Food Ingredients Recognition Through Multi-label Learning

Marc Bolaños[1,2(✉)], Aina Ferrà[1], and Petia Radeva[1,2]

[1] Universitat de Barcelona, Barcelona, Spain
{marc.bolanos,aferrama10.alumnes,petia.ivanova}@ub.edu
[2] Computer Vision Center, Bellaterra, Spain

Abstract. Automatically constructing a food diary that tracks the ingredients consumed can help people follow a healthy diet. We tackle the problem of food ingredients recognition as a multi-label learning problem. We propose a method for adapting a highly performing state of the art CNN in order to act as a multi-label predictor for learning recipes in terms of their list of ingredients. We prove that our model is able to, given a picture, predict its list of ingredients, even if the recipe corresponding to the picture has never been seen by the model. We make public two new datasets suitable for this purpose. Furthermore, we prove that a model trained with a high variability of recipes and ingredients is able to generalize better on new data, and visualize how it specializes each of its neurons to different ingredients.

1 Introduction

People's awareness about their nutrition habits is increasing either because they suffer from some kind of food intolerance; they have mild or severe weight problems; or they are simply interested in keeping a healthy diet. This increasing awareness is also being reflected in the technological world. Several applications exist for manually keeping track of what we eat, but they rarely offer any automatic mechanism for easing the tracking of the nutrition habits [2]. Tools for automatic food and ingredient recognition could heavily alleviate the problem.

Since the reborn of Convolutional Neural Networks (CNNs), several works have been proposed to ease the creation of nutrition diaries. The most widely spread approach is food recognition [8]. These proposals allow to recognize the type of food present in an image and, consequently, could allow to approximately guess the ingredients contained and the overall nutritional composition. The main problem of these approaches is that no dataset covers the high amount of existent types of dishes worldwide (more than 8,000 according to Wikipedia).

On the other hand, a clear solution for this problem can be achieved if we formulate the task as an ingredients recognition problem instead [6]. Although tens of thousands of types of dishes exist, in fact they are composed of a much smaller number of ingredients, which at the same time define the nutritional composition of the food. If we formulate the problem from the ingredients recognition perspective, we must consider the difficulty of distinguishing the presence

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 394–402, 2017.
https://doi.org/10.1007/978-3-319-70742-6_37

of certain ingredients in cooked dishes. Their visual appearance can greatly vary from one dish to another (e.g. the appearance of the ingredient 'apple' in an 'apple pie', an 'apple juice' or a 'fresh apple'), and in some cases they can even be invisible at sight without the proper knowledge of the true composition of the dish. An additional benefit of approaching the problem from the ingredients recognition perspective is that, unlike in food recognition, it has the potential to predict valid outputs on data that has never been seen by the system.

In this paper, we explore the problem of food ingredients recognition from a multi-label perspective by proposing a model based on CNNs that allows to discover the ingredients present in an image even if they are not visible to the naked eye. We present two new datasets for tackling the problem and prove that our method is capable of generalizing to new data that has never been seen by the system. Our contributions are four-fold. (1) Propose a model for food ingredients recognition; (2) Prove that by using a varied dataset of images and their associated ingredients, the generalization capabilities of the model on never seen data can be greatly boosted; (3) Delve into the inner layers of the model for analysing the ingredients specialization of the neurons; and (4) Release two datasets for ingredients recognition.

This paper is organized as follows: in Sect. 2, we review the state of the art; in Sect. 3, explain our methodology; in Sect. 4, we present our proposed datasets, show and analyse the results of the experiments performed, as well as interpret the predictions; and in Sect. 5, we draw some conclusions.

2 Related Work

Food analysis. Several works have been published on applications related to automatic food analysis. Some of them proposed food detection models [1] in order to distinguish when there is food present in a given image. Others focused on developing food recognition algorithms, either using conventional hand-crafted features, or powerful deep learning models [8]. Others have applied food segmentation [11]; use multi-modal data (i.e. images and recipe texts) for recipe recognition [15]; tags from social networks for food characteristics perception [9]; food localization and recognition in the wild for egocentric vision analysis [3], etc.

