the best performance on food recognition. Bossard et al. [17] created an image dataset called Food-101, which contains 101 types of food images. They presented a method based on Random Forests to mine discriminative visual components and could efficiently classify with an accuracy rate of 50.8%.

In recent years, CNN is also widely used in food recognition and provides better performance than the conventional methods. Bossard et al. [17] trained a deep CNN from scratch on Food-101 dataset using the architecture of AlexNet model (proposed by Krizhevsky et al. [18]) and achieved 56.4% top-1 accuracy. Their proposed a new method based on Random Forest outperforms state-of-the-art methods on food recognition. In [8], Kagaya et al. also trained CNN for food recognition and the experimental results showed that the CNN outperformed all the other baseline classical approaches by achieving an average accuracy of 73.7% for 10 classes. Kawano et al. [19] used CNN as a feature extractor and achieved state-of-the-art best accuracy of 72.3% on the UEC-FOOD-100 [20] dataset, which contains 100 classes of Japanese food. They used the pre-trained AlexNet model as a feature extractor and integrated both CNN features and Fisher Vector encoded conventional SIFT and color features. Yanai et al. [21] fine-tuned the AlexNet model and achieved the best results on public food datasets so far, with top-1 accuracy of 78.8% for UEC-FOOD-100 dataset and 67.6% for UEC-FOOD-256 [22] (another Japanese food image dataset with 256 classes). Their works showed that the recognition performance on small image datasets like UEC-FOOD-256 and UEC-FOOD-100 (both of which contained 100 images for each class) can be boosted by fine-tuning the CNN network which was pre-trained on a large dataset of similar objects. Myers et al. [23] presented the Im2Calories system for food recognition which extensively used CNN-based approaches. The architecture of GoogLeNet [24] was used in

their work and a pre-trained model was fine-tuned on Food-101. The resulting model has a top-1 accuracy of 79% on Food-101 test set.

## 3. CONVOLUTIONAL NEURAL NETWORK

Over the last few years, due to the advancements in the deep learning, especially in the convolutional neural networks, the accuracy in identifying and recognizing images has been increased drastically. This is not only because larger datasets but also new algorithms and improved deep architectures [24]. Convolutional Neural Network (CNN) is also known as LeNet due to its inventor [25]. CNN mainly comprises convolutional layers, pooling layers and sub-sampling layers followed by fully-connected layers. The very first architecture of CNN [7] takes an input image and applies convolution followed by sub-sampling. After two such computations, the data is fed into the fully connected neural network, where it performs the classification task [7]. The main advantage of CNN is the ability to learn the high-level efficient features and in addition to that, it is robust against small rotations and shifts.

Significant progress has been made on this basic design of CNN and it has been extended by increasing the number of layers [26], size of layers [27] and better activation function, e.g., ReLU [28] to yield the best results on various challenges related to object classification, recognition and computer vision.

In this paper, we use GoogLeNet model, which was developed recently based on deep convolutional neural network, in order to classify food/non-food images and then recognize the food images as one of the 11 categories defined in 4.2. GoogLeNet is an efficient deep neural network architecture, which has a new level of organization called "Inception Module". It consists of convolutions and maxpooling operation and there are nine such modules in GoogLeNet architecture. Fully-connected layers are being replaced with parallel convolutions that operate on the same input layer. The 1×1 convolutions at the bottom of the module reduce the number of inputs and hence decreases the computation cost dramatically. It also captures the correlated features of an input image in the same region. Where as, image patterns are responded by 3×3 and 5×5 convolutions at larger scales. Feature maps which are being produced by all the convolutions are concatenated to form the output [24]. GoogLeNet uses 12 times fewer parameters than [28] which was the winning architecture in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 and also performs significantly better in terms of accuracy [24].

## 4. DATASET

We have created two image datasets, named **Food-5K** and **Food-11**, used for the experiments on food/non-food classification and category recognition respectively. Both datasets are split into three subsets, for the purpose of training, validation and evaluation respectively[1]. In addition, another dataset created by [9] was used in our experiments to evaluate the performance of our model on food/non-food classification. Descriptions of all the datasets are given below.

[1]The datasets are publicly accessible in http://mmspg.epfl.ch/food-image-datasets.

