Fall 10-16-2021

Machine learning approaches for dietary assessment

Pedro Martins
CISeD – Research Centre in Digital Services, Polytechnic of Viseu, pedromom@estgv.ipv.pt

Filipe Sá
CISeD – Research Centre in Digital Services, Polytechnic of Viseu, filipe.sa@estgv.ipv.pt

Maryam Abbasi
CISUC - Centre for Informatics and Systems, University of Coimbra, maryam@dei.uc.pt

Machine learning approaches for dietary assessment

[1]Pedro Martins, [1]Filipe Sá, [2]Maryam Abbasi

[1] CISeD – Research Centre in Digital Services,
Polytechnic of Viseu, Portugal
[2]CISUC - Centre for Informatics and Systems,
University of Coimbra, Portugal

pedromom@estgv.ipv.pt, filipe.sa@estgv.ipv.pt maryam@dei.uc.pt

Abstract

When considering individuals with dietary limitations, automatic food recognition, and assessment are paramount. Smartphone-oriented applications are convenient and handy when dish recognition and the elements inside are required. Machine learning (deep learning) applied to image recognition, alongside other classification techniques (for example, bag-of-words), are possible approaches to tackle this problem. The current most promising approach to the classification problem is deep leaning, which requires high computation for training, but it is an extremely fast and computationally light classifier. Since the requirement for the classifiers to be as accurate as possible, the humans must also be considered as the classifier. This work tests and compares deep-learning methods bag-of-words applied to computer vision, and the human visual system. Results show that deep learning is better when considering a low number of food categories. However, with more food categories, the human overcomes the machine algorithms.

Keywords – Machine learning, Food recognition, Human visual system, Image classification.

1 INTRODUCTION

Mobile devices with relatively good computational characteristics play an essential role in computer vision and assisting in better and healthy lifestyle dietary assessment. In food recognition, the first step is to automatically identify the plate containing the food and ignore all surroundings. Then, isolate food elements in the dish, consider the semantics (e.g., rice with beans, egg yolk, and white), and finally classify the food. Based on extracted data, eating habits can improve with any application to track the consumed food calories, recommending better lifestyle decisions.

Since Convolution Neural Networks (CNNs) were improved, image classification and object recognition became possible with reasonable accuracy. By matching CNNs with the current mobile computing capacity, manual identification of food can be replaced by an automatic classification, merely pointing the camera of a smartphone to the food plate. In this context, the challenge is to have a CNN that can at least match the humans. Otherwise, no point/trust is using an automated process to identify food elements on a plate. Therefore, it is critical to study if CNNs can outperform humans, and also if they are more efficient than other traditional approaches to the problem.

The selected traditional approach to test was the bag-of-words. The bag-of-words model can be applied to the representation of images in computer vision by treating image features as terms. A bag of words is a sparse vector of the frequency counts of terms in text classification; that is, a sparse histogram of the vocabulary (Sarwar et al., 2019). As a more state-of-the-art deep-learning approach, the same setup is tested with CNNs, namely with GoogLeNet (Al-Qizwini et al., 2017), Inception-v3 (Szegedy et al., 2016) and Resnet101 (Varga et al., 2018). Additionally, the human visual system is also testes and used as a baseline to compare with the other methods. It also includes training, learning new dishes of food e.g., for European food (Asiatic foods are unknown), and then identifying them. This paper describes the architectures of the bag-of-words pipeline and CNNs, as well as the created survey for training and testing the food recognition capacity of a selected group of humans.

The organization of this paper is in the following four sections. Section 2 resumes the vast related-work in the field. Section 3, describes the implementation of the approaches to test the bag-of-words, deep-learning, and human. Additionally, there is an explanation for the tests, training set, and the experimental setup. Finally, Section 4 discuss the experimental results, and Section 5 concludes the study with a review of the accomplishments and discuss future work guidelines.