Multi-Label learning. Multi-label learning [13] consists in predicting more than one output category for each input sample. Thus, the problem of food ingredients recognition can be treated as a multi-label learning problem. Several works [14] argued that, when working with CNNs, they have to be reformulated for dealing with multi-label learning problems. Some multi-label learning works have already been proposed for restaurant classification. So far, only one paper [6] has been proposed related to ingredients recognition. Their dataset, composed of 172 food types, was manually labelled considering visible ingredients only, which limits it to find 3 ingredients on average. Furthermore, they propose a double-output model for simultaneous food type recognition and multi-label ingredients recognition. Although, the use of the food type for optimizing the model limits

its capability of generalization only to seen recipes and food types. This fact becomes an important handicap in a real-world scenario when dealing with new recipes. As we demonstrate in Sects. 4.3 and 4.4, unlike [6], our model is able to: (1) recognize the ingredients appearing in unseen recipes (see Fig. 1b); (2) learn abstract representations of the ingredients directly from food appearance (see Fig. 2); and (3) infer invisible ingredients.

Interpreting learning through visualization. Applying visualization techniques is an important aspect in order to interpret what has been learned by our model. The authors in [17] have focused on proposing new ways of performing this visualization. At the same time, they have proven that CNNs have the ability to learn high level representations of the data and even hidden inter-related information, which can help us when dealing with ingredients that are apparently invisible in the image.

3 Methodology

Deep multi-ingredients recognition. Most of the top performing CNN architectures have been originally proposed and intended for the problem of object recognition. At the same time, they have been proven to be directly applicable to other related classification tasks and have served as powerful pre-trained models for achieving state of the art results. In our case, we compared either using the InceptionV3 [12] or the ResNet50 [7] as the basic architectures for our model. We pre-trained it on the data from the ILSVRC challenge [10] and modified the last layer for applying a multi-label classification over the N possible output ingredients. When dealing with classification problems, CNNs typically use the softmax activation in the last layer. The softmax function allows to obtain a probability distribution for the input sample x over all possible outputs and thus, predicts the most probable outcome, $\hat{y}_x = \arg\max_{y_i} P(y_i|x)$.

The softmax activation is usually combined with the categorical cross-entropy loss function L_c during model optimization, which penalizes the model when the optimal output value is far away from 1:

$$L_c = -\sum_x \log(P(\hat{y}_x|x)). \tag{1}$$

In our model, we are dealing with ingredients recognition in a multi-label framework. Therefore, the model must predict for each sample x a set of outputs represented as a binary vector $\hat{Y}_x = \{\hat{y}_x^1, ..., \hat{y}_x^N\}$, where N is the number of output labels and each $\hat{y}_x^i$ is either 1 or 0 depending if it is present or not in sample x. For this reason, instead of softmax, we use a sigmoid activation function:

$$P(y_i|x) = \frac{1}{1 - \exp^{-f(x)_i}} \tag{2}$$

which allows to have multiple highly activated outputs. For considering the binary representation of $\hat{Y}_x$, we chose the binary cross-entropy function L_b [5]:

$$L_b = -\sum_{x}\sum_{i}^{N}(\hat{y}_x^i \cdot log(P(y_i|x)) + (1 - \hat{y}_x^i) \cdot log(1 - P(y_i|x))) \tag{3}$$

which during backpropagation rewards the model when the output values are close to the target vector $\hat{Y}_x$ (i.e. either close to 1 for positive labels or close to 0 for negative labels).

4 Results

In this section, we describe the two datasets proposed for the problem of food ingredients recognition. Later we describe our experimental setup and at the end, we present the final results obtained both for ingredients recognition on known classes as well as recognition results for generalization on samples never seen by the model.

4.1 Datasets

In this section we describe the datasets proposed for food ingredients recognition and the already public datasets used.

