Research Highlights (Required)

- We investigate the use of CNN-based features for the purpose of food classification and retrieval.
- We compare different CNN architectures and we found that the most suitable architecture is a Residual Network with 50 layers (ResNet-50).
- We evaluate features extracted from the ResNet-50 fine-tuned and trained on food datasets having different food-domain representativeness.
- We introduce a new benchmark food database, Food-475, which contains 475 food classes and 247,636 images.
- Our results show that the most robust features for food classification and retrieval are those obtained from the ResNet-50 fine-tuned on the Food-475 database.

CNN-based Features for Retrieval and Classification of Food Images

Gianluigi Ciocca[a,*], Paolo Napoletano[a], Raimondo Schettini[a]

[a]*DISCo (Dipartimento di Informatica, Sistemistica e Comunicazione), Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy*

ABSTRACT

Features learned by deep Convolutional Neural Networks (CNNs) have been recognized to be more robust and expressive than hand-crafted ones. They have been successfully used in different computer vision tasks such as object detection, pattern recognition and image understanding. Given a CNN architecture and a training procedure, the efficacy of the learned features depends on the domain-representativeness of the training examples. In this paper we investigate the use of CNN-based features for the purpose of food recognition and retrieval. To this end, we first introduce the Food-475 database, that is the largest publicly available food database with 475 food classes and 247,636 images obtained by merging four publicly available food databases. We then define the food-domain representativeness of different food databases in terms of the total number of images, number of classes of the domain and number of examples for class. Different features are then extracted from a CNN based on the Residual Network with 50 layers architecture and trained on food databases with diverse food-domain representativeness. We evaluate these features for the tasks of food classification and retrieval. Results demonstrate that the features extracted from the Food-475 database outperform the other ones showing that we need larger food databases in order to tackle the challenges in food recognition, and that the created database is a step forward toward this end.

1. Introduction

Automatic food recognition is an important task for automatic or semi-automatic daily dietary monitoring. Nowadays, technology can support the users in keep tracks of their food consumption in a more user friendly way allowing for a more comprehensive daily dietary monitoring. Computer vision techniques can help to build systems to automatically locate and recognize diverse foods as well as to estimate the food quantity. Many works exist in the literature that exploit hand-crafted visual features for food recognition and quantity estimation both for desktop and for mobile applications (He et al., 2014; Nguyen et al., 2014; Bettadapura et al., 2015; Ciocca et al., 2015; Akpro Hippocrate et al., 2016; Pouladzadeh et al., 2016; Mezgec and Koroušić Seljak, 2017). Features learned by deep Convolutional Neural Networks (CNNs) have been recognized to be more robust and expressive than hand-crafted ones. They have been successfully used in different computer vision tasks such as object detection, pattern recognition and image

*Corresponding author: Tel.: +39 0264487922; fax: +0-000-000-0000;
e-mail: ciocca@disco.unimib.it (Gianluigi Ciocca), napoletano@disco.unimib.it (Paolo Napoletano), schettini@disco.unimib.it (Raimondo Schettini)

understanding. A number of studies have investigated the use of deep neural networks also for food recognition as in Yanai and Kawano (2015); Martinel et al. (2016); Fu et al. (2017); Ciocca et al. (2017a,b); Mezgec and Koroušić Seljak (2017); Ciocca et al. (2018). The most common food recognition paradigm is classification. This parading requires a set of annotated images (Bianco et al. (2013, 2015)). However, in real applications the amount of examples needed to train a robust classifier may not be always available. In this case, the food retrieval paradigm can be used to find similar foods among the available ones and to suggest a possible food class. Retrieval can be also exploited by humans to ease the tedious task of annotating food images. Regardless of the paradigm chosen, existing food recognition methods are usually benchmarked on a single database. Although this methodology is useful for evaluation and comparison, this could limit the generalization properties of the recognition algorithms. One of the reasons is that databases may have an unintentional bias such as: similar image acquisition conditions across image classes, similar compositions, point of view, etc (Tommasi et al., 2017; Torralba and Efros, 2011). Having a more heterogeneous food database would help to perform a more general and less database-specific recognition benchmark. This would also allow the research community to better evaluate the goodness of the existing and future recognition approaches. Moreover, a more heterogeneous database would help to train a CNN that can be then used to extract features that are more robust than features obtained from a CNN trained on a specific food database.

The contribution of the present work is twofold. Firstly we introduce a very large and heterogeneous food database obtained by carefully merging databases from the state-of-the-art and thus creating the largest food database available in the literature with 475 food classes and 247,636 images. This database, denoted as Food-475, is an evolution of the Food-524 database that we presented in Ciocca et al. (2017b). Food-524 contains 524 food classes obtained by syntactically merging food class names of four existing databases. This means that food classes denoted with a different name but representing the same food dish were considered as separate classes. Food-475, instead, contains 475 food classes obtained after applying a semi-automatic merging procedure that considers semantically equivalent food classes.

Secondly, we evaluate different CNN-based features learned from a CNN trained on different, publicly available, databases. We categorize the databases based on their food-domain representativeness. We define the food-domain representativeness of a database in terms of total number of images, number of food classes, and average number of images per classes. Among the considered food databases, the features learned on Food-475 and Food-524 databases, are the best performing ones in all our experiments. The results of these experiments confirm our intuition that, in order to have robust features for both food classification and retrieval, we need a large database, such as Food-475, for training that is truly representative for the food domain.

2. Related work

In this section we describe previous work in food image classification and retrieval with a special attention to CNN-based approaches.

One of the first works that used Deep Learning within the context of food recognition was by Kawano and Yanai (2014b). The food images are described with the features extracted from the FC7 layer of an AlexNet-style architecture pretrained on ImageNet. Also combinations of CNN-based features and hand-crafted features are considered. Images are then classified using a Support Vector Machine (SVM). The approach is evaluated on the UECFOOD-100 (Matsuda et al., 2012) and UECFOOD-256 (Kawano and Yanai, 2014a) datasets. Yanai and Kawano (2015), evaluated different CNN techniques for food recognition. These techniques include using network pre-trained with the large-scale ImageNet data, fine-tuned network for food classification, and the use of the activation features extracted from the CNN. The fine-tuning techniques achieve the best overall results on the UECFOOD databases as well as on the Food-101 database (Bossard et al., 2014). In Ciocca et al. (2017a) the AlexNet network is used as feature extraction module for classification of food images acquired in a canteen environment. Classification is performed either using k-NN or SVM classifier. The learned features outperforms all the hand-crafted features considered even though they were not specifically learned on food images.