2 RESUMED RELATED WORK

Food recognition is a challenging problem because of deformable objects and high intra-class variation. Different shapes, textures, and colors that food can assume, create demanding obstacles for machine-learning to identify dishes and their content.

Authors in (Baxter, 2012) and (Yang et al., 2010) use machine learning combined with statistics and spatial relationships between ingredient labels to detect food, whereas ingredients are statistically related to each other to increase classification accuracy.

In (Joutou and Yanai, 2009), authors propose an automatic food image recognition system for recording people's eating habits. They apply the Multiple Kernel Learning method to integrate several kinds of image features such as color, texture, and SIFT adaptively, reaching in their experimental results a 61.34% classification rate for 50 types of foods. Also, in (Matsuda et al., 2012), candidate regions are detected and using bag-of-words combined with SIFT, spatial pyramid, histograms, and Gabor texture, experimental results show 56% accuracy when using ten classes of food types.

In the last decade, deep learning raised with Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, presenting AlexNet, a CNN oriented to recognize objects in images. Results with AlexNet outcome the, at the time state-of-the-art, and they have become frequent in all kinds of tasks, such as gender recognition from facial images (Arora and Bhatia, 2018).

CNNs stand out from other traditional methods when processing images by including convolution layers to extract and filter images' features automatically. The concept of back-propagation is also introduced in CNNS and extended to all convolution layers, which allow extracting information from images automatically (Shen et al., 2016). New CNNs are developed along with research to improve the accuracy of Alexnet. Resnet is an evolution of Alexnet, by adding residual blocks to manage gradients that vanish in the first layers on back-propagation execution, increasing depth, Resnet can improve its accuracy (Cheng et al., 2017) (He et al., 2016).

By introducing sparse connections between layers and using different filter sizes, the CNN Inception can capture more features with a more varied degree of detail to push performance, both in terms of speed and accuracy (Szegedy et al., 2015) (Baldassarre et al., 2017).

One of the motivations for this work is the fact that authors using CNNs applied to food recognition were able to achieve better results than other approaches in general. Authors in (Kawano and Yanai, 2014a) present a technique using Deep Convolutional Neural Network to boost food recognition accuracy significantly by integrating it with conventional hand-crafted image features, Fisher Vectors with HoG and Color patches. Results show they have achieved 72.26% as the top-1 accuracy and 92.00% as the top-5 accuracy for the 100-class food dataset, UEC-FOOD100, which outperforms the best classification accuracy of this dataset reported so far, 59.6%.

Authors in (Pouladzadeh and Shirmohammadi, 2017) and (Kawano and Yanai, 2013) propose similar mobile apps for food recognition, both using deep-learning. Their solution requires the user to build a bounding box around the elements of food manually — performed experiments with 7000 food images, from 30 different classes (FooD dataset). Results show an average recall of 90.98%, the precision of 93.05%, and accuracy of 94.11% compared to 50.8% to 88% accuracy of other food recognition systems. In general, the results are striking good. However, they rely too much on user action:

- The user needs to interact with the app and manually set the bounding box. Thus, this makes the approach incomparable with others that require no interaction from the user (except taking the photo).
- The dataset is not comprehensive. It includes easily recognizable items, such as everyday fruits.
- It also has only 30 different classes of foods, which is too less, and it influences the accuracy.

Other authors, (Yanai and Kawano, 2015), examined the efficacy of the deep convolutional neural network, applied to food recognition. In their work, they search for the best combination between pre-training with the large-scale ImageNet data, fine-tuning, and activation features extracted from the pre-trained deep convolutional neural network. Results show that using a pre-trained network, including 1000 food-related categories, was the best method, achieving 78.77% as accuracy for UEC-FOOD100 and 67.57% for UECFOOD256 (Kawano and Yanai, 2014b).

