Food vs Non-Food Classification

Francesco Ragusa[1], Valeria Tomaselli[2], Antonino Furnari[1],
Sebastiano Battiato[1], Giovanni Maria Farinella[1]
[1]University of Catania - Department of Mathematics and Computer Science
[2]STMicroelectronics - Advanced System Technology - Computer Vision
francescoragusa@outlook.com, valeria.tomaselli@st.com,
{furnari,battiato,gfarinella}@dmi.unict.it

ABSTRACT

Automatic understanding of food is an important research challenge. Food recognition engines can provide a valid aid for automatically monitoring the patient's diet and food-intake habits directly from images acquired using mobile or wearable cameras. One of the first challenges in the field is the discrimination between images containing food versus the others. Existing approaches for food vs non-food classification have used both shallow and deep representations, in combination with multi-class or one-class classification approaches. However, they have been generally evaluated using different methodologies and data, making a real comparison of the performances of existing methods unfeasible. In this paper, we consider the most recent classification approaches employed for food vs non-food classification, and compare them on a publicly available dataset. Different deep-learning based representations and classification methods are considered and evaluated.

CCS Concepts

•Computing methodologies → Image representations; Object recognition;

Keywords

Food vs Non-Food classification; Image Understanding

1. INTRODUCTION

Automatic food understanding is becoming more and more important to provide services for self-nutrition [1, 2]. User's feeding habits have to be taken into account to try to combat obesity, which is dramatically increasing also in childhood [3]. Having a daily record of actual intake of food has a crucial importance to provide the user with a personalized diet and beneficial information on the food intake balance. More specifically, the procedure of measuring energy and nutrients intake requires the record of all food consumed by an individual, the identification of the portion size and the determination of the frequency with which each food is eaten. Standard approaches for food intake monitoring demand the user to self-report all the food he/she eats and to recognize and describe also quantities. As pointed-out by different studies, self-reporting is often inaccurate [4, 5, 6, 7]. Many studies have shown that the collection of accurate dietary intake data is hampered by under-reporting of food intake [8]. Under-reporting in large nutritional surveys ranges from 18 to 54% of the whole sample, and it can widely vary due to different criteria used to identify under-reporters and also to non-uniformity of under-reporting across populations (e.g., men and women, children and adults). Moreover, it has been frequently observed that under-reporting is greater in overweight and obese individuals than those of healthy weight [9].

Thanks to the great diffusion of low cost image acquisition devices (e.g., smartphones and wearable cameras), food is nowadays one of the most photographed objects. This allows to create a food-log [10, 11] by simply taking snapshots of meals. It should be noted that snapshots could be acquired either actively, using a smartphone [12], or automatically, using an always-on wearable camera [13]. The first step to address in order to build a food-logging system is the discrimination between food and non-food images [14, 15, 16]. It should be noted that, in this paper, we tackle the problem of food vs non-food discrimination, which is different from food detection. Indeed, while in the former case, the algorithm takes an image as input and predicts whether the image contains food or not, in the latter one, the algorithm should be able to detect where food is located in the image, providing as output a bounding box or a pixel-wise mask [17]. Images classified as "containing food" could be further analyzed by automated systems in order to count calories [16] or simply stored and organized to create a food log [10] to be analysed by nutritionists in order to perform a better monitoring of the patient's diet and to understand food intake habits.

Despite the availability of different approaches for food vs non-food discrimination in the literature, they have been generally evaluated using different methodologies and on different datasets, making a direct comparison unfeasible. Considering their promising performances in many computer vision problems, including food analysis [18, 16, 19], in this paper we benchmark the main deep-learning-based representations on the task of food vs non-food image classification. We also consider different classification methods employed in the recent literature, including the softmax classification embedded in CNNs [16, 18, 19], standard binary SVM clas-

MADiMa'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2986035.2986041

sification [14], and one-class SVM classification [17]. Experiments are performed on public datasets and by considering the challenging evaluation scheme proposed in [17]. The best among the considered models achieves an overall accuracy of 94.86% with balanced True Positive Rate (ability to recognize food images) and True Negative Rate (ability to recognize non-food images) scores of 94.28% and 95.59% respectively. An online demo of the proposed system along with the models required for the discrimination are available at the URL *http://iplab.dmi.unict.it/demofood/*.