Hassannejad et al. (2016), used the Google's image recognition architecture Inception V3. The network, composed of 54 layers, was designed to tackle the ImageNet's ILSVRC15 and it was fine tuned for classifying food images on the UECFOOD databases. The network is able to greatly surpass the performances of previous approaches. Also the approach of Liu et al. (2016), DeepFood, is based on the Inception structure. In this case, the Inception module is modified by introducing 1×1 convolutional layers to reduce the input dimension to the next layers. These modifications allow a less complex network but with some loss in performances. Martinel et al. (2016) devised the WIde-Slice Residual Network (WISeR) designed to specifically handle structures that can be found in food images. The network is able to outperform the Inception V3 architecture. CNNs can be used to tackle different tasks simultaneously. Chen and Ngo (2016) used this ability to build a deep convolutional neural network architecture for simultaneous food ingredients recognition and food categorization. Food classification results are not improved with respect to the state-of-the art, but the food ingredients recognition results are promising. The more complex CNNs have many parameters and require much time to train. Mezgec and Koroušić Seljak (2017) proposed a modified version of the Alexnet architecture (NutriNet) which uses fewer parameter compared with the original design, and is trained on a very large food database of more than 130,000 images. The proposed architecture performs slightly worse than methods based on the Residual Network architecture ResNet (He et al., 2016). Finally, most of the CNN approaches in the state-of-the-art are evaluated on single databases. In our previous experiments we investigated the use of a very large and heterogeneous food database in order to design a more robust food classification

Table 1. Performances of food classification methods using deep learning techniques.

Database	Network	Reference	Top-1 (%)	Top-5 (%)
UECFOOD-100	DeepFoodCam	Kawano and Yanai (2014b)	72.26	92.00
	DeepFood	Liu et al. (2016)	76.30	94.60
	CNN-FOOD(ft)	Yanai and Kawano (2015)	78.48	94.85
	ResNet(APL)	Fu et al. (2017)	80.60	95.90
	Inception V3	Hassannejad et al. (2016)	81.45	97.27
	MultiTaskCNN	Chen and Ngo (2016)	82.12	97.29
	WISeR	Martinel et al. (2016)	89.58	99.23
UECFOOD-256	DeepFood	Liu et al. (2016)	54.70	81.50
	DeepFoodCam	Kawano and Yanai (2014b)	63.77	85.82
	CNN-FOOD(ft)	Yanai and Kawano (2015)	67.57	88.97
	ResNet(APL)	Fu et al. (2017)	71.20	91.10
	Inception V3	Hassannejad et al. (2016)	76.17	92.58
	WISeR	Martinel et al. (2016)	83.15	95.45
Food-101	CNN-FOOD(ft)	Yanai and Kawano (2015)	70.41	-
	DeepFood	Liu et al. (2016)	77.40	93.70
	ResNet(APL)	Fu et al. (2017)	78.50	94.10
	Inception V3	Hassannejad et al. (2016)	88.28	96.88
	WISeR	Martinel et al. (2016)	90.27	98.71
ChinFood1000	ResNet(APL)	Fu et al. (2017)	44.10	68.40
VIREO	MultiTaskCNN	Chen and Ngo (2016)	82.12	97.29
NurtiNet-DB	NutriNet	Mezgec and Koroušić Seljak (2017)	86.72	-
Food-524	ResNet-50	Ciocca et al. (2017b)	81.34	95.45

approach Ciocca et al. (2017b). The database, named Food-524, contains 524 food classes for a total of more than 240,000 images.

Table 1 summarizes different food classification approaches. The Top-1 and Top-5 accuracies are reported along with the food databases used in the evaluation and the underlying CNN architecture.

Compared to food classification, there are few CNN-based approaches about food retrieval so we have included the most recent methods using hand-crafted features as well. In Farinella et al. (2016) different image representations are evaluated on the UNICT-FD1200 database that has been specifically created for the task of food retrieval. The database contains 1,200 food categories and each food plate is acquired multiple times under different geometric and photometric conditions. The image representations are based on SIFT, Textons and LBP features. Textons obtained the best results among the representations. Ciocca et al. (2017b) evaluated a CNN-based image representation, extracted from a CNN model based on the ResNet-50 architecture and fine-tuned on the Food-524 database. The results obtained on the UNICT-FD1200 confirm the strengths of the learned features with respect to hand-crafted ones.

Table 2 lists the works in the literature specifically dealing with food retrieval.

Table 2. Performances of food retrieval methods using both deep learning techniques and hand crafted features.

Reference	Features	Database	MAP (%)
Farinella et al. (2016)	Bag of SIFT 1200	UNICT-FD1200	29.14
	Textons (MR8) - RGB - Global	UNICT-FD1200	77.00
	Textons (Schmidt) - Lab - Global	UNICT-FD1200	90.06
Ciocca et al. (2017b)	F-ResNet-50 (ImageNet)	UNICT-FD1200	94.15
	F-ResNet-50 (Food-524)	UNICT-FD1200	96.56

3. From Food-524 to Food-475

In this section we introduce the Food-475 database that is an evolution of the Food-524 database. The Food-524 was created by combining the databases Food-50 by Joutou and Yanai (2009), Food-101 by Bossard et al. (2014), UECFOOD-256 by Kawano and Yanai (2014a) and VIREO by Chen and Ngo (2016). In Ciocca et al. (2017b) we combined these databases and merged duplicated classes via *syntactic* analysis of the labels. Food-524 resulted in one of the largest food database available. However, some equivalent food classes, that passed the syntactic analysis, still remains in the database. In this work, before using it in our experiments of food recognition, we decided to further process the images and merge *sematically* equivalent food classes. Belonging to these classes are those with different names for the same food either because the names are translated differently (e.g. "gyoza" vs. "jaiozi" vs "fried dumplings", or "xiaolongbao" vs. "steamed bun") or because the naming convention used in different databases are quite diverse (e.g. "fish & chip" vs. "fish and chips", or "dish consisting of stir-fried potato,eggplant and green pepper" vs. "fried potato, green pepper & eggplant"). See Figure 1 for some visual examples.