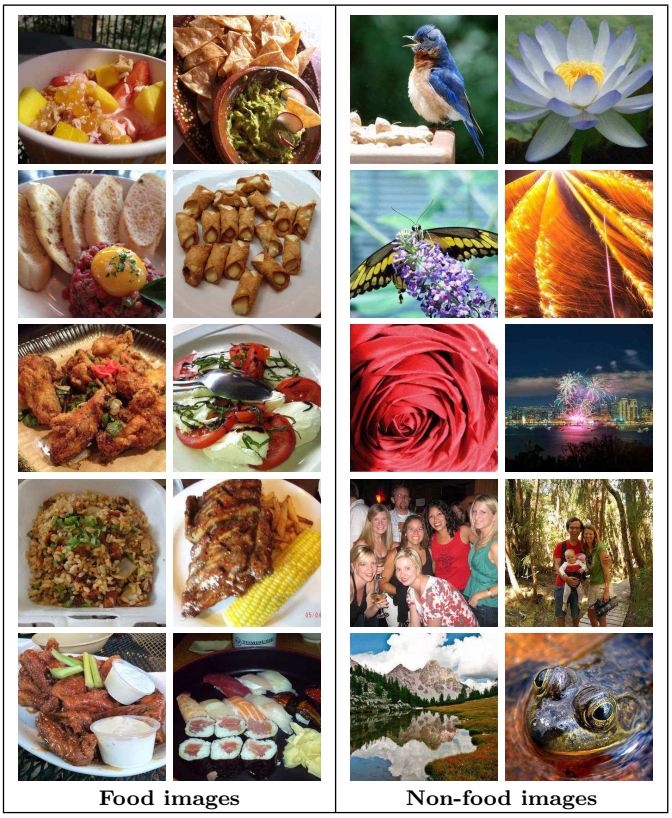



**Figure 1: Example images of Food-5K dataset.**

### 4.1 Dataset 1: Food-5K

Food-5K contains 2,500 food images and 2,500 non-food images, resulting in a total of 5,000 images. The food images were selected from already existing and publicly available food image datasets, including Food-101 [17], UEC-FOOD-100 [20] and UEC-FOOD-256 [22]. The food images were selected in such a way that they could cover a wide variety of food items. This could help to train a strong classifier that can detect food images with a wide variety. In addition, images containing other objects or people in which food is not even the main target are also considered as food image. Every image was visually inspected by us such that it is distinguishable by a human observer in terms of its belongingness to one of the two classes: food and non-food.

For non-food images, we randomly selected 2,500 from existing image datasets consisting of general non-food objects or humans. These datasets include Caltech101 [30], Caltech256 [31], the Images of Groups of People [32] and Emotion6 [33]. We tried to cover a wide range of contents in the non-food images and included some non-food images visually similar to food, thus increasing the difficulty of classification task. For the training phase, we used 3,000 images with 1,500 for food and 1,500 for non-food. The rest of the dataset was equally divided into two subsets, with 500 images for each class in each subset, for validation and evaluation respectively. Figure 1 shows some examples of food and non-food images in Food-5K.


**Figure 2: Example images of Food-11 dataset.**

**Table 1: Food items and number of images in Food-11.**

| Category | Example items | Training | Validation | Evaluation |
|---|---|---|---|---|
| Bread | Bread, burger, pizza, pancakes, etc. | 994 | 362 | 368 |
| Dairy products | Milk, yogurt, cheese, butter, etc. | 429 | 144 | 148 |
| Dessert | Cakes, ice cream, cookies, chocolates, etc. | 1500 | 500 | 500 |
| Egg | Boiled and fried eggs, and omelette. | 986 | 327 | 335 |
| Fried food | French fries, spring rolls, fried calamari, etc. | 848 | 326 | 287 |
| Meat | Raw or cooked beef, pork, chicken, duck, etc. | 1325 | 449 | 432 |
| Noodles/Pasta | Flour/rice noodle, ramen, and spaghetti pasta. | 440 | 147 | 147 |
| Rice | Boiled and fried rice. | 280 | 96 | 96 |
| Seafood | Fish, shellfish, and shrimp; raw or cooked. | 855 | 347 | 303 |
| Soup | Various kinds of soup. | 1500 | 500 | 500 |
| Vegetable/Fruit | Fresh or cooked vegetables, salad, and fruits. | 709 | 232 | 231 |
| Total | | 9866 | 3430 | 3347 |