The remainder of the paper is organized as follows. Section 2 discusses the related works. Section 3 introduces the considered features and classification methods. Section 4 describes the experiments, whereas Section 5 reports the results. Finally, Section 6 concludes the paper.

2. RELATED WORKS

The literature related to food vs non-food classification have considered both shallow [14, 17] and deep [16, 18, 19] representations, and employed either multi-class [19] or one-class [17] classification techniques. In [14], Kitamura et al. addressed the problem of discriminating between food and non-food images by training an SVM classifier on features based on shapes (e.g., circles and rectangles), color histograms, and DCT coefficients. The system has been extended in [15] with the introduction of a Bayesian framework using the same color and shape-based features. Kagaya et al. [18] used deep learning for food detection and recognition, by employing both food and non-food images to train the discriminative model. Similar approaches have been employed in [16] and [19], where respectively the GoogLeNet [20] and Network in Network [21] models pre-trained on ImageNet were fine-tuned on two different food vs non-food datasets. Farinella et al. [17] exploited one-class classification methods [22] on multiple shallow image representations. Using a one-class classifier allows to learn from food images only, avoiding the inclusion of non-food images in the training procedure.

As mentioned before, previous approaches have been evaluated using different experimental settings and different data, so a direct comparison of such methods, to select the best approach to be employed in a real application, is quite impossible. In this paper we provide an extensive evaluation of the main classification approaches and CNN models employed in past works, to design a robust food vs non-food classifier. Following [17], we use public data and perform evaluations in challenging settings, in order to allow both reproducibility of the results and future comparisons.

3. FEATURES AND CLASSIFICATION

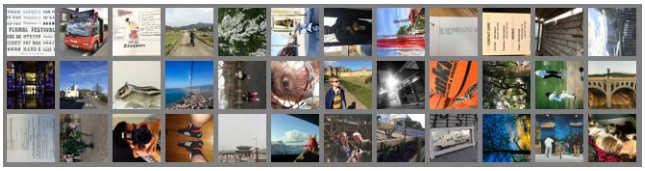
We address the task of discriminating between food and non-food images as a classification problem. In our analysis, we will also consider different representations and classification schemes. Let $\mathcal{T} = (\{I_1, I_2, \cdots, I_n\}, \{l_1, l_2, \cdots, l_n)\})$ be a given training set, where $\{I_1, I_2, \cdots, I_n\}$ are training images belonging to the set of all possible images $\mathcal{I}$ ($I_i \in \mathcal{I}, \forall i | 1 \leq i \leq n$) and $\{l_1, l_2, \cdots, l_n\}$ are training labels ($l_i = 1$ if I_i contains food, $l_i = 0$ otherwise). Let $\phi : \mathcal{I} \rightarrow \Re^d$ be a representation function from the set of all possible images to the d-dimensional representation space $\Re^d$ and let $\mathcal{S} = (\{x_1, x_2, \cdots, x_n\}, \{l_1, l_2, \cdots, l_n)\})$ be the set of labeled training samples, where $x_i = \phi(I_i), \forall i | 1 \leq i \leq n$. A given classifier $\mathcal{C}$ should be able to learn from the set of training samples $\mathcal{S}$ in order to assign the correct label $\overline{l}$ to a new (not previously observed) image $\overline{I}$. This scheme can be applied employing a binary or one-class SVM classifier with a selected kernel, or considering Convolutional Neural Networks. In practice, a binary SVM classifier would learn the model and optimize the involved hyperparameters from the training samples $\mathcal{S}$. When CNNs are used for both feature extraction and classification, the learning and representation steps are performed jointly and the model is learned directly from the training set of images $\mathcal{T}$. When one-class classifiers are employed, the training set $\mathcal{S}$ contains only positive samples with all labels equal to 1, while a set of negative samples can still be used to optimize the possible hyper-parameters of the model.