Food101 [4] is one of the most widely extended datasets for food recognition. It consists of 101,000 images equally divided in 101 food types.

Ingredients101[1] is a dataset for ingredients recognition that we constructed and make public in this article. It consists of the list of most common ingredients for each of the 101 types of food contained in the Food101 dataset, making a total of 446 unique ingredients (9 per recipe on average). The dataset was divided in training, validation and test splits making sure that the 101 food types were balanced. We make public the lists of ingredients together with the train/val/test split applied to the images from the Food101 dataset.

Recipes5k[2] is a dataset for ingredients recognition with 4,826 unique recipes composed of an image and the corresponding list of ingredients. It contains a total of 3,213 unique ingredients (10 per recipe on average). Each recipe is an alternative way to prepare one of the 101 food types in Food101. Hence, it captures at the same time the intra-class variability and inter-class similarity of cooking recipes. The nearly 50 alternative recipes belonging to each of the 101 classes were divided in train, val and test splits in a balanced way. We make also public this dataset together with the splits division. A problem when dealing with the 3,213 raw ingredients is that many of them are sub-classes (e.g. ‘sliced tomato’ or ‘tomato sauce’) of more general versions of themselves (e.g. ‘tomato’).

[1] http://www.ub.edu/cvub/ingredients101/.

[2] http://www.ub.edu/cvub/recipes5k/.

Thus, we propose a simplified version by applying a simple removal of overly-descriptive particles[3] (e.g. 'sliced' or 'sauce'), resulting in 1,013 ingredients used for additional evaluation (see Sect. 4.3).

We must note the difference between our proposed datasets and the one from [6]. While we consider any present ingredient in a recipe either visible or not, the work in [6] only labelled manually the visible ingredients in certain foods. Hence, a comparison between both works is infeasible.

4.2 Experimental Setup

Our model was implemented in Keras[4], using Theano as backend. Next, we detail the different configurations and tests performed. **Random prediction**: (baseline) a set of K labels are generated uniformly distributed among all possible outputs. K depends on the average number of labels per recipe in the corresponding dataset. **InceptionV3 + Ingredients101**: InceptionV3 model pre-trained on ImageNet and adapted for multi-label learning. **ResNet50 + Ingredients101**: ResNet50 model pre-trained on ImageNet and adapted for multi-label learning. **InceptionV3 + Recipes5k**: InceptionV3 model pre-trained on InceptionV3 + Ingredients101. **ResNet50 + Recipes5k**: ResNet50 model pre-trained on ResNet50 + Ingredients101.

4.3 Experimental Results

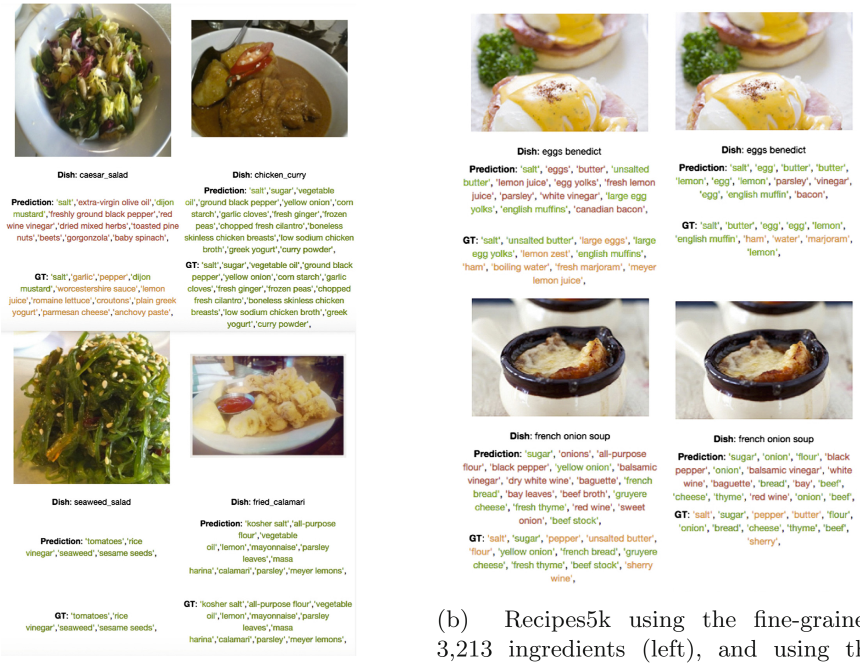
In Table 1, we show the ingredient recognition results on the Ingredients101 dataset. In Fig. 1a some qualitative results are shown. Both the numerical results and the qualitative examples prove the high performance of the models in most of the cases. Note that although a multi-label classification is being applied, considering that all the samples from a food class share the same set of ingredients, the model is indirectly learning the inherent food classes. Furthermore, looking at the results on the Recipes5k dataset in Table 2 (top), we can see that the very same model obtains reasonable results even considering that it was