## 4.2 Dataset 2: Food-11

Food-11 dataset consists of 16,643 images grouped into 11 categories, which basically cover the major types of food that people consume in daily life. We defined the food categories by adopting and modifying the major food groups defined by United States Department of Agriculture (USDA) [34]. The 11 categories are: *Bread*, *Dairy products*, *Dessert*, *Egg*, *Fried food*, *Meat*, *Noodles/Pasta*, *Rice*, *Seafood*, *Soup* and *Vegetable/Fruit*. The dataset was mainly collected from existing food image datasets including Food-101 [17], UEC-FOOD-100 [20] and UEC-FOOD-256 [22]. For certain categories (*Diary products* and *Vegetable/Fruit*), we downloaded images from social media sites, Flickr and Instagram. For each food category, we tried to include different food items in order to increase the difficulty of recognition. Apart from this, only those images whose main content is food of that particular category were selected. The concrete example food items in each category, and the number of images for each subset are listed in Table 1. Figure 2 shows example food images of the 11 categories.

## 4.3 Dataset 3: IFD

In [9], Kagaya built a dataset called Instagram Food/Non-Food Dataset (IFD) from search results of #tag "food" in Instagram and manually annotated with food and non-food labels. The dataset consists of 4,230 food images and 5,428 non-food images. In [9], the food/non-food classification experiments conducted on IFD dataset resulted in a maximum accuracy of 95.1%. We used this dataset in our experiments to evaluate the performance of our trained model and to compare with the classification results in [9].

# 5. EXPERIMENTAL RESULTS

This section describes the experiments on food/non-food classification and food category recognition carried out using different datasets. In our experiments, we used Caffe [35] as

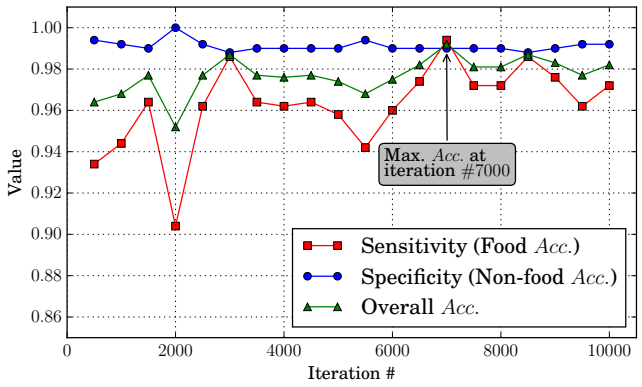
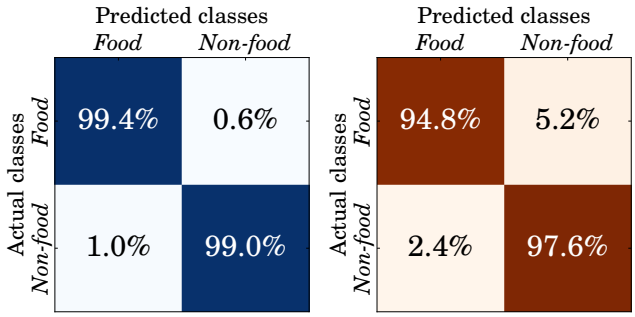



**Figure 3: Food/non-food classification results on Food-5K dataset.**




(a) Food-5K dataset (b) IFD dataset [9]

**Figure 4: Confusion matrix of food/non-food classification results on two different image datasets.**

the CNN library, which is one of the most popular frameworks for deep convolution neural network. A pre-trained GoogLeNet model has been applied and fine-tuned using our dataset in both food/non-food classification and category recognition. In particular we provide details on how the refinement of the model was achieved.

## 5.1 Food/Non-food Classification

Food/Non-food classification, or food image detection, is one of the initial and important steps for image-based dietary assessment. To classify food and non-food images, we used a pre-trained GoogLeNet model from [36] and fine-tuned it using the training subset of Food-5K dataset. Fine-tuning process takes a pre-trained model, adapts the architecture, and resumes training from the already learned model weights. When fine-tuning a pre-trained GoogLeNet model, we can choose the layers whose parameters should be updated. We have not used any pre-processing and post-processing steps. Firstly, we made the following basic changes in the CNN GoogLeNet model:

- All the three output layers names have been changed, e.g., "loss3/classifier" was changed to "loss3/classifier_Food". The reason for changing the layers names is that there should not be any conflict when the original weights are being read from the pre-trained model.