Considering their promising results in the task of food vs non-food discrimination [16, 18, 19], we consider deep representations in our experiments. In particular, we consider the following CNN architectures: the Network in Network model [21], the AlexNet model [23], and the VGG-S model [24]. All considered models have been pretrained on the ImageNet dataset. The AlexNet model is a standard CNN architecture which has already been used for food vs non-food discrimination [18]. The Network in Network model proposed in [21] is a memory efficient network which achieves state of the art performances on the ImageNet dataset. This model has been used for food vs non-food classification in [19]. The considered VGG model has shown state of the art performances on many classification tasks and usually outperforms the standard AlexNet architecture as discussed in [24]. Given the unavailability of large-scale food datasets, we do not train the considered models from scratch. Instead, we take advantage of models pretrained on the ImageNet dataset and consider transfer learning techniques as done by other authors [16]. Specifically, we consider two basic transfer learning techniques: 1) using the pre-trained models as simple feature extractors and 2) fine-tuning the models. The former method consists in propagating the input image into the network and extracting as features the activation values contained in the penultimate layer of the network. For the AlexNet [23] and VGG models [24], such layer is referred to as $fc7$ and its dimensionality is equal to 4096. In the case of the Network in Network model [21], we extract activations from the penultimate layer ($cppp7$). Extracted values are then passed through an average pooling layer to obtain vectors of 1024 features (see [21] for details). Once features are extracted from the training set, an SVM classifier is trained on top of them. In the case of fine-tuning, the last layer of the network containing 1000 units is substituted with a new layer containing only two nodes. The weights of the new units are initialized with gaussian noise ($\sigma = 0.01$) and the biases are initialized to zero. The learning rate of the new units is set to 10 times the base learning rate (i.e., the learning rate for all other units) and the standard training procedure based on gradient-descent is resumed starting from the weights pretrained on ImageNet. We would like to note that we consider also combining both transfer learning methods, i.e., combining fine-tuned models with an SVM classifier.

We also consider three different classification schemes which have been employed in the literature: 1) SoftMax classification: when networks are fine-tuned, the output probability is used for classification; 2) One Class SVM: following [17], we

(a) UNICT-FD889

(b) Flickr-Food



(c) Flickr-NonFood

Figure 1: Some examples from the three considered datasets.

train a one-class classifier considering only positive samples; 3) Binary SVM: we train a standard binary SVM classifier using both positive (food) and negative (non-food) samples. While the first method is restricted to the case in which CNNs are fine-tuned, the second and third methods are applied when CNNs are used just for feature extraction, as well as when models are finetuned.

4. EXPERIMENTAL SETTINGS

Experiments follow the evaluation pipeline suggested in [17]. In order to assess the generalization ability of the methods, three different datasets are used for the training and testing procedures. The first dataset is named UNICT-FD889 and has been proposed in [25, 1][1]. It contains 3583 food images related to 889 different plates acquired multiple times to introduce variability in real-world scenarios during meals (e.g., by using an iPhone). It should be noted that, differently from other datasets (e.g., Food-101 [26]), each image in UNICT-FD889 is a close-up of the acquired meal, so that images contain mainly food whereas the presence of other objects is limited. The second dataset considered in the experiments has been proposed in [17] and consists of 4805 food images. All images have been downloaded from Flickr and manually inspected to check the actual presence of food. The dataset contains only images acquired with an iPhone at their original resolution. Differently from UNICT-FD889, the acquisition settings for this dataset are less constrained and images can occasionally contain also other objects not related to food. This dataset is referred to as Flickr-Food. The third considered dataset has been acquired from Flickr by the same authors [17] with similar modalities. It contains 8005 non-food images belonging to different scene types and depicting different objects. This dataset is referred to as Flickr-NonFood. Figure 1 shows some examples from the considered datasets.

[1]The UNICT-FD889 dataset is available at the URL *http://iplab.dmi.unict.it/madima2015/*

