Learning CNN-based Features for Retrieval of Food Images

Gianluigi Ciocca, Paolo Napoletano(✉), and Raimondo Schettini

DISCo (Dipartimento di Informatica, Sistemistica e Comunicazione), Università degli Studi di Milano-Bicocca, Viale Sarca 336, 20126 Milano, Italy
{ciocca,napoletano,schettini}@disco.unimib.it

Abstract. Recently a huge amount of work has been done in order to develop Convolutional Neural Networks (CNNs) for supervised food recognition. These CNNs are trained to classify a predefined set of food classes within a specific food dataset. CNN-based features have been largely experimented for many image retrieval domains and to a lesser extent to the food domain. In this paper, we investigate the use of CNN-based features for food retrieval by taking advantage of existing food datasets. To this end, we have built the Food524DB, the largest publicly available food dataset with 524 food classes and 247,636 images by merging food classes from existing datasets in the state of the art. We have then used this dataset to fine tune a Residual Network, ResNet-50, which has demonstrated to be very effective for image recognition. The last fully connected layer is finally used as feature vector for food image indexing and retrieval. Experimental results are reported on the UNICT-FD1200 dataset that has been specifically design for food retrieval.

Keywords: Food retrieval · Food dataset · Food recognition
CNN-based features

1 Introduction

Recently, food recognition received a considerable amount of attention due to the importance of monitoring food consumption for a balanced and healthy diet. To this end, computer vision techniques can help to build systems to automatically recognize diverse foods and 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 as well as mobile applications [1,3,17,27,28].

With the advent of practical techniques for training large convolutional neural networks, hand-crafted features are being reconsidered in favor of learned ones [30]. 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. It is not surprising that

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 426–434, 2017.
https://doi.org/10.1007/978-3-319-70742-6_41

a number of studies have investigated the use of deep neural networks for food recognition as well. Table 1 shows the most notable works on food recognition using deep learning techniques along with the datasets on which they have been evaluated their performances in terms of Top-1 and Top-5 classification accuracy.

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

Reference	Network	Dataset	Top-1 (%)	Top-5 (%)
Kawano et al. [22]	DeepFoodCam	UECFOOD-100	72.26	92.00
		UECFOOD-256	63.77	85.82
Yanai et al. [32]	DCNN-FOOD(ft)	UECFOOD-100	78.48	94.85
		UECFOOD-256	67.57	88.97
		Food-101	70.41	-
Liu et al. [23]	DeepFood	UECFOOD-100	76.30	94.60
		UECFOOD-256	54.70	81.50
		Food-101	77.40	93.70
Hassannejad et al. [15]	Inception V3	UECFOOD-100	81.45	97.27
		UECFOOD-256	76.17	92.58
		Food-101	88.28	96.88
Martinel et al. [25]	WISeR	UECFOOD-100	89.58	99.23
		UECFOOD-256	83.15	95.45
		Food-101	90.27	98.71
Chen et al. [6]	MultiTaskDCNN	UECFOOD-100	82.12	97.29
		VIREO	82.05	95.88

A Convolutional Neural Network technique requires a large dataset to build a classification model. To overcome this, often previously pre-trained models on a different dataset are fine tuned using a small sized dataset specific for the current classification task. Since the larger and heterogeneous the dataset is, the more the network can be used to learn powerful models, for the food retrieval task, we have decided to create a very large food dataset starting from existing ones. We have analyzed the public datasets and merged some of them depending on their availability and characteristics thus creating the largest food dataset available in the literature with 524 food classes and 247,636 images. The lowest number of images for a given class is 100 while the largest is about 1,700. We exploit this dataset for learning robust features for food retrieval using a Residual Network. Our intuition is that, having this dataset more food classes than the ones used in previous works, the network should be more powerful, generalizes better and thus the extracted features should be more expressive.

Table 2. List of food datasets used in the literature. S: Single instance food images. M: Multi-instance food images.