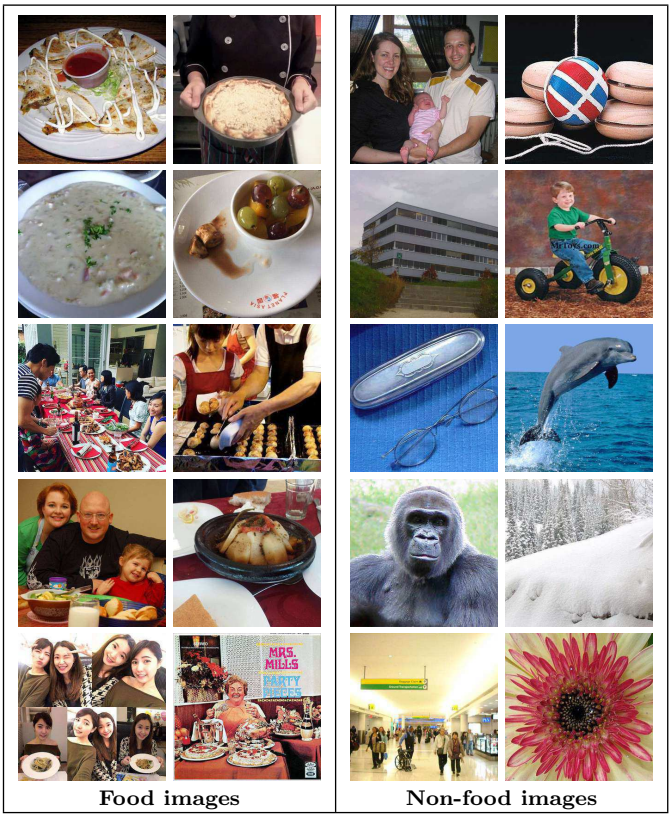



**Figure 5: Examples of correctly classified food and non-food images in Food-5K dataset.**

- Number of output layers has been changed from 1000 to 2 as we have only 2 classes: food and non-food.
- The base learning rate `Base_lr` has been changed to 0.01 and learning rate policy is `polynomial`.
- The maximum number of iteration, `Max_iter`, has been changed to 10000.

Then we set up two configurations to fine-tune the GoogLeNet model, with one only updating the parameters of the last two layers and the other for the last six layers. The overall classification accuracies of the two configurations for different iterations are shown in Table 2, with the overall accuracy *Acc.* defined as follows:

$$Acc. = \frac{TP + TN}{TP + FP + TN + FN}, \quad (1)$$

where $TP$, $FP$, $TN$ and $FN$ refer to true positive, false positive, true negative and false negative respectively. In most of the cases especially for higher number of iterations, higher accuracy is achieved on the second setup of fine-tuning i.e. the last six layers of GoogLeNet model. Therefore, we kept using the setup of fine-tuning the last six layers in the remaining experiments.

Figure 3 shows the detailed results of food/non-food classification on the evaluation subset of Food-5K, for all the iterations. In the result, the sensitivity, or true positive rate, indicates the rate of correctly detected food images. While, the specificity, or true negative rate, refers to the rate of correctly detected non-food images. From Figure 3, a maximum accuracy rate of 99.2% was achieved on evaluation

Table 2: Classification accuracy of two different fine-tuning configurations.

| Iteration # | 1,000 | 2,000 | 3,000 | 4,000 | 5,000 | 6,000 | 7,000 | 8,000 | 9,000 | 10,000 |
|---|---|---|---|---|---|---|---|---|---|---|
| Fine-tuning last 2 layers | **0.976** | **0.969** | 0.953 | **0.983** | 0.972 | 0.970 | 0.979 | 0.980 | 0.978 | 0.979 |
| Fine-tuning last 6 layers | 0.968 | 0.952 | **0.987** | 0.976 | **0.974** | **0.975** | **0.992** | **0.981** | **0.983** | **0.982** |


**Figure 6: Misclassified food and non-food images in Food-5K dataset.**

dataset at iteration #7000, with sensitivity and specificity of 99.4% and 99.0% respectively.