Manually comparing every pair of classes would require to inspect 56,316 unique class pairs (ignoring the intra-class food pairs). In order to cope with this number, we devised a semi-automatic procedure in order to speed up the identification of the classes to be merged. The procedure is defined as follows:

1. Every image in the database is described in terms of CNN-based features as in Ciocca et al. (2017b);
2. Every image is compared against all the other images in the database using the Euclidean distance;
3. The images are sorted according to their distance from the query and the top k images are selected (we set k=50). We consider the top k images as the most similar to the query;
4. A co-occurence matrix M is constructed by accumulating the evidences that an image of class i is similar (i.e. in the top k) to an image of class j (i.e. the query). We accumulate these evidences for all the queries and then transform them into probabilities $p(i, j)$ by dividing the number of occurrences of the given pair of labels against the total number of occurrences in the matrix M;

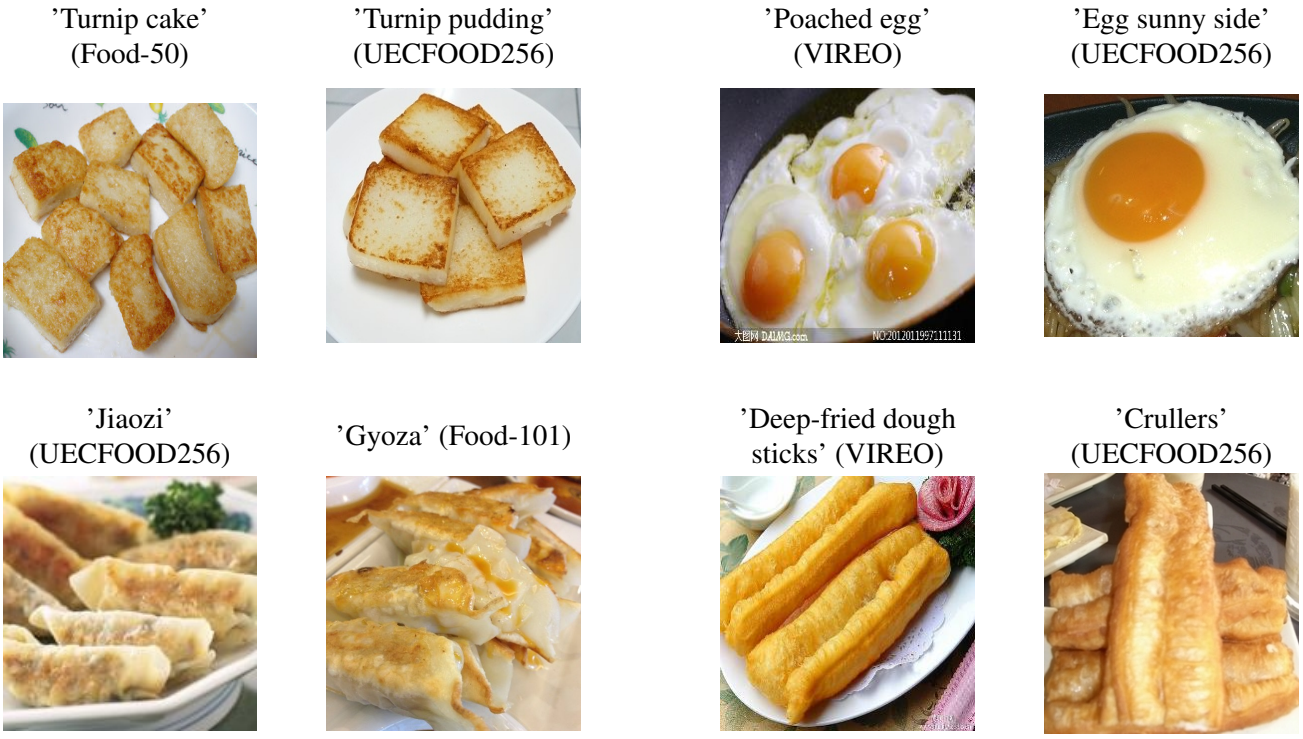

Fig. 1. Examples of *semantically* related food classes that can be found in the Food-524 database.

5. Starting from the highest $p(i, j)$ with $i \neq j$, the images in the corresponding classes are visually inspected to verify that they are indeed semantically equivalent;
6. A list of "to-be-merged" classes is compiled.

With the above procedure we evaluated 10,385 class pairs corresponding to about 18% of the total number of pairs, and covering about 85% of the total number of food co-occurrences. At the end of the inspection, we were able to identify 49 class pairs that are semantically equivalent. These pairs are then merged thus reducing the original 524 food classes to 475 classes. Figure 1 shows some examples of merged food class pairs. During the inspection phase, we also found some very challenging classes with images exhibiting strong visual similarities that are very difficult to disambiguate even for humans. Figure 2 shows some examples of these classes. For instance, the two scrambled eggs images are quite similar (top left pairs) and can be taken for the same food. However the loofah and bitter melon pieces, under close inspection, present small differences in shape. The donuts and bagels (top right pairs) have the same shape and, depending on the dressing, they may have the same visual appearance. The last two image pairs (bottom row) can be distinguished only by the ingredients since they are similarly prepared and presented. This kind of food classes have not been merged. The database can be downloaded from the following link `http://www.ivl.disco.unimib.it/activities/food475db/`.

Fig. 2. Examples of visually challenging food classes containing images difficult to disambiguate.

4. Proposed CNN-based features for food classification and retrieval

CNNs are a class of learnable architectures adopted in many domains such as image recognition, image annotation, image retrieval etc. (Schmidhuber, 2015). CNNs are usually composed of several layers, each involving linear as well as non-linear operators, that are learned jointly, in an end-to-end manner, to solve a particular tasks. A CNN architecture for image classification includes several convolutional layers followed by one or more fully connected layers. The output of the CNN is the output of the last fully connected layer. The number of output nodes is equal to the number of image classes (Krizhevsky et al., 2012).

A CNN that has been trained for solving a given task can be also adapted to solve a different task. It is not always possible to train an entire CNN from scratch, because it is relatively rare to have a dataset of sufficient size. It is common to use a CNN that is pre-trained on a very large dataset. For instance the ImageNet dataset, which contains 1.2 million images with 1000 categories (Deng et al., 2009). The pre-trained network is then used either as an initialization or as a fixed feature extractor for the task of interest (Razavian et al., 2014; Vedaldi and Lenc, 2014). If the network is used as feature extractor, the pre-trained CNN performs all the multilayered operations and, given an input image, the feature vector is the output of one of the last network layers (Vedaldi and Lenc, 2014). The use of CNNs as feature extraction method has demonstrated to be very effective in many pattern recognition applications (Razavian et al., 2014; Napoletano, 2018; Bianco et al., 2017; Cusano et al., 2016).

The first notably CNN architecture that has showed very good performance upon previous methods on the image classification task is the AlexNet (Krizhevsky et al., 2012) After the success of AlexNet, many other deeper architectures have been proposed

such as: VGGNet (Simonyan and Zisserman, 2014), GoogleNet (Szegedy et al., 2015), and Residual Networks (ResNet) (He et al., 2016). ResNet architectures has demonstrated to be very effective on the ILSVRC 2015 (ImageNet Large Scale Visual Recognition Challenge) validation set with a top-1 recognition accuracy of about 80%.