When analyzing results from (Yanai and Kawano, 2015; Hassannejad et al., 2016; Ciocca et al., 2017) and others, it was evident that accuracy is far from the optimal 100%. Based on our previous works over this topic (Caldeira et al., 2019), CNNs outperform humans with a low number of classes of foods. However, if we increase the number of food classes, humans are better. Hence, in this work, we question whether using an approach based on bag-of-words will optimize the efficiency of the CNN.

3 METHODOLOGY

Food classification using deep-learning techniques is the most used approach when classifying food dishes. Another way of getting the work done is by using humanbased identification, where a human interacts to identify each food. The human-based approach is essential to use as a comparison baseline. Succeeding sections describe how the problem (identify and classify food) approached to compare and withdraw conclusions.

Our implementation of the algorithm bag-of-words is in Matlab, using the bag-offeatures object, which allows implementing specific image feature extractors. The hardware setup for all computational, experimental work consisted of an Intel i7 4.7GHz processor, 32GB RAM, Asus Nvidia GeForce Gtx 1070 8GB, Windows 10 64 bits, MatLab R2018a.

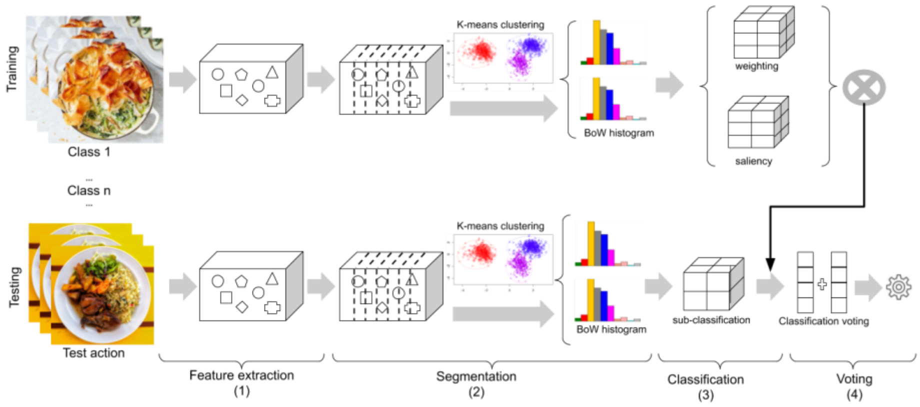


Figure 1: Bag-of-words architecture sketch

3.1 Bag-of-Words architecture

Bag-of-words is a machine learning classification technique. The process rep rents images as words, generating a histogram of the visual representation that occurs in the image (Csurka et al., 2004). Such histograms are used for the preparation of an image recognition classifier. In general, the bag-of-words extract features to represent the image (using a training dataset), then creates a visual vocabulary (bag of features), then classifies the image.

Figure 1 depicts the bag-of-words architectural steps to extract features and classify an image. In (1) is represented the feature extraction, for both training and testing pipelines. These features are described in the bag-of-words approach, by vectors of features. In this step, oriented to the proposed research, were extracted features such as texture (GLCM binary patterns), color histograms, geometry features of regions (6 layers the start of step (2)), and also SURF. Using k-means algorithm images features from the training dataset are extracted and clustered, and this allows obtaining k feature vectors. The k feature vectors represent the centroid of each class feature from the training dataset (2) visual words are grouped and separated by similar characteristics and defined as a vocabulary histogram. In the classification step (3), a trained artificial neural network combines all weighted vectors and saliencies that are analyzed based on a ReLU function. Feature visualization is the same as the forward pass of the deep convolutional network, which only leaves the input’s positive components. After step (3), the already trained neural network feeds the testing images in step (4) and label pictures. Once the BOW trained, the method identifies and removes features from the training images provided every new picture for classification and generates the histogram for the picture occurring in codewords. Afterwards, the qualified classifier is used to label the picture as one of the groups (step (4)).

3.2 Deep learning architecture(s)

With the objective of comparing the accuracy of CNNs a set of networks were selected for testing: GoogleNet (Tang et al., 2017), Inception-v3 (Nivrito et al., 2016), and Resnet101 (Doersch and Zisserman, 2017).