Figure 5 shows some examples of correctly detected food and non-food images for iteration #7000. It can be seen that some images that have even very small regions of food are correctly classified as food and some food-like non-food images are correctly classified as non-food, e.g., the fake Macaron lookalike. Figure 6 shows the incorrectly classified food and non-food images for iteration #7000. Some non-food images that were classified as food images are highly similar to food images and those food images classified as non-food images are either ambiguous or containing a very small region of food. Figure 4(a) shows the confusion matrix of food/non-food classification on our own dataset Food-5K.

To further evaluate the performance of our fine-tuned model on food/non-food classification, we ran our model on the other two datasets: Food-11 dataset created by us, and Instagram Food/Non-Food Dataset (IFD) by Kagaya et al. [9]. For both datasets, we tested our classifier on iteration #7000.

For Food-11 dataset, we ran our food/non-food classifier on all the 16,643 food images and 16,127 of them were correctly detected as food images, which results in a detection rate of 96.9%. Note that there are only food images in Food-11 dataset, and therefore the accuracy is just the rate of correctly detected food images. Figure 7 shows some examples of detected and undetected food images in Food-11 dataset.

For IFD dataset [9], we evaluated our model on 500 food and 500 non-food randomly selected images. The classification result is shown as confusion matrix in Figure 4(b). Among all the 500 food images, 474 (94.8%) were correctly classified as food, while 488 (97.6%) out of 500 non-food images were correctly classified as non-food. This resulted in an overall accuracy of 96.4%, which is slightly higher than the maximum accuracy of 95.1% obtained in [9].

## 5.2 Food Category Recognition

Correctly recognizing the type of a food in a food image is another crucial step for a dietary assessment system.


**Figure 7: Example of detected and undetected food images in Food-11 dataset.**

The aim of food categorization is to let the system either directly estimate the nutritional value of a food item using the general information about the food category, or further classify the food item into sub-category to have a better estimation. In this experiment, we used Food-11 dataset to train and test a CNN model on food category recognition. As explained in Section 4.2, the food images in Food-11 have been categorized into 11 classes and Table 1 shows the number of images in each category for training, validation and evaluation. Our task here was to classify each food image into one of the 11 categories. For this purpose, the pre-trained GoogLeNet model [36] was applied and the last six layers were fine-tuned on the training set of Food-11. We have not used any pre-processing and post-processing steps. Following changes have been made in the CNN GoogLeNet model:

- All the three output layers have been renamed, e.g., "loss3/classifier" was changed to "loss3/classifier_FoodReco", for the same reason as food/non-food classification in Section 5.1.
- The number of output layers has been changed from 1000 to 11 as we have 11 classes.


**Figure 8: Misclassified food and non-food images in Kagaya's IFD dataset [9].**

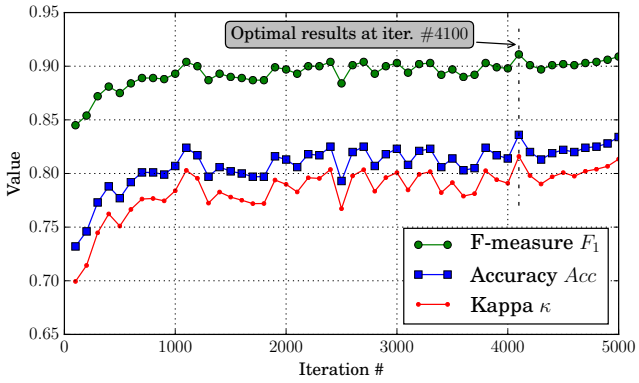



**Figure 9: Accuracy of food recognition on Food-11 dataset.**

- The base learning rate `Base_lr` has been changed to 0.001 and learning rate policy is `polynomial` .
- The maximum number of iteration, `Max_iter`, has been changed to 40,000.

We used three metrics to evaluate the performance of food recognition: 1) overall accuracy *Acc.*, 2) F-measure $F_1$ [37], and 3) Cohen's kappa coefficient $\kappa$ [38]. Specially, the Cohen's kappa coefficient is a numerical evaluation of inter-rater agreement which takes into account not only the observed classification accuracy but also the accuracy that any random classifier would be expected to achieve, namely, random accuracy. It is especially useful in evaluation of classification when the number of images in different categories are not the same.