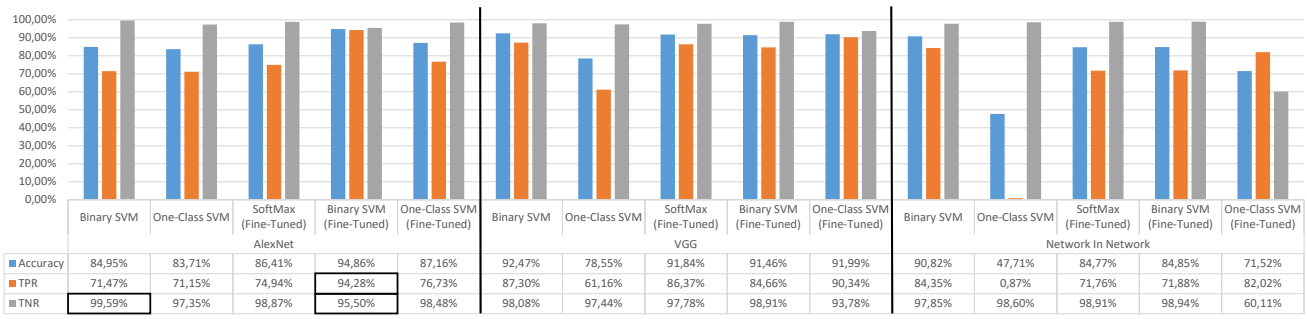
In our experiments, the Flickr-NonFood dataset is randomly divided into two asymmetrical halves containing 3583 and 4422 samples respectively. This asymmetrical split is considered in order to obtain a balanced training set (i.e., with an equal number of positive and negative samples). Training is hence performed using the 3583 food images contained in UNICT-F889 and the 3583 non-food images contained in the first half of the Flickr-NonFood dataset. Evaluations are performed on the 4805 food images contained on the Flickr-Food dataset and the 4422 non-food images contained in the second half of the Flickr-NonFood dataset. As pointed out in [17], such challenging training/testing scheme allows to assess the generalization capability of the considered methods since positive samples in training and test sets have been acquired by different sources and with different modalities, while negative samples are heterogeneous per se.

All experiments have been performed using the Caffe framework [27] and LibSVM library [28]. The implementation of [29] is considered for one-class SVM. The BVLC reference CaffeNet model has been used as AlexNet implementation. During training, standard data augmentation techniques are employed resizing the input images to 256×256 pixels and cropping/mirroring them according to what suggested in [23]. When CNNs are fine-tuned, one third of the training set is used for validation, while the remaining two thirds are used for training. Fine-tuning was terminated when the validation accuracy stopped increasing. After the fine-tuning is complete, the model corresponding to the epoch with the highest validation accuracy is considered for reference. We consider a sigmoid kernel for training both the one-class and binary SVM classifiers. The best parameters (namely, γ value for the sigmoid kernel, ν value for the one-class SVM, and C cost for the binary SVM) have been selected using a grid search with 3-fold validation. It should be noted that in the experiments related to one-class classification just positive samples are used to learn the model. Positive and negative samples are used jointly in order to chose the optimal set of hyperparameters.

5. RESULTS

Figure 2 summarizes the experimental results. For each classifier and representation combination, we report the following performance scores: the accuracy of the system (i.e., the fraction of correct predictions), the True Positive Rate (TPR) (i.e., the fraction of correctly classified food images), and the True Negative Rate (TNR) (i.e., the fraction of correctly classified non-food images). Best performances are reported as boxed numbers in Figure 2.

Pre-trained CNNs allow to obtain already good results when they are simply used for feature extraction and coupled to a binary SVM classifier (first column for each model in Figure 2). In particular, advanced models such as VGG and Network In a Network yield very good results ($> 90\%$). However, it should be noted that these approaches always achieve unbalanced results, with TNR values (i.e., non-food class) higher than TPR values (i.e., food class). This is probably due to the fact that all networks have been pre-trained on a dataset (i.e., ImageNet) mainly containing objects and few food samples. Keeping using CNNs just for feature extraction and substituting the binary SVM classifiers with one-class classifiers as suggested in [17] (second column for each model in Figure 2), leads in general to less balanced and worse results. In particular, the features extracted with