Table 1. Ingredients recognition results obtained on the dataset Ingredients101. Prec stands for *Precision*, Rec for *Recall* and F_1 for F_1 *score*. All measures reported in %. The best test results are highlighted in boldface.

	Validation			Test		
	Prec	Rec	F_1	Prec	Rec	F_1
Random prediction	2.05	2.01	2.03	2.06	2.01	2.04
InceptionV3 + Ingredients101	80.86	72.12	76.24	83.51	**76.87**	80.06
ResNet50 + Ingredients101	84.80	67.62	75.24	**88.11**	73.45	**80.11**

[3] https://github.com/altosaar/food2vec.

[4] www.keras.io.


Dish: caesar_salad
Prediction: 'salt','extra-virgin olive oil','dijon mustard','freshly ground black pepper','red wine vinegar','dried mixed herbs','toasted pine nuts','beets','gorgonzola','baby spinach',
GT: 'salt','garlic','pepper','dijon mustard','worcestershire sauce','lemon juice','romaine lettuce','croutons','plain greek yogurt','parmesan cheese','anchovy paste',
Dish: chicken_curry
Prediction: 'salt','sugar','vegetable oil','ground black pepper','yellow onion','corn starch','garlic cloves','fresh ginger','frozen peas','chopped fresh cilantro','boneless skinless chicken breasts','low sodium chicken broth','greek yogurt','curry powder',
GT: 'salt','sugar','vegetable oil','ground black pepper','yellow onion','corn starch','garlic cloves','fresh ginger','frozen peas','chopped fresh cilantro','boneless skinless chicken breasts','low sodium chicken broth','greek yogurt','curry powder',
Dish: seaweed_salad
Prediction: 'tomatoes','rice vinegar','seaweed','sesame seeds',
GT: 'tomatoes','rice vinegar','seaweed','sesame seeds',
Dish: fried_calamari
Prediction: 'kosher salt','all-purpose flour','vegetable oil','lemon','mayonnaise','parsley leaves','masa harina','calamari','parsley','meyer lemons',
GT: 'kosher salt','all-purpose flour','vegetable oil','lemon','mayonnaise','parsley leaves','masa harina','calamari','parsley','meyer lemons',
Dish: eggs benedict
Prediction: 'salt', 'eggs', 'butter', 'unsalted butter', 'lemon juice', 'egg yolks', 'fresh lemon juice', 'parsley', 'white vinegar', 'large egg yolks', 'english muffins', 'canadian bacon',
GT: 'salt', 'unsalted butter', 'large eggs', 'large egg yolks', 'lemon zest', 'english muffins', 'ham', 'boiling water', 'fresh marjoram', 'meyer lemon juice',
Dish: eggs benedict
Prediction: 'salt', 'egg', 'butter', 'butter', 'lemon', 'egg', 'lemon', 'parsley', 'vinegar', 'egg', 'english muffin', 'bacon',
GT: 'salt', 'butter', 'egg', 'egg', 'lemon', 'english muffin', 'ham', 'water', 'marjoram', 'lemon',
Dish: french onion soup
Prediction: 'sugar', 'onions', 'all-purpose flour', 'black pepper', 'yellow onion', 'balsamic vinegar', 'dry white wine', 'baguette', 'french bread', 'bay leaves', 'beef broth', 'gruyere cheese', 'fresh thyme', 'red wine', 'sweet onion', 'beef stock',
GT: 'salt', 'sugar', 'pepper', 'unsalted butter', 'flour', 'yellow onion', 'french bread', 'gruyere cheese', 'fresh thyme', 'beef stock', 'sherry wine',
Dish: french onion soup
Prediction: 'sugar', 'onion', 'flour', 'black pepper', 'onion', 'balsamic vinegar', 'white wine', 'baguette', 'bread', 'bay', 'beef', 'cheese', 'thyme', 'red wine', 'onion', 'beef',
GT: 'salt', 'sugar', 'pepper', 'butter', 'flour', 'onion', 'bread', 'cheese', 'thyme', 'beef', 'sherry',