Name	Year	#Images	#Classes	Type	Reference
Food50	2009	5,000	50	S	[20]
PFID	2009	1,098[a]	61[a]	S	[7]
TADA	2009	50/256	-	S, M	[24]
Food85[b]	2010	8,500	85	S	[18]
Chen	2012	5,000	50	S	[8]
UECFOOD-100	2012	9,060	100	S, M	[26]
Food-101	2014	101,000	101	S	[5]
UECFOOD-256[c]	2014	31,395	256	S, M	[21]
UNICT-FD889	2014	3,583	889	S	[14]
Diabetes	2014	4,868	11	S	[2]
UMPCFood-101[d]	2015	90,993	101	S	[31]
UNIMIB2015	2015	1,000 × 2	15	M	[9]
UNICT-FD1200[e]	2016	4,754	1,200	S	[13]
UNIMIB2016	2016	1,027	73	M	[10]
VIREO	2016	110,241	172	S	[6]
Food524DB	2017	247,636	524	S	-

[a] Numbers refer to the baseline dataset.
[b] Includes Food50.
[c] Includes UECFOOD-100.
[d] Includes same classes of Food-101.
[e] Includes UNICT-FD889.

2 CNN-based Features for Food Retrieval

Domain adaptation, also known as transfer learning or fine tuning, is a machine learning procedure designed to adapt a classification model trained on a set of data to work on a different set of data. The importance and the usefulness of a domain adaptation process has been largely discussed in the food recognition literature [4,11,12,22,23,25,32]. Taking inspiration from these works, in this paper we fine-tuned a CNN architecture using a large, heterogeneous, food dataset, namely the Food524DB. The rational behind the creation of the Food524DB is that building a robust food recognition algorithm requires a large image dataset of different food instances.

2.1 The Food524DB Food Dataset

Table 2 summarizes the characteristic of the food datasets that can be found in the literature. For each dataset, we have reported its size, the number of food classes and the type of images it contains: either single, i.e. each image depict a single food category, or multi, i.e. the images can contain multiple food classes.

We decided to consider only datasets publicly available, with many food classes, and, most importantly, where each food category is represented by at least 100 images. After having analyzed the available datasets, we finally selected Food50, Food-101, UECFOOD-256, and VIREO. Since UECFOOD-256 contains multi-food instance images, we extracted from these images each food region using the bounding boxes provided in the ground truth. The combined dataset is thus composed of 247,636 images grouped in 579 food classes making this dataset the largest and most comprehensive food dataset available nad that can be used for training food classifiers. Some food classes are present in more than one of the four datasets. For example both the UECFOOD-256 and Food-101 contain the "apple_pie" category; UECFOOD-256 contains the "beef noodle" category while the VIREO dataset contains the "Beef noodles" category. In order to remove these redundancies we applied a category merging procedure based on the category names. After this procedure, the number of food classes in our dataset that we named Food524DB is reduced to 524 as reported in the last row of Table 2.