Due to its remarkable performances, the CNN-based features proposed in this paper have been obtained by exploiting a deep residual architecture. Residual architectures are based on the idea that each layer of the network learns residual functions with reference to the layer inputs instead of learning unreferenced functions. He et al. (2016) demonstrate that such architectures is easier to optimize and it gains accuracy also when the depth increase considerably. Our reference network architecture is based on the ResNet-50 which represents a good trade-off between depth and performance. The activations of the neurons in the fully connected layer are used as features for the retrieval of food images. The resulting feature vectors have size 2,048 components.

5. Food classification and retrieval experiments

Our experiments are organized into several steps:

1. We compared different notable CNN architecture with the aim to confirm that the ResNet-50 is the best performing one for food classification of the Food-475; (cf. Subsection 5.1);
2. We compared a fine-tuned ResNet-50 with a trained from the scratch (cf. Subsection 5.1);
3. We compared six ResNet-50 fine-tuned on the following databases for the food classification task: Food-50, UECFOOD-256, Food-101, VIREO, Food-524 and Food-475 (cf. Subsection 5.2);
4. We evaluated the features extracted using the fine-tuned CNNs obtained in the previous step for both food classification (cf. Subsection 5.3) and retrieval (cf. Subsection 5.4) tasks of the UNICT-FD1200 and Food-475 databases. In this evaluation we also considered features extracted from a pre-trained CNN on ILSVRC15 and the ones obtained by considering features pooling strategies.

5.1. Best performing CNN architecture choice

As stated in section 4, we proposed to extract CNN-based features exploiting the ResNet-50 architecture since residual networks proved very effective for classification in different application domains.

To validate our choice, we compared the ResNet-50 architecture against some of the most notable CNN architectures in the literature. To this end we focused only on the food classification task exploiting the fine-tuning training strategy as previously done in Martinel et al. (2016); Kawano and Yanai (2014b); Fu et al. (2017). Our assumption is that a robust food classification CNN is also able to extract robust features for food retrieval.

We used 75% of the Food-475 database for fine-tuning and 25% for performance evaluation. The network architecture included in the evaluation are: AlexNet (Krizhevsky et al., 2012), the reference Caffe implementation of the ImageNet (Jia et al.,

2013), GoogleNet (Szegedy et al., 2015), the very deep networks VGG-16 and VGG-19 (Simonyan and Zisserman, 2014), and the Inception V3 (Szegedy et al., 2016).

After having experimented with different settings, we found that the best classification results can be obtained by fine-tuning the networks via stochastic gradient descent with a mini-batch of 16 images, initial learning rate of 0.001 that decreases of a factor 10 at every 20K iterations. All the networks have been fine-tuned within the Caffe framework (Jia et al., 2014) on a PC equipped with a Tesla NVIDIA K40 GPU.

Results, in terms of Top-1 and Top-5 classification accuracy, are reported in Table 3. Classification accuracies of our ResNet-50 network are also reported. As it can be seen, the ResNet-50 exhibits the best results for the Top-1 with an accuracy of 81.59%. The runner up is the Inception V3 network with 74.46%. The VGG networks have similar results with the accuracy of the VGG-16 half a point better than the VGG-19 (73.94% against 73.57%). The AlexNet architecture shows the worse results with a 61.61% accuracy that is more than 20% percentage points lower than the accuracy of the ResNet-50. The Top-5 results exhibits a similar behavior as the Top-1 results.

These results confirm our intuition that the ResNet-50 is the most suitable architecture to be exploited for classification and feature extraction for food recognition and retrieval tasks.

As a further experiment, we compared the fine-tuned ResNet-50 with a trained from scratch one by using the same evaluation setup as before. We trained this network via stochastic gradient descent with a mini-batch of 24 images, initial learning rate of 0.1 that decreases of a factor 10 at every 50K iterations. The network have been trained within the Caffe framework on a Server machine equipped with two Tesla NVIDIA K80 GPUs.

For this network configuration, ResNet-50-S in Table 3, we obtained a classification accuracy of 69.45% that is more than 12 percentage points lower than accuracy of the fine-tuned ResNet-50. This is probably due to the fact that some of the food categories in the Food-475 database contain too few images to effectively train the network.

In the following, we will present experiments carried out using the fine-tuned ResNet-50 network.

5.2. Building food-domain CNN-based features

To evaluate different food-domain CNN-based features we compared six ResNet-50 networks fine-tuned on the following databases: Food-50, UECFOOD-256, Food-101, VIREO, Food-524 and Food-475. Each database has different characteristics: a different number of food classes, ranging from 50 to 524, and a different average number of examples for each class, ranging from about 100 to about 1,000. Considering these characteristics, the databases under consideration can be categorized in terms of food-domain representativeness, that is the more classes and examples for each food class are included in the database and the more the database is representative of the food domain. Figure 3 is a diagram showing the number of food classes (size of the

Table 3. Classification accuracy of the Food-475 database using different CNN architectures and transfer learning. 75% of the total number of images of each database is used for fine tuning and 25% is used for test. Suffix -S means that the network is trained from scratch.

Network	Top-1 (%)	Top-5 (%)
AlexNet (Food-475)	61.10	84.74
Caffe-Reference (Food-475)	61.43	85.20
GoogleNet (Food-475)	71.75	91.28
VGGNet-16 (Food-475)	73.94	92.28
VGGNet-19 (Food-475)	73.57	93.72
InceptionV3 (Food-475)	74.46	92.95
ResNet-50 (Food-475)	**81.59**	**95.50**
ResNet-50-S (Food-475)	69.45	91.01

Table 4. ResNet-50 classification accuracy on the food databases under consideration. 75% of the total number of images of each database is used for training and 25% is used for test.

Training and Test database	**Top-1 (%)**	**Top-5 (%)**
ResNet-50 (Food-50)	93.84	99.44
ResNet-50 (UECFOOD-256)	71.70	91.33
ResNet-50 (Food-101)	82.54	95.79
ResNet-50 (VIREO)	85.86	97.32
ResNet-50 (Food-524)	81.34	95.45
ResNet-50 (Food-475)	**81.59**	**95.50**

circle), average number of image for each class (x axis) and total number of images (y axis). For sake of comparison, the diagram includes all the databases considered and the ILSVRC15 database (ImageNet) Russakovsky et al. (2015), which does not contain food images but has been widely used as basis for domain transfer learning. From the diagram is quite clear that the ImageNet database, that is the smallest possible circle, contains the lowest number of food classes, the highest total number of images and highest number of images for each class. Food-475 and Food-524 contain the highest number of food images among all food databases, while Food-50 is the smallest food database considered. We have divided each database in approximately 75% training and 25% test according to the splits provided by the authors of the corresponding papers. In the case of Food-524 and Food-475, the training and test split has been obtained by merging the training and test sets of Food-50, UECFOOD-256, Food-101, and VIREO. All the networks have been fine-tuned using the same parameters as in the Subsection 5.1.