(a) Ingredients101 samples.

(b) Recipes5k using the fine-grained 3,213 ingredients (left), and using the 1,013 simplified ingredients (right).

Fig. 1. Our method's results. TPs in green, FPs in red and FNs in orange. (Color figure online)

Table 2. Ingredients recognition results on Recipes5k (top) and on Recipes5k simplified (bottom). Prec stands for *Precision*, Rec for *Recall* and F_1 for F_1 *score*. All measures reported in %. Best test results are highlighted in boldface.

	Validation			Test		
	Prec	Rec	F_1	Prec	Rec	F_1
Random prediction	0.33	0.32	0.33	0.54	0.53	0.53
InceptionV3 + Ingredients101				23.80	18.24	20.66
ResNet50 + Ingredients101				26.28	16.85	20.54
InceptionV3 + Recipes5k	36.18	20.69	26.32	35.47	**21.00**	**26.38**
ResNet50 + Recipes5k	38.41	19.67	26.02	**38.93**	19.57	26.05
Random prediction	6.27	6.29	6.28	6.14	6.24	6.19
InceptionV3 + Ingredients101				44.01	34.04	38.39
ResNet50 + Ingredients101				47.53	30.91	37.46
InceptionV3 + Recipes5k	56.77	31.40	40.44	55.37	31.52	40.18
ResNet50 + Recipes5k	56.73	28.07	37.56	**58.55**	28.49	38.33
InceptionV3 + Recipes5k simplified	53.91	42.13	47.30	53.43	**42.77**	**47.51**

not specifically trained on that dataset. Note that only test results are reported for the models trained on Ingredients101 because we only intend to show its generalization capabilities on new data.

Comparing the results with the models specifically trained on Recipes5k, it appears that, as expected, a model trained on a set of samples with high variability of output labels is more capable of obtaining high results on never seen recipes. Thus, it is more capable of generalizing on unseen data.

Table 2 (bottom) shows the results on the Recipes5k dataset with a simplified list of ingredients. Note that for all tests, the list was simplified only during the evaluation procedure for maintaining the fine-grained recognition capabilities of the model, with the exception of *Inception V3 + Recipes5k simplified*, where the simplified set was also used for training. The simplification of the ingredients list enhances the capabilities of the model when comparing the results, reaching more than 40% in the F_1 metric and 47.5% also training with them.

Figure 1b shows a comparison of the output of the model either using the fine-grained or the simplified list of ingredients. Overall, although usually only a single type of semantically related fine-grained ingredients (e.g. 'large eggs', 'beaten eggs' or 'eggs') appears at the same time in the ground truth, it seems that the model is inherently learning an embedding of the ingredients. Therefore, it is able to understand that some fine-grained ingredients are related and predicts them at once in the fine-grained version (see waffles example).