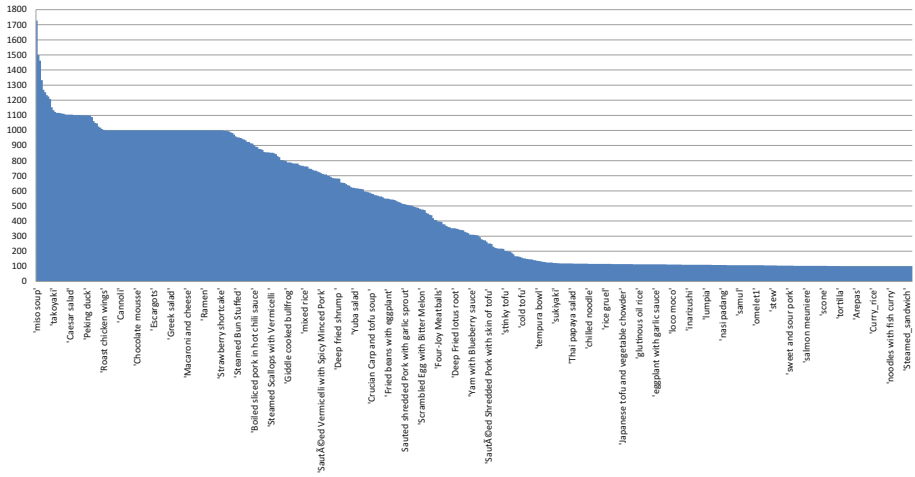

Fig. 1. Distribution of the cardinalities of the Food524DB food classes. Names are shown one every ten.

The sizes of the 524 food classes are reported in Fig. 1. The smallest food category contains 100 images; 241 classes have size between 100 and 199 images, 58 classes have size between 200 and 499 images, 113 have size between 500 and 999 images, and 112 have more than 1,000 images. The top-5 largest classes are: "Miso Soup" with 1,728 images; "Rice" with 1,499 images; "Spaghetti alla Bolognese" with 1,462 images; "Hamburger" with 1,333 images; and "Fried Rice" with 1,269 images. The Food524db is publicly available at http://www.ivl.disco.unimib.it/activities/food524db/.

2.2 CNN-based Food Features

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. Such architectures demonstrated to be easier to optimize and to gain accuracy by considerably increasing the depth [16].

Our network architecture is based on the ResNet-50 which represents a good trade-off between depth and performance. ResNet-50 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% [16]. We did not train the ResNet-50 from the scratch on Food524DB because the number of images for each class is not enough. As in previous work on this topic [22,25], we started from a pre-trained ResNet-50 on ILSVRC2012 scene image classification dataset [29]. The Food524DB dataset has been split in 80% of training data and 20% of test data. During the fine-tuning stage each image has been resized to 256×256 and a random crop has been taken of 224×224 size. We augmented data with the horizontal flipping. During the test stage we considered a single central 224×224 crop from the 256×256-resized image.

The ResNet-50 has been trained via stochastic gradient descent with a mini-batch of 16 images. We set the initial learning rate to 0.01 with learning rate update at every 5 K iterations. The network has been trained within the Caffe [19] framework on a PC equipped with a Tesla NVIDIA K40 GPU. The classification accuracy of the ResNet-50 fine-tuned with the Food524DB dataset is 69.52% for the Top-1, and 89.61% for the Top-5.

In the following experiments, the ResNet-50 is then used as feature extractor. 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.

3 Food Retrieval Experiments

We have evaluated the classification performances of our network on the UNICT-FD1200 dataset, chosen because it was specifically designed for food retrieval. The UNICT-FD1200 dataset is composed by 4,754 images and 1,200 distinct dishes of food of different nationalities. We followed the evaluation procedures described in the original paper [13]. Specifically, the food dataset 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 dataset are considered. The overall retrieval performances are measured as the average on the three splits.

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 retrieved images [13]. For the retrieval

task, 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 is one correct image to be retrieved. We also report the Top-1 recognition accuracy.

Table 3 shows the retrieval results obtained on the UNICT-FD1200 dataset. We compare the performances of the features extracted with the fine-tuned network, "Activations ResNet-50 (Food524DB)" against those obtained with the original network, "Activation ResNet-50 (ImageNet)", and against the hand-crafted features used in [13]. As it can be seen the using the fine tuned network outperform all the other methods in the classification task as well as in the retrieval task. As expected the learned features greatly outperforms the hand-crafted ones. The fine tuning of the ResNet-50 improves the retrieval results of 3% for the Top-1 and of 2.4% for the mAP. Figure 2 shows the $P(n)$ curves of the methods in Table 3. It can be appreciated how the CNN-based features are able to effectively retrieve the relevant images in the first position.

Table 3. Classification and retrieval results on the UNICT-FD1200 dataset.