Before Resnet, neural networks struggled with problems in gradients ignored and then removed during the back-propagation learning process, affecting the number of layers. Thus, Resnet architecture came to improve how deep-learning-networks training uses

Stochastic Gradient Descent through the residual modules (Wang et al., 2017). The residual modules are sub-architectural-blocks, which feed both the next layer and layers that are two or three hops distance. Passing forward the residual modules, instead of discarding them, enables deep-learning-networks to minimize vanishing gradients problems, and to classification more appropriately.

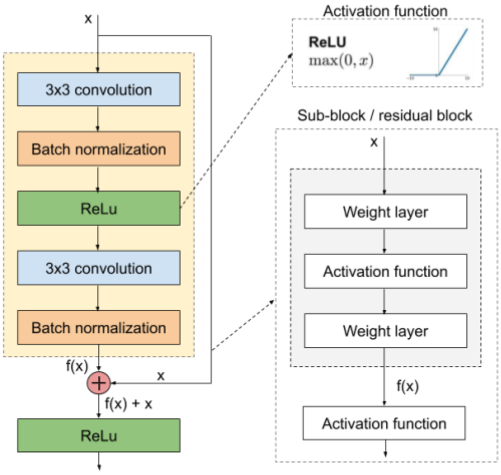


Figure 2: Resnet block architecture

Figure 2 shows a small illustration of Resnet architecture. Several layers compose Resnet, and at each layer, the output of the previews is added.

Resnet follows VGG's full 3 × 3 convolutional layer design. The residual block has two 3 × 3 convolutional layers with the same number of output channels. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function.

Then, these two convolution operations add the input directly before the final ReLU activation function. This kind of design requires that the output of the two convolutional layers be of the same shape as the input so that they can be added together. Let us focus on a local neural network, as depicted. Denote the input by x. The ideal mapping by learning is $f(x)$, to be used as the input to the activation function. The portion within the dotted-line box must directly fit the mapping $f(x)$.

GoogleNet first introduces the inception architecture concept, and later the CNN inception improves it. Inceptionv3 CNN introduces a special inception module is adopted to improve model performance, which is a multi-level feature extraction that computes 1x1, 3x3 and 5x5 convolutions, all in the same network module. In the inception module, convolutional layers with different filter sizes are computed in parallel. Resulting features are concatenated before passed to the next layer. The increase in features significantly increases the learning power of the model (Szegedy et al., 2016). Also, pooling operations are essential for the Inception convolutional network; hence, a parallel pooling path is added to reduce the amount of data and the computation time (grid size reduction). Softmax activation outputs the probabilities of each class. The architecture for the inceptionv3 model is described in Figure 3. Note that the outputs of the inception filters are all stacked and used as input to the next layer.

In the experimental testing setup, the number of layers for the architectures, Resnet, Inception, and Googlenet, were 101, 48 and 22, followed by training for a larger number of epochs (300) to ensure the convergence of the network. All the three CNNs were pre-trained (with a selected set of images) to classify images (from another distinct collection of images) from the food dataset UECFOOD256 (Kawano and Yanai, 2014b). The training and classification were either all 256 categories or 16 randomly selected categories. The training rate was initially configured to 0.05, validating every four interaction, preliminary conclusion indicates accuracy convergence in every run.