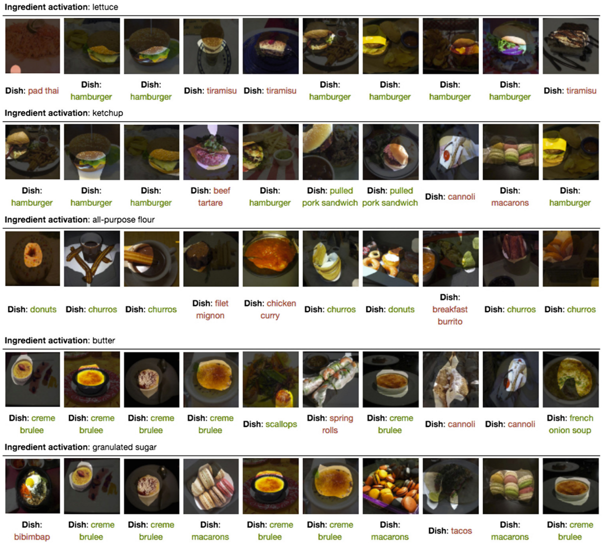


Fig. 2. Visualization of neuron activations. Each row is associated to a specific neuron from the network. The images with top activation are shown as well as the top ingredient activation they have in common. The name of their respective food class is only for visualization purposes and is displayed in green if the recipe contains the top ingredient. Otherwise, it is shown in red. (Color figure online)

4.4 Neuron Representation of Ingredients

When training a CNN model, it is important to understand what it is able to learn and interpret from the data. To this purpose, we visualized the activations of certain neurons of the network in order to interpret what is it able to learn.

Figure 2 shows the results of this visualization. As we can see, it appears that certain neurons of the network are specialized to distinguish specific ingredients. For example, most images of the 1st and 2nd rows illustrate that the characteristic shape of a hamburger implies that it will probably contain the ingredients 'lettuce' and 'ketchup'. Also, looking at the 'granulated sugar' row, we can see that the model learns to interpret the characteristic shape of *creme brulee* and *macarons* as containing sugar, although it is not specifically seen in the image.

5 Conclusions and Future Work

Analysing both the quantitative and qualitative results, we can conclude that the proposed model and the two datasets published offer very promising results for the multi-label problem of food ingredients recognition. Our proposal allows to obtain great generalization results on unseen recipes and sets the basis for applying further, more detailed food analysis methods. As future work, we will create a hierarchical structure [16] relationship of the existent ingredients and extend the model to utilize this information.

References

1. Aguilar, E., Bolaños, M., Radeva, P.: Exploring food detection using CNNs. In: Proceedings of the 16th International Conference on Computer Aided Systems Theory, pp. 242–243. Springer (2017)
2. Aizawa, K., Ogawa, M.: FoodLog: multimedia tool for healthcare applications. IEEE MultiMedia **22**(2), 4–8 (2015)
3. Bolaños, M., Radeva, P.: Simultaneous food localization and recognition. In: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR) (2016)
4. 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
5. Buja, A., Stuetzle, W., Shen, Y.: Loss functions for binary class probability estimation and classification: structure and applications. Working draft, November 2005
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. 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)

8. Martinel, N., Foresti, G.L., Micheloni, C.: Wide-slice residual networks for food recognition. arXiv preprint arXiv:1612.06543 (2016)
9. Ofli, F., Aytar, Y., Weber, I., al Hammouri, R., Torralba, A.: Is saki# delicious? the food perception gap on instagram and its relation to health. arXiv preprint arXiv:1702.06318 (2017)
10. Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
11. Shimoda, W., Yanai, K.: CNN-based food image segmentation without pixel-wise annotation. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 449–457. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_55
12. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
13. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. Int. J. Data Warehouse. Min. **3**(3), 1–13 (2006)
14. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: a unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2016)
15. 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)
16. Wu, H., Merler, M., Uceda-Sosa, R., Smith, J.R.: Learning to make better mistakes: semantics-aware visual food recognition. In: Proceedings of the 2016 ACM on Multimedia Conference, pp. 172–176. ACM (2016)
17. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)