Figure 9 shows the overall accuracy, F-measure and Cohen's kappa coefficient on the evaluation subset of Food-11 with respect to the number of iterations. The maximum accuracy of 83.5% was achieved on evaluation dataset at iteration #4100, where we also obtained the maximum values of F-measure and kappa coefficient, 0.911 and 0.816 respectively. The high value of Cohen's kappa coefficient (0.816) also indicates that the trained classifier performs significantly better than any random classifier. Due to time constraints, we had to stop evaluating the results on the evaluation dataset after iteration #5000, as the accuracy on the validation dataset did not show any significant improvement.

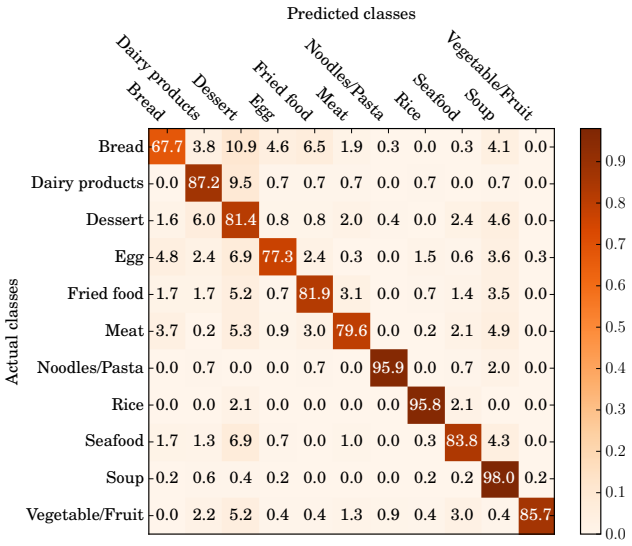



**Figure 10: Confusion matrix of food recognition. Values of the matrix are in percentage.**

The confusion matrix of recognition results at iteration #4100 is shown in Figure 10. Among all the classes, *Noodles/Pasta*, *Rice* and *Soup* give the best recognition accuracies, higher than 95%. This is because the food images in each category have their own common characteristics in either shape or color and therefore are easier to be identified. However, we notice that some types of food images are error-prone, e.g., *Bread*, *Egg* and *Meat*, accuracies of which are lower than 70%. Those three types of food are also the ones that have highly mixed food items in our dataset. For instance, category *Egg* contains boiled egg, fried egg and omelette, which are highly different in appearance. Besides, many of those images have the main food mixed with other food items, e.g., meat with salad. Interestingly, we observe that *Dessert* and *Soup* are the two target classes most likely to be misclassified. In 7 classes (*Bread*, *Dairy*, *Egg*, *Seafood*, *Meat*, *Fried food* and *Vegetable/Fruit*), more than 5% of their testing images were incorrectly classified as *Dessert*. This is because *Dessert* is the category that has the most mixed items in our dataset, and many of them could be visually similar to other food. Besides, more than 4% of images in *Bread*, *Dessert*, *Meat* and *Seafood* were misclassified as *Soup*. By checking some of images misclassified as soup, we found most of them have round-shaped elements such as plate or round bread. Most *Soup* images also have the similar round-shaped plates or containers.

According to the confusion matrix in Figure 10, we list the top 10 misclassified class pairs and show two example images for each in Figure 11. By observing the incorrectly classified images, we found that misclassification mostly happen in the following two cases:

1. Images within different classes have similar appearance, shape or color.
2. Images have more than one type of food items mixed.

Considering the fact that each image category in Food-11 dataset contains different food items with certain varieties, and that our training dataset is not significantly large, the


**Figure 11: Top 10 misclassified category pairs and example images. The percentage number indicates the proportion of incorrectly classified images in all testing images for the particular category.**

results we obtained ($Acc. = 0.835$, $F_1 = 0.911$ and $\kappa = 0.816$) seem promising.