Table 4 shows the classification accuracy of the six ResNet-50 fine-tuned using the various food databases for both training and test. In the case of Food-50, UECFOOD-256, Food-101, and VIREO, the Top-1 and Top-5 accuracies are coherent with those obtained from state-of-the-art methods based on the same database. These results are not surprising if evaluated taking into account the food dataset characteristics showed in Figure 3. The results of ResNet-50 trained on the Food-524 database here reported differs

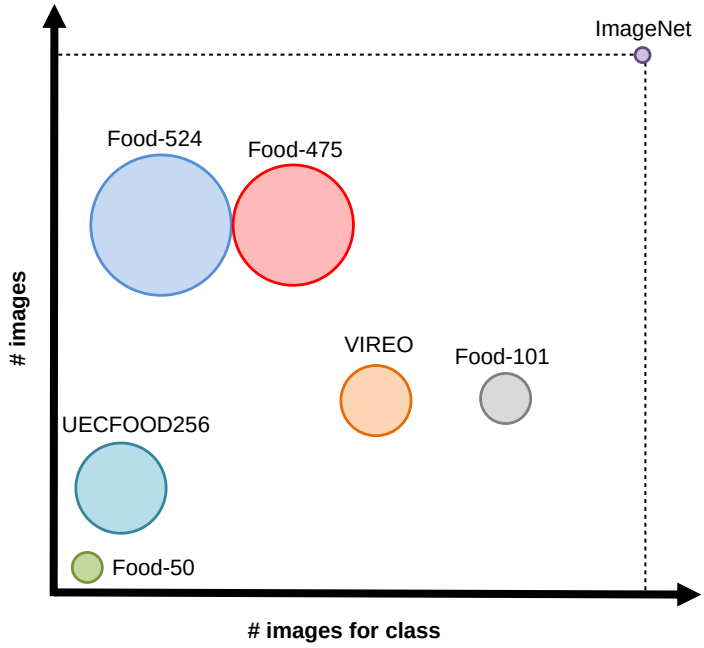

Fig. 3. Diagram of food-domain representativeness. The size of the circle is proportional to the number of food classes. The x axis represents the average number of images for each class. The y axis represents the total number of images contained in the database.

from the results reported in our previous work Ciocca et al. (2017b), where we trained the network using different parameters.

As baseline we have also considered the features extracted from a ResNet-50 pre-trained on the ImageNet database. Finally, for the sake of comparison, we have also experimented a meta feature vector obtained by combining the features extracted from each CNN trained on Food-50, UECFOOD-256, Food-101, and VIREO. We have experimented max and average pooling as ways to combine, and we report here only the results obtained with the best performing one, that is average pooling (AVG-POOL). Figure 4 shows how these features are obtained. In total we evaluated eight different CNN-based features.

The next two sections report the results of the eight CNN-based features evaluated for the tasks of food classification and food retrieval respectively. The evaluations are performed on the Food-475 and UNICT-1200 databases. The UNICT-FD1200 database (Farinella et al., 2016) has been chosen because it was specifically designed for food retrieval. It is composed of 4,754 images and 1,200 distinct dishes of food of different nationalities.

5.3. Food classification using CNN-based features

In order to test the effectiveness of the extracted CNN-based features for food classification, we chose to use a very simple k-Nearest Neighbour (kNN) classifier with $k = 1$. Features are compared using the Euclidean distance. As evaluation measure, we adopted the top-1 classification accuracy, that is the percentage of images correctly classified with respect to the total number of images. In the case of the Food-475 database, we considered its test set that contains approximately the 25% of the entire database, that is 65,404 images. The classification experiment was conducted as follows. Let N be the number of available test images. We considered each image as a test sample and performed the k-NN classification using the remaining $N-1$ images as training samples.

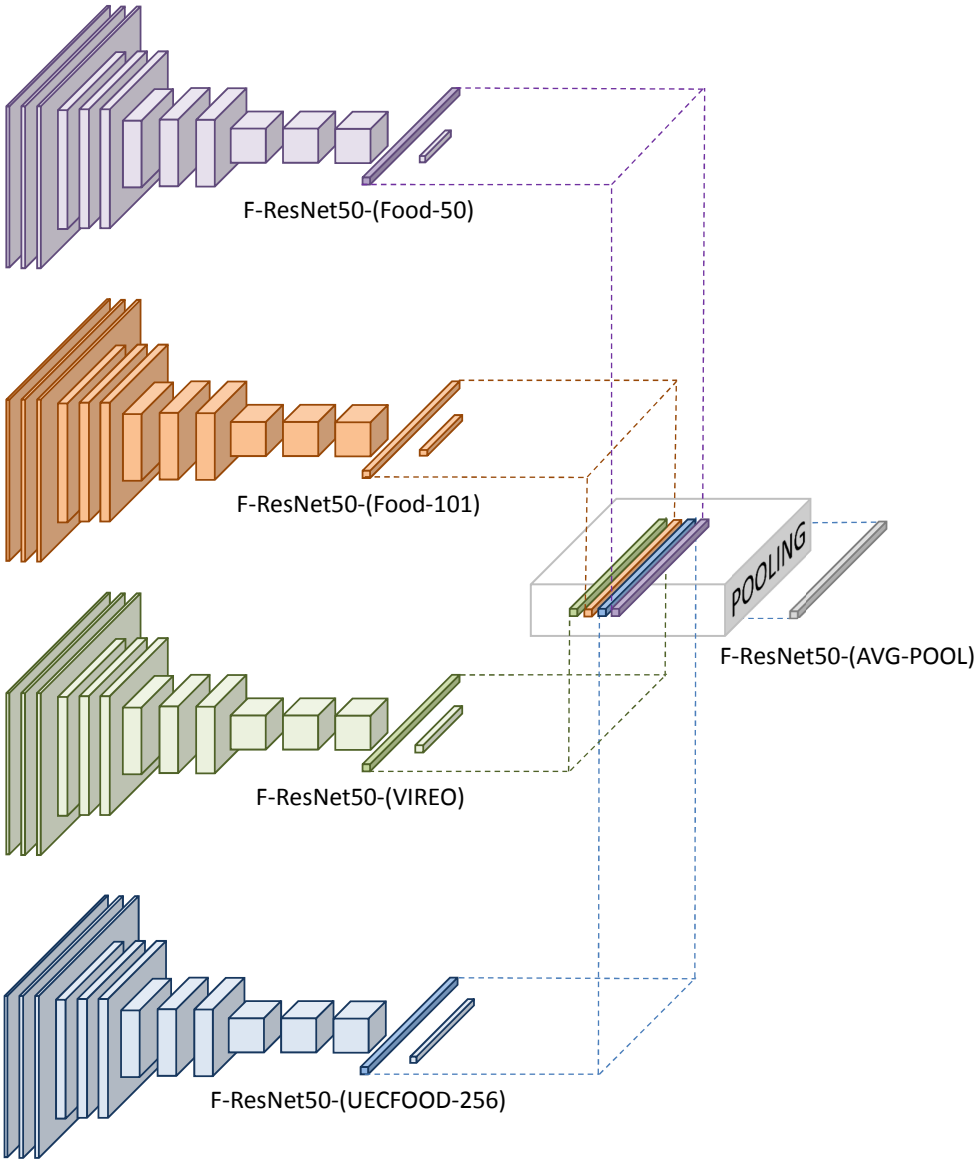

Fig. 4. Conceptual diagram of the average pooling (AVG-POOL) strategy for CNN-based feature extraction.