	AlexNet					VGG					Network In Network				
	Binary SVM	One-Class SVM	SoftMax (Fine-Tuned)	Binary SVM (Fine-Tuned)	One-Class SVM (Fine-Tuned)	Binary SVM	One-Class SVM	SoftMax (Fine-Tuned)	Binary SVM (Fine-Tuned)	One-Class SVM (Fine-Tuned)	Binary SVM	One-Class SVM	SoftMax (Fine-Tuned)	Binary SVM (Fine-Tuned)	One-Class SVM (Fine-Tuned)
Accuracy	84,95%	83,71%	86,41%	94,86%	87,16%	92,47%	78,55%	91,84%	91,46%	91,99%	90,82%	47,71%	84,77%	84,85%	71,52%
TPR	71,47%	71,15%	74,94%	94,28%	76,73%	87,30%	61,16%	86,37%	84,66%	90,34%	84,35%	0,87%	71,76%	71,88%	82,02%
TNR	99,59%	97,35%	98,87%	95,50%	98,48%	98,08%	97,44%	97,78%	98,91%	93,78%	97,85%	98,60%	98,91%	98,94%	60,11%

Figure 2: Results of the experiments. Values indicating best performances are boxed. Left: results related to the AlexNet model. Middle: results related to the VGG model. Right: results related to the Network In Network model.

the Network in Network model are unable to correctly classify food images, with a TPR of only 0.87%. Fine-tuning the networks and using the output of the softmax layer directly for classification (third column for each model in Figure 2) brings some minor improvements over the binary SVM in the case of AlexNet and generally worse results with the other models. Interestingly, training a binary SVM classifier on top of the representations extracted with the fine-tuned models (fourth column for each model in Figure 2) leads to big improvements in the case of the AlexNet model (7% to 10% over other classification approaches), and comparable results when other models are employed. Substituting the binary SVM classifier with a one-class classifier on fine-tuned models (fifth column for each model in Figure 2) does not leads to consistent improvements. Figure 3 shows some success/failure examples related to the best performing model, i.e., fine-tuned AlexNet + binary SVM classifier.

In our experiments, the best performing method is achieved coupling a fine-tuned AlexNet model with a binary SVM classifier. The overall accuracy is 94.86%, and performances are balanced with respect to both food and non-food classes, with a True Positive Rate (i.e., food class) of 94.28% and a True Negative Rate (i.e., non-food class) of 95.50%. It should be noted that this combination outperforms the softmax classifier included in the fine-tuned network by a big margin (about 11%). This discrepancy, is probably due to the fact that the training process of CNNs is probabilistic and optimizes representation and classifier jointly, whereas training a binary SVM model is a deterministic process which optimizes just the classifier component. Interestingly, while in general the VGG model outperforms the AlexNet architecture on different image classification tasks [24], in our experiments, the AlexNet model outperforms VGG when fine-tuning is coupled with a binary SVM classifier. We would like to note that the AlexNet model is much lighter and fast than VGG, which involves savings in space and computational time. The AlexNet model seems to be the only one benefiting from fine-tuning. This is probably due to the higher capacity of VGG, which makes fine-tuning with few samples prone to overfitting, as well as to the known difficulties in fine-tuning the Network in Network model due to the absence of fully connected layer as discussed in [20, 30]. This leaves us with the intuition that the less specialized features of the AlexNet model are more prone to adapt to a quite different task as food vs non-food discrimination. An online demo based on the best results obtained in this paper (i.e., fine-tuned AlexNet + binary SVM classifier), is available at the URL *http://iplab.dmi.unict.it/demofood/*.

(a) Food as Food (b) Non-Food as Food

(c) Food as Non-Food (d) Non-Food as Non-Food

Figure 3: Some success/failure cases related to the best performing model (fine-tuned AlexNet + binary SVM). (a) food images correctly classified. (b) non-food images erroneously classified as food. (c) food images erroneously classified as non-food. (d) non-food images correctly classified.

6. CONCLUSION

Food vs non-food classification have been exploited using both shallow and deep representation and employing different classification approaches. Since such methods have usually been evaluated on non-public data, with different methodologies, in this paper we have benchmarked the main deep representation and classification techniques in order to design a robust and efficient food vs non-food classifier. Results show that the combination of a fine-tuned AlexNet model and a binary SVM classifier gives the best results. Future works will investigate how the proposed method can be optimized to reduce the required memory and computational resources. Moreover, exploitation of this model as a first step towards a food-log application from wearable cameras will be considered.

7. REFERENCES

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

[2] G. M. Farinella, M. Moltisanti, and S. Battiato. Classifying food images represented as bag of textons. In *IEEE International Conference on Image Processing*, pages 5212–5216, 2014.