## 6. CONCLUSION

In this paper, we applied a pre-trained GoogLeNet model based on CNN architecture on the tasks of food/non-food image classification and food category recognition. We constructed two image datasets from publicly available datasets and social media, and fine-tuned the GoogLeNet model using our datasets. The experimental results show the overall accuracy of 99.2% on food/non-food image classification and 83.6% on food categorization. The main reason for not achieving a high recognition accuracy on certain types of food images is complex mixture of food items in image and highly visual similarities between some images across categories. As a future direction, we aim at recognizing food items in images with a multi-label approach, namely, using top-$n$ as prediction output, and integrating contextual information to improve the accuracy and compare it with different architectures such as AlexNet, VGG, and ResNet. Further investigation will be done based on the different transfer learning schemes such as locking layers, etc. We will also work on the estimation of food items quantity and weight in order to finally estimate their nutritional value.

## 7. ACKNOWLEDGMENTS

This research work was supported by EPFL Food Science and Nutrition Center funding in the framework of NutriTake project, as well as partial funding from Swiss National Foundation for Scientific Research project LEADME 200020-149259. We also appreciate and acknowledge constructive comments from anonymous reviewers.

## 8. REFERENCES

[1] Giovanni Maria Farinella, Dario Allegra, Marco Moltisanti, Filippo Stanco, and Sebastiano Battiato. Retrieval and classification of food images. *Computers in Biology and Medicine*, 77:23 – 39, 2016.

[2] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178, 2006.

[4] Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image Classification with the Fisher Vector: Theory and Practice. *International Journal of Computer Vision*, 105(3):222–245, December 2013.

[5] Keigo Kitamura, Toshihiko Yamasaki, and Kiyoharu Aizawa. Food log by analyzing food images. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 999–1000, New York, NY, USA, 2008. ACM.

[6] J. Nie, Z. Wei, W. Jia, L. Li, J. D. Fernstrom, R. J. Sclabassi, and M. Sun. Automatic detection of dining plates for image-based dietary evaluation. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 4312–4315, Aug 2010.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[8] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. Food detection and recognition using convolutional neural network. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 1085–1088, New York, NY, USA, 2014. ACM.

[9] Hokuto Kagaya and Kiyoharu Aizawa. *Highly Accurate Food/Non-Food Image Classification Based on a Deep Convolutional Neural Network*, pages 350–357. Springer International Publishing, Cham, 2015.

[10] Taichi Joutou and Keiji Yanai. A food image recognition system with multiple kernel learning. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 285–288, Nov 2009.

[11] H. Hoashi, T. Joutou, and K. Yanai. Image recognition of 85 food categories by feature fusion. In

*Multimedia (ISM), 2010 IEEE International Symposium on*, pages 296–301, Dec 2010.

[12] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang. Pfid: Pittsburgh fast-food image dataset. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 289–292, Nov 2009.

[13] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar. Food recognition using statistics of pairwise local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2249–2256, June 2010.

[14] Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D. Abowd, and Irfan A. Essa. Leveraging context to support automated food recognition in restaurants. *CoRR*, abs/1510.02078, 2015.

[15] M. H. Rahmana, M. R. Pickering, D. Kerr, C. J. Boushey, and E. J. Delp. A new texture feature for improved food recognition accuracy in a mobile phone based dietary assessment system. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 418–423, July 2012.

[16] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp. Analysis of food images: Features and classification. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2744–2748, Oct 2014.

[17] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[19] Yoshiyuki Kawano and Keiji Yanai. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 589–593. ACM, 2014.

[20] 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.

[21] Keiji Yanai and Yoshiyuki Kawano. Food image recognition using deep convolutional network with pre-training and fine-tuning. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[22] 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.

[23] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. Im2calories: towards an automated mobile vision food diary. In *ICCV*, 2015.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Haohan Wang and Bhiksha Raj. A survey: Time travel in deep learning space: An introduction to deep learning models and how deep learning models evolved from the initial ideas. *CoRR*, abs/1510.04781, 2015.

[26] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013.

[27] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[28] Z. Zhong, L. Jin, and Z. Xie. High performance offline handwritten chinese character recognition using googlenet and directional feature maps. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pages 846–850, Aug 2015.

[29] Alex R. Kuefler. Merging recurrence and inception-like convolution for sentiment analysis.

[30] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, pages 178–178, June 2004.

[31] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[32] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.

[33] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, pages 860–868. IEEE, 2015.

[34] Marion Nestle. *Food politics: How the food industry influences nutrition and health*, volume 3. Univ of California Press, 2013.

[35] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

[36] https://github.com/BVLC/caffe/tree/master/models/bvlc_googlenet.

[37] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[38] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.