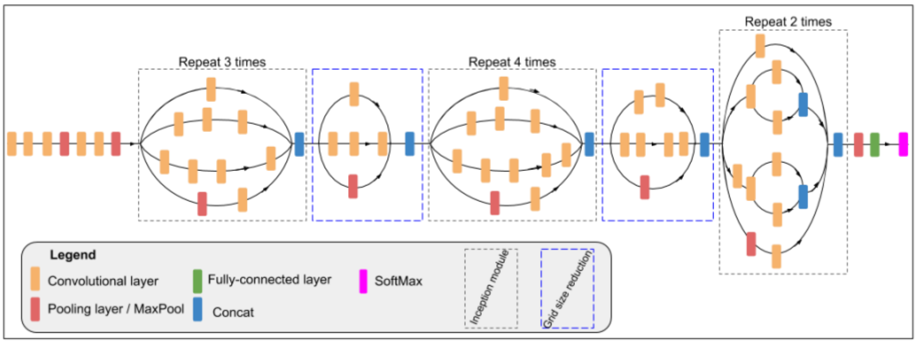


Figure 3: Simplified inception architecture, ConvNets model

Procedures to classify new food images consisted of loading the food image to be categorized, extract features automatically (using the CNN convolutions layers) and giving learned weights to each feature. As a result, food is classified as one of those types, ending with a group of probabilities that the submitted food is one of the before learned ones. Note that, an independent dataset is used to evaluate accuracy.

3.3 Baseline survey

A survey in Asian foods was mounted and delivered to humans without previews knowledge in these food classes. A test group of 20 subjects was accepted (the ones not knowing at least 80% of the dishes) all in the age rank between 25 and 40 years old, all with European nationality. As control questions, two items in the survey were universal foods, such as Pizza.

Training is necessary before testing humans. The challenge then is to train the subject to be able to pick the correct name of food dishes presented to him, and the accuracy is measured as the percentage of the right choices. For this reason, the survey was divided into two stages.

A training stage was given to the human subjects, consisting of showing a set of total 98 screens for the 16 food categories (a number that was deemed sufficiently small to allow humans to keep the attention and simultaneously not too small to be too easy). Each screen shows seven images for the chosen food category and one image from other food categories, in random order. Then the test subject had to indicate the one that did not belong to the respective food category. This allows the test subjects to learn the names and key characteristics of each food category. Figure 4 shows a set of dishes from the same category (identified by the title), and one image that does not belong (i.e., the user needs to identify the dish image that does not fit) marked with an ”x”.

In the second stage, image classification, a set of images, based on 32 screens, is shown in random order, from the 16 trained food dish classes, each with 16 class label options, where only one is correct. The test subjects need to select the right label for the image. Figure 6 shows one of the screens used to test the human classification accuracy. In the figure, on the left side, is given the dish image belonging to an unknown class, and on the right, the dish class options for the respondent to fill with the correct option on the and then follows to the next screen.

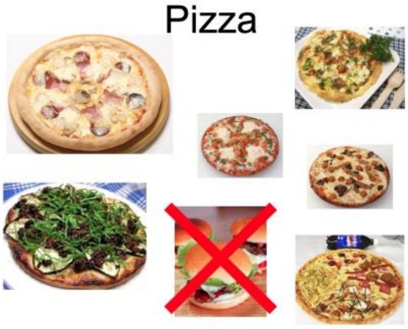


Figure 4: Human, training image example food class "Pizza"

4 RESULTS

The training time for Resnet101 was 8018 minutes for 256 food categories and almost 226 minutes for 16 classes. Figure 5 compares the accuracy of the methods: Human, Resnet101, GoogleNet, Inception v3, Bag-of-words (500 words vocabulary), and Bagof-words (1000 words vocabulary), all with two approaches, 16 food categories and 256 food categories (except for the human survey). In the bag-of-words, results show that increasing the number of codewords does not bring a significant accuracy performance, and still, the bag-of-words was worse than CNNs and humans. Results also show that for 16 food categories, CNNs were better than humans. A more realistic test, using 256 food categories, indicates that CNNs performance decreases significantly. The best performing CNN was Resnet with 95% and 75% accuracy for 16 and 256 food categories, respectively.