For the UNICT-FD1200 database, we followed the original evaluation protocol. Specifically, the food database is divided into a training set of 1,200 images and in a test set with the remaining ones. The three training/test splits provided by the authors of the database are considered. The overall classification accuracy is measured as the average accuracy on the three splits.

Classification results are shown in Table 5. For both databases the Top-1 accuracy suggests that CNN-based features obtained using Food-475 and Food-524 perform better than CNN-based features extracted form the ImageNet even though the training database is very large. In the case of the UNICT-FD1200 database the improvement obtained by the CNN-based features trained using Food-475, i.e. F-ResNet-50 (Food-475), with respect to ImageNet features, denoted as F-ResNet-50 (ImageNet), is not so high and it is about 5%. This is due to the fact that the UNICT-FD1200 database is made of many classes but a low number of examples of classes (4.7 on average). For this reason, the classification task of UNICT-FD1200 is not as challenging as in the case

Table 5. Classification results on the UNICT-FD1200 and Food-475 databases using k-NN.

CNN-based Features	UNICT-FD1200 Top-1 (%)	Food-475 Top-1 (%)
F-ResNet-50 (ImageNet)	91.84	40.46
F-ResNet-50 (Food-50)	91.26	37.76
F-ResNet-50 (UECFOOD-256)	94.54	42.17
F-ResNet-50 (Food-101)	95.31	57.95
F-ResNet-50 (VIREO)	94.96	57.92
F-ResNet-50 (AVG-POOL)	95.98	53.69
F-ResNet-50 (Food-524)	**96.56**	67.78
F-ResNet-50 (Food-475)	96.49	**68.01**

Table 6. Retrieval results on the UNICT-FD1200 and Food-475 databases.

CNN-based Features	UNICT-FD1200 mAP (%)	Food-475 mAP (%)
F-ResNet-50 (ImageNet)	94.15	7.43
F-ResNet-50 (Food-50)	93.76	7.57
F-ResNet-50 (UECFOOD-256)	96.25	8.84
F-ResNet-50 (Food-101)	96.79	19.97
F-ResNet-50 (VIREO)	96.54	24.92
F-ResNet-50 (AVG-POOL)	97.29	15.81
F-ResNet-50 (Food-524)	**97.71**	30.03
F-ResNet-50 (Food-475)	97.66	**31.56**

of Food-475.

The Top-1 accuracy reached by the CNN-based features trained using Food 50, that are denoted as F-ResNet-50 (Food-50), is quite similar to the top-1 accuracy reached by F-ResNet-50 (ImageNet). This is explained by fact that the Food-50 database contains a very low number of food classes (see also Figure 3). In the case of the Food-475 databases the improvement with respect to no food-domain features is quite high and it is about 28%. ResNet-50 trained using Food-475 and Food-524 achieve similar classification accuracy. The results obtained in both databases using the aggregated features, denoted as F-ResNet-50 (AVG-POOL), are lower than the F-ResNet-50 (Food-475) and F-ResNet-50 (Food-524). This shows that is more effective to have features obtained on merged databases, than merging the features themselves.

5.4. Food retrieval using CNN-based features

The retrieval experiments are also conducted on the UNICT-FD1200 and Food-475 databases. For the UNICT-FD1200, as in the original paper, the images in the training set are considered as database images, while the images in the test set are the queries. Moreover, for each query there are up to four correct images to be retrieved. For the experiments on the Food-475 database, retrieval is performed using only the test set and a one-vs-rest approach. An image of the test set is considered a query, and the remaining images in the test set as the target database. All the images in the test set have been evaluated as queries. The retrieval performances are measured using the $P(n)$ quality metric and the Mean Average Precision (MAP). The $P(n)$ is based on the top n criterion: $P(n) = Q_n/Q$, where Q is the number of queries (test images) and Q_n the number of correct queries among the first n

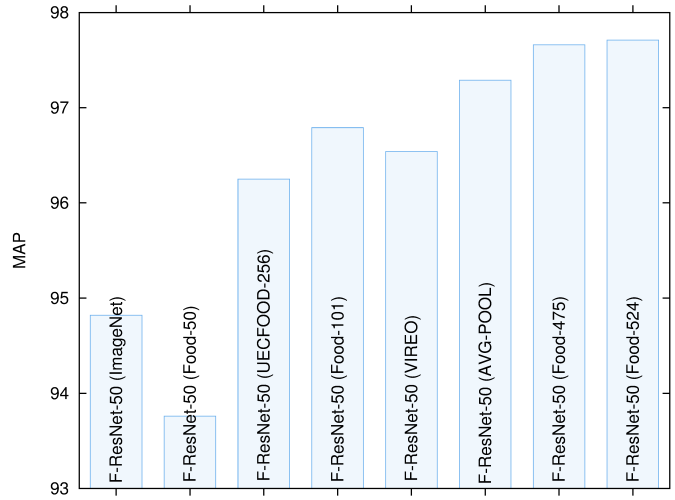

Fig. 5. Comparison of retrieval accuracies of the different CNN-based features on the UNICT-FD1200 database.

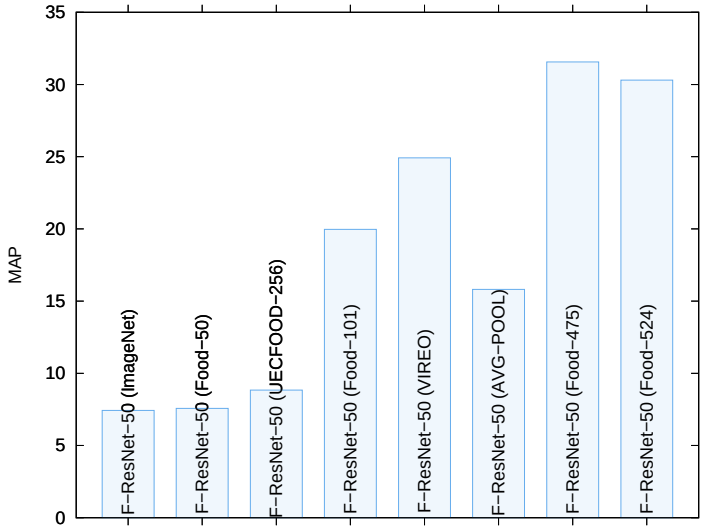

Fig. 6. Comparison of retrieval accuracies of the different CNN-based features on the Food-475 database.

retrieved images (Farinella et al., 2016).