[3] C.E. Collins, J. Watson, and T. Burrows. Measuring dietary intake in children and adolescents in the context of overweight and obesity. *International journal of obesity*, 34(7):1103–1115, 2010.

[4] F. Kong and J. Tan. Dietcam: Automatic dietary assessment with mobile camera phones. *Pervasive and Mobile Computing*, 8(1):147–163, 2012.

[5] L. Arab, D. Estrin, D. H. Kim, J. Burke, and J. Goldman. Feasibility testing of an automated image-capture method to aid dietary recall. *European Journal of Clinical Nutrition*, 65(10):1156–1162, 2011.

[6] F. Zhu, M. Bosch, I. Woo, S. Y. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp. The use of mobile devices in aiding dietary assessment and evaluation. *Selected Topics in Signal Processing*, 4:756–766, 2010.

[7] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney. Recognition and volume estimation of food intake using a mobile device. In *Applications of Computer Vision*, pages 1–8, 2009.

[8] S. Kye, S.-O. Kwon, S.-Y. Lee, J. Lee, B.-H. Kim, H.-J. Suh, and H.-K. Moon. Under-reporting of energy intake from 24-hour dietary recalls in the korean national health and nutrition examination survey. *Osong Public Health and Research Perspectives*, 5:85–91, 2014.

[9] K. L. Rennie, A. Coward, and S. A. Jebb. Estimating under-reporting of energy intake in dietary surveys using an individualised method. *British Journal of Nutrition*, 97:1169–1176, 2007.

[10] K. Aizawa. Multimedia foodlog: Diverse applications from self-monitoring to social contributions. *ITE Transactions on Media Technology and Applications*, 1:214–219, 2013.

[11] E. S. Sazonov and J. M. Fontana. A sensor system for automatic detection of food intake through non-invasive monitoring of chewing. *Sensors Journal*, 12:1340–1348, 2012.

[12] D. Ravi, B. Lo, and G.-Z. Yang. Real-time food intake classification and energy expenditure estimation on a mobile device. In *Wearable and Implantable Body Sensor Networks*, pages 1–6, 2015.

[13] E. Thomaz, A. Parnami, I. Essa, and G. D. Abowd. Feasibility of identifying eating moments from first-person images leveraging human computation. In *SenseCam & Pervasive Imaging Conference*, pages 26–33, 2013.

[14] K. Kitamura, T. Yamasaki, and K. Aizawa. Foodlog: Capture, analysis and retrieval of personal food images via web. In *Workshop on Multimedia for Cooking and Eating Activities*, pages 23–30, 2009.

[15] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa. Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Transactions on Multimedia*, 15:2176–2185, 2013.

[16] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy. Im2calories: Towards an automated mobile vision food diary. In *International Conference on Computer Vision*, December 2015.

[17] G. M. Farinella, D. Allegra, F. Stanco, and S. Battiato. On the exploitation of one class classification to distinguish food vs non–food images. In *Multimedia Assisted Dietary Management*, volume 9281 of *Lecture Notes in Computer Science*, 2015.

[18] H. Kagaya, K. Aizawa, and M. Ogawa. Food detection and recognition using convolutional neural network. In *ACM International Conference on Multimedia*, pages 1085–1088, 2014.

[19] H. Kagaya and K. Aizawa. Highly accurate food/non-food image classification based on a deep convolutional neural network. In *New Trends in Image Analysis and Processing*, pages 350–357, 2015.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[21] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*, 2013.

[22] S.S. Khan and M.G. Madden. A survey of recent trends in one class classification. In *Irish Conference on Artificial Intelligence and Cognitive Science*, pages 188–197, 2010.

[23] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[24] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, volume 2, page 8, 2011.

[25] G. M. Farinella, D. Allegra, and F. Stanco. A benchmark dataset to study the representation of food images. In *International Workshop on Assistive Computer Vision and Robotics in conjunction with ECCV*, 2014.

[26] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. 2014.

[27] Berkeley Vision and Learning Center (BVLC). Caffe. http://caffe.berkeleyvision.org/.

[28] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines, 2001.

[29] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13:1443–1471, 2001.

[30] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015.