Representation	Top-1 (%)	mAP (%)
Bag of SIFT 12000 [13]	21.81	29.14
Textons (MR8) - RGB - Global [13]	71.55	77.00
Textons (Schmidt) - Lab - Global [13]	87.44	90.06
Activations ResNet-50 (ImageNet)	91.84	94.15
Activations ResNet-50 (Food524DB)	**94.96**	**96.56**

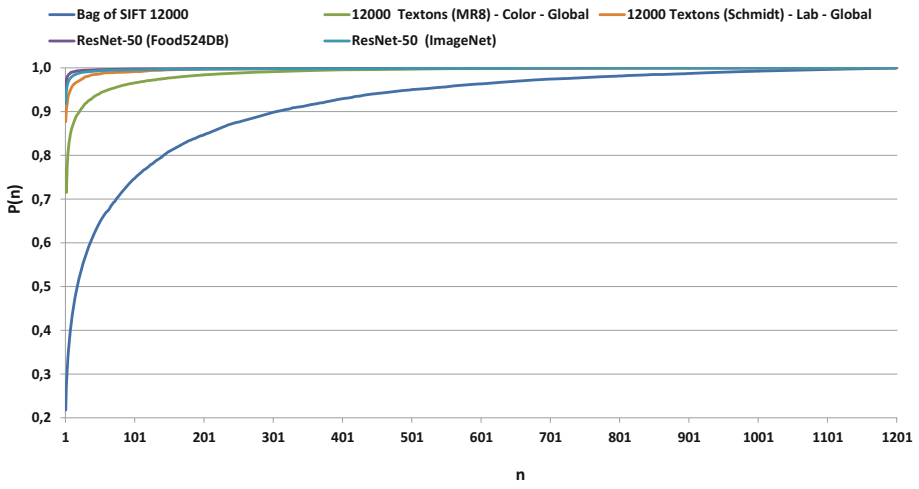

Fig. 2. P(n) curves of the methods in Table 3.

4 Conclusions

In this paper we investigated the use of CNN-based features for food retrieval. In order to accomplish this task we have created the Food524DB dataset by merging food classes from existing datasets in the state of the art. To date, Food524DB is the largest publicly available food dataset with 524 food classes and 247,636 images. The proposed CNN-based features have been obtained from a Residual Network (ResNet-50) fine tuned on Food524DB. The evaluation have been carried out on the UNICT-FD1200 dataset, that is a specific food retrieval dataset with 1,200 classes. Results demonstrated the powerful of the proposed CNN-based features with respect to CNN-based features extracted from the same network architecture trained on scene images and with respect to the state of the art features evaluated on the same dataset.

Acknowledgements. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

1. Akpro Hippocrate, E.A., Suwa, H., Arakawa, Y., Yasumoto, K.: Food weight estimation using smartphone and cutlery. In: Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems, IoT of Health 2016, pp. 9–14. ACM (2016)
2. Anthimopoulos, M.M., Gianola, L., Scarnato, L., Diem, P., Mougiakakou, S.G.: A food recognition system for diabetic patients based on an optimized bag-of-features model. IEEE J. Biomed. Health Inf. **18**(4), 1261–1271 (2014)
3. Bettadapura, V., Thomaz, E., Parnami, A., Abowd, G., Essa, I.: Leveraging context to support automated food recognition in restaurants. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 580–587 (2015)
4. Bianco, S., Ciocca, G., Napoletano, P., Schettini, R., Margherita, R., Marini, G., Pantaleo, G.: Cooking action recognition with iVAT: an interactive video annotation tool. In: Petrosino, A. (ed.) ICIAP 2013. LNCS, vol. 8157, pp. 631–641. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41184-7_64
5. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 446–461. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_29
6. Chen, J., Ngo, C.W.: Deep-based ingredient recognition for cooking recipe retrieval. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 32–41. ACM (2016)
7. Chen, M., Dhingra, K., Wu, W., Yang, L., Sukthankar, R., Yang, J.: PFID: pittsburgh fast-food image dataset. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 289–292. IEEE (2009)
8. Chen, M.Y., Yang, Y.H., Ho, C.J., Wang, S.H., Liu, S.M., Chang, E., Yeh, C.H., Ouhyoung, M.: Automatic chinese food identification and quantity estimation. In: SIGGRAPH Asia 2012 Technical Briefs, p. 29. ACM (2012)

9. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition and leftover estimation for daily diet monitoring. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 334–341. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_41
10. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments and results. IEEE J. Biomed. Health Inf. **21**(3), 588–598 (2017)
11. Cusano, C., Napoletano, P., Schettini, R.: Intensity and color descriptors for texture classification. In: IS&T/SPIE Electronic Imaging, p. 866113. International Society for Optics and Photonics (2013)
12. Cusano, C., Napoletano, P., Schettini, R.: Combining local binary patterns and local color contrast for texture classification under varying illumination. JOSA A **31**(7), 1453–1461 (2014)
13. Farinella, G.M., Allegra, D., Moltisanti, M., Stanco, F., Battiato, S.: Retrieval and classification of food images. Comput. Biol. Med. **77**, 23–39 (2016)
14. Farinella, G.M., Allegra, D., Stanco, F.: A benchmark dataset to study the representation of food images. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 584–599. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16199-0_41
15. Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., Cagnoni, S.: Food image recognition using very deep convolutional networks. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, MADiMa 2016, pp. 41–49. ACM (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
17. He, Y., Xu, C., Khanna, N., Boushey, C., Delp, E.: Analysis of food images: features and classification. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 2744–2748 (2014)
18. Hoashi, H., Joutou, T., Yanai, K.: Image recognition of 85 food categories by feature fusion. In: IEEE International Symposium on Multimedia (ISM) 2010, pp. 296–301. IEEE (2010)
19. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
20. Joutou, T., Yanai, K.: A food image recognition system with multiple kernel learning. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 285–288. IEEE (2009)
21. Kawano, Y., Yanai, K.: Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8927, pp. 3–17. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16199-0_1
22. Kawano, Y., Yanai, K.: Food image recognition with deep convolutional features. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2014 Adjunct, pp. 589–593 (2014)
23. Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V., Ma, Y.: DeepFood: deep learning-based food image recognition for computer-aided dietary assessment. In: Chang, C.K., Chiari, L., Cao, Y., Jin, H., Mokhtari, M., Aloulou, H. (eds.) ICOST 2016. LNCS, vol. 9677, pp. 37–48. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39601-9_4

24. Mariappan, A., Bosch, M., Zhu, F., Boushey, C.J., Kerr, D.A., Ebert, D.S., Delp, E.J.: Personal dietary assessment using mobile devices, vol. 7246, pp. 72460Z-1–72460Z-12 (2009)
25. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. arXiv preprint arXiv:1612.06543 (2016)
26. Matsuda, Y., Hoashi, H., Yanai, K.: Recognition of multiple-food images by detecting candidate regions. In: 2012 IEEE International Conference on Multimedia and Expo (ICME), pp. 25–30 (2012)
27. Nguyen, D.T., Zong, Z., Ogunbona, P.O., Probst, Y., Li, W.: Food image classification using local appearance and global structural information. Neurocomputing **140**, 242–251 (2014)
28. Pouladzadeh, P., Kuhad, P., Peddi, S.V.B., Yassine, A., Shirmohammadi, S.: Food calorie measurement using deep learning neural network. In: IEEE International Instrumentation and Measurement Technology Conference, pp. 1–6 (2016)
29. 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.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
30. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 806–813 (2014)
31. Wang, X., Kumar, D., Thome, N., Cord, M., Precioso, F.: Recipe recognition with large multimodal food dataset. In: 2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6. IEEE (2015)
32. Yanai, K., Kawano, Y.: 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 (2015)