Table 6 shows the retrieval results obtained on the UNICT-FD1200 and Food-475 databases. As it can be seen, the features computed on the Food-524 and Food-475 databases outperforms the other ones. Again, the more heterogeneous food database perform better than no food-domain features in both databases. In the case of UNICT-FD1200 the improvement is quite low, about 3%, while in the case of Food-475 the improvement is about 25%. As in the case of the classification task, features computed on the Food-524 and Food-475 databases have small performance differences.

Figures 5 and 6 summarize the MAP behaviour in both UNICT-FD1200 and Food-475 databases. In both cases can be observed that the larger is a food database the higher is the retrieval accuracy. Figures 7 and 8 show the $P(n)$ curves of the CNN-based features considered for both the UNICT-FD1200 and Food-475 databases. To make the curves more readable we show just a portions of them. It can be appreciated how the features extracted from databases with high food-domain representativeness are

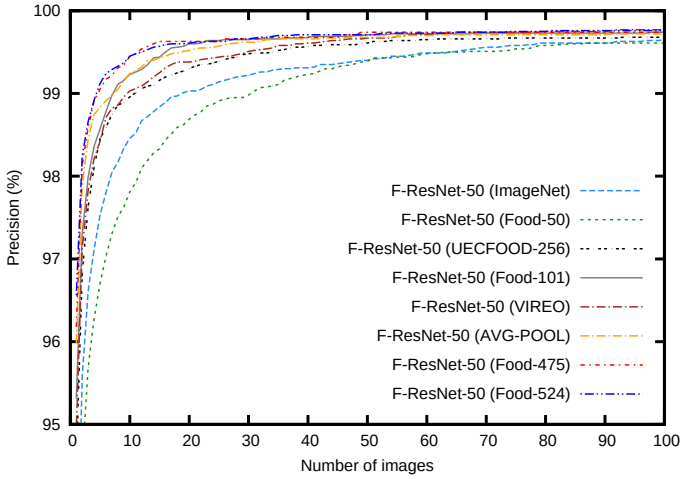

Fig. 7. Comparison between P(n) curves of the CNN-based features in the case of the UNICT-FD1200 database.

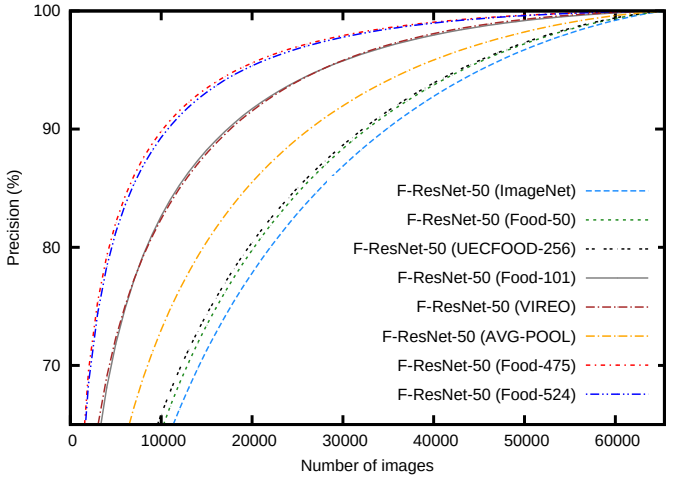

Fig. 8. Comparison between P(n) curves of the CNN-based features in the case of the Food-475 database.

able to effectively retrieve the relevant images in the first positions.

From the results in Table 6, and Figures 5, 6, 7, and 8, it is quite evident that the higher is the food-domain representativeness of the database is, the more the features learned using such database greatly improve the precision of the retrieval with respect to the other methods, allowing more relevant images to be returned in the first positions. As already noted in the classification task, the meta features, that is F-ResNet-50 (AVG-POOL), does not perform better than the F-ResNet-50 (Food-475) or F-ResNet-50 (Food-524).

To further evaluate the usefulness of these two features we have experimented with the pseudo-relevance feedback retrieval scheme. Following this scheme, after the initial query, the first n items returned by the system are considered as relevant to the initial query and then used to re-query the system. The final list of returned items is obtained by combining each list returned with respect to each query. As a result, if the initial query returns a high number of relevant items in the first positions, the result of the new query is likely more accurate. We performed these experiments on the Food-475 database because the UNICT-FD1200

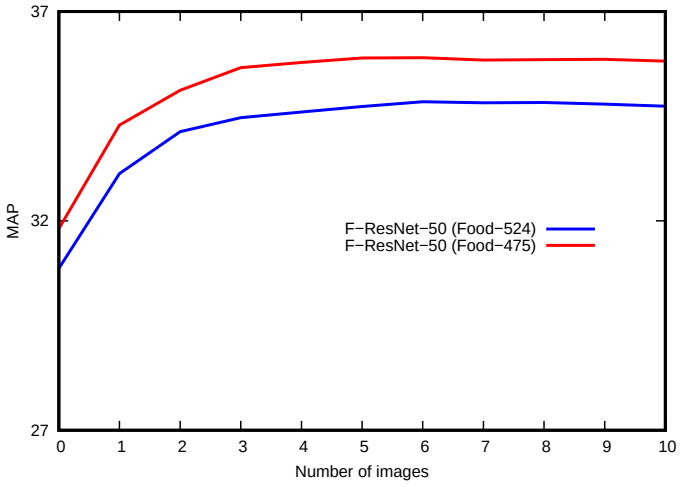

Fig. 9. Pseudo-relevance feedback experiments performed on the Food-475 database.

database contains, on average, only 4.7 images for each food class. Figure 9 shows the results obtained by experimenting the pseudo-relevance feedback (PRF) by exploiting different values of relevant images n ranging from 1 to 10. The use of PRF leads to an improvement of about 4% with two relevant images added to the initial query. Then the improvement remain constant as the number of relevant images added increase.

6. Conclusion

In this paper we present an evaluation of CNN-base features for food classification and retrieval. We use a Residual Network with 50 layers as a reference architecture to learn food-domain features. We argue that, in order to have robust features for food-related tasks, we need food-domain representative food database. To test this, we consider different food databases publicly available and categorize them according to their food-domain representativeness that we expressed through the total number of images, number of classes of the domain and number of examples for class. We also introduce a new food database, Food-475, that is a refinement of our previously proposed Food-524 food database. This database exhibits a higher food-domain representativeness with respect to the other databases considered. The features learned on the proposed database outperforms those learned on other food databases and on the very large ImageNet image database. Specifically, experiments on the test sets of the Food-475 and UNICT-FD1200 databases show that these features exhibits large improvements in accuracy for food classification and recognition tasks. This demonstrates that the more is representative the database for food domain and the more is accurate the recognition and retrieval of features obtained from a CNN trained on that database. The Food-475 is the largest, publicly available, database in the state of the art. We think that it will be of great interest for the scientific community for the design of more robust food recognition and retrieval algorithms. For this reason we will soon make the database available for download.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. Published in the context of the project FooDesArt: Food Design Arte - L'Arte del Benessere , CUP (Codice Unico Progetto - Unique Project Code): E48I16000350009 - Call "Smart Fashion and Design", cofunded by POR FESR 2014-2020 (Programma Operativo Regionale, Fondo Europeo di Sviluppo Regionale - Regional Operational Programme, European Regional Development Fund).

References

Akpro Hippocrate, E.A., Suwa, H., Arakawa, Y., Yasumoto, K., 2016. Food weight estimation using smartphone and cutlery, in: Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems, ACM. pp. 9–14.

Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G., Essa, I., 2015. Leveraging context to support automated food recognition in restaurants, in: Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on, pp. 580–587.

Bianco, S., Celona, L., Napoletano, P., Schettini, R., 2017. On the use of deep learning for blind image quality assessment. Journal of Signal, Image and Video Processing -.

Bianco, S., Ciocca, G., Napoletano, P., Schettini, R., 2015. An interactive tool for manual, semi-automatic and automatic video annotation. Computer Vision and Image Understanding 131, 88–99.

Bianco, S., Ciocca, G., Napoletano, P., Schettini, R., Margherita, R., Marini, G., Pantaleo, G., 2013. Cooking action recognition with ivat: an interactive video annotation tool, in: Int. Conf. on Image Analysis and Processing, Springer. pp. 631–641.

Bossard, L., Guillaumin, M., Van Gool, L., 2014. Food-101–mining discriminative components with random forests, in: Computer Vision–ECCV 2014, pp. 446–461.

Chen, J., Ngo, C.W., 2016. Deep-based ingredient recognition for cooking recipe retrieval, in: Proc. of the 2016 ACM on Multimedia Conference, ACM. pp. 32–41.

Ciocca, G., Napoletano, P., Schettini, R., 2015. Food recognition and leftover estimation for daily diet monitoring, in: New Trends in Image Analysis and Processing – ICIAP 2015 Workshops, pp. 334–341.

Ciocca, G., Napoletano, P., Schettini, R., 2017a. Food recognition: a new dataset, experiments and results. IEEE Journal of Biomedical and Health Informatics 21, 588–598.

Ciocca, G., Napoletano, P., Schettini, R., 2017b. Learning cnn-based features for retrieval of food images, in: New Trends in Image Analysis and Processing – ICIAP 2017, pp. 426–434.

Ciocca, G., Napoletano, P., Schettini, R., 2018. Ivlfood-ws: Recognizing food in the wild using deep learning, in: International Conference on Consumer Electronics - Berlin (ICCE-Berlin), IEEE. pp. –.

Cusano, C., Napoletano, P., Schettini, R., 2016. Combining multiple features for color texture classification. Journal of Electronic Imaging 25, 1–9.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 248–255.

Farinella, G.M., Allegra, D., Moltisanti, M., Stanco, F., Battiato, S., 2016. Retrieval and classification of food images. Computers in Biology and Medicine 77, 23–39.

Fu, Z., Chen, D., Li, H., 2017. Chinfood1000: A large benchmark dataset for chinese food recognition, in: International Conference on Intelligent Computing, Springer. pp. 273–281.

Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., Cagnoni, S., 2016. Food image recognition using very deep convolutional networks, in: Proceedings of the 2Nd International Workshop on Multimedia Assisted Dietary Management, ACM. pp. 41–49.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

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

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 .

Jia, Y., et al., 2013. An open source convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM International Conference on Multimedia (Orlando, Florida, USA), pp. 675–678.

Joutou, T., Yanai, K., 2009. A food image recognition system with multiple kernel learning, in: Image Processing (ICIP), 2009 16th IEEE International Conference on, IEEE. pp. 285–288.

Kawano, Y., Yanai, K., 2014a. 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.

Kawano, Y., Yanai, K., 2014b. Food image recognition with deep convolutional features, in: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 589–593.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Ma, Y., 2016. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment, in: Proceedings of the 14th International Conference on Inclusive Smart Cities and Digital Health - Volume 9677, pp. 37–48.

Martinel, N., Foresti, G.L., Micheloni, C., 2016. Wide-slice residual networks for food recognition. arXiv preprint arXiv:1612.06543 .

Matsuda, Y., Hoashi, H., Yanai, K., 2012. Recognition of multiple-food images by detecting candidate regions, in: Multimedia and Expo (ICME), 2012 IEEE International Conference on, pp. 25–30.

Mezgec, S., Koroušić Seljak, B., 2017. Nutrinet: A deep learning food and drink image recognition system for dietary assessment. Nutrients 9, 657.

Napoletano, P., 2018. Visual descriptors for content-based retrieval of remote-sensing images. International Journal of Remote Sensing 39, 1–34.

Nguyen, D.T., Zong, Z., Ogunbona, P.O., Probst, Y., Li, W., 2014. Food image classification using local appearance and global structural information. Neurocomputing 140, 242–251.

Pouladzadeh, P., Kuhad, P., Peddi, S.V.B., Yassine, A., Shirmohammadi, S., 2016. Food calorie measurement using deep learning neural network, in: IEEE International Instrumentation and Measurement Technology Conference, pp. 1–6.

Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: an astounding baseline for recognition, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on, pp. 512–519.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 211–252.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks 61, 85–117.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE CVPR conference, pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T., 2017. A deeper look at dataset bias, in: Domain Adaptation in Computer Vision Applications. Springer, pp. 37–55.

Torralba, A., Efros, A.A., 2011. Unbiased look at dataset bias, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 1521–1528.

Vedaldi, A., Lenc, K., 2014. Matconvnet – convolutional neural networks for matlab. CoRR abs/1412.4564.

Yanai, K., Kawano, Y., 2015. Food image recognition using deep convolutional network with pre-training and fine-tuning, in: 2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW), pp. 1–6.