Figure 7 shows some class precision values from classification of: Humans (16 food classes); Resnet (256 categories); Inceptionv3 (256 categories); GoogleNet (256 categories); and Bag-of-words (1000 code-words). The analysis of the values in chat, Figure 7 allow concluding that the hardest classification categories for humans and neural networks are different, which lead to humans and CNNs to have distinct behaviors when classifying types of food types.

A global analysis of results (Figure 5 and 7) shows that CNNs are more accurate than bag-of-words, simultaneously show that, as the number of categories to learn increase (16 to 256 food categories) the accuracy decreases. Consequently, humans knowledge continues to outperform the accuracy of CNNs when identifying different food types.

5 CONCLUSIONS

In this work, bag-of-words is compared with CNNs and with humans. The experimental comparison process allows us to conclude that CNNs are more accurate than the bag-of-words, while at the same time showing CNNs accuracy decreases when the number of food categories increases (considering 16 to 256 food groups). As a result, the efficiency of CNNs tends to be exceeded by human intelligence when considering various food types. Although different classifications of people and deep learning are robust, human and CNNs attitudes are distinct when classifying various food types.

Future work includes better research on how humans learn and classify food dishes, so that, those learning details can also be reinforced on CNNs. Additionally, other image variation factors of impact, such as illumination, camera angles, and shadows, need to be perfected in CNNs. Finally, we propose the integration into mobile apps of these learning capabilities applied not only to photos but also video, for instance, using approaches such as YOLO (You Only Look Once) (Du, 2018).

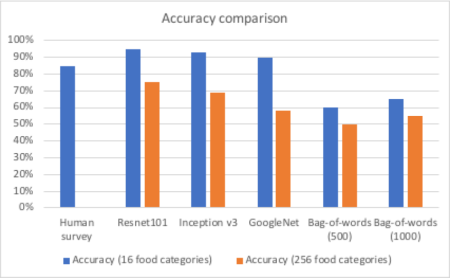
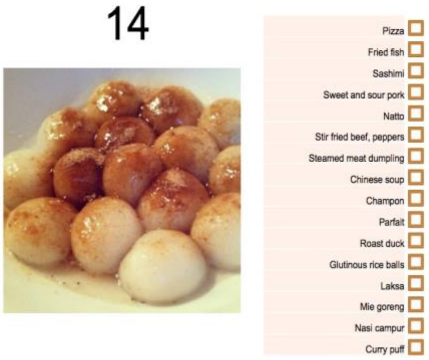


Figure5:Accuracycomparison.



Pizza □
Fried fish □
Sashimi □
Sweet and sour pork □
Natto □
Stir fried beef, peppers □
Steamed meat dumpling □
Chinese soup □
Champon □
Parfait □
Roast duck □
Glutinous rice balls □
Laksa □
Mie goreng □
Nasi campur □
Curry puff □

Figure 6: Human, classification screen.

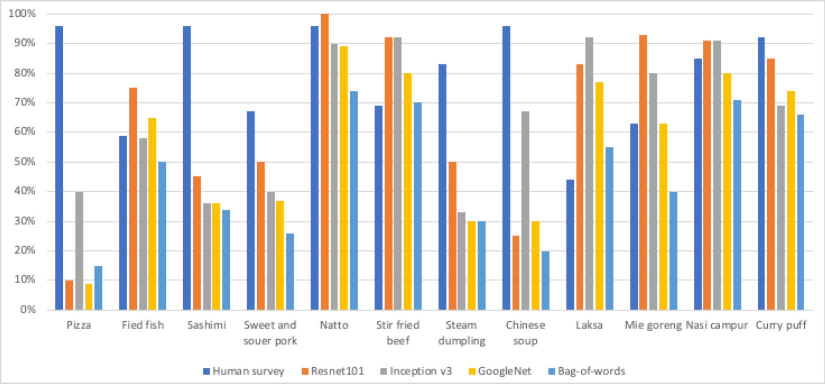


Figure 7: Some food classification accuracy's

ACKNOWLEDGEMENTS

"This work is funded by National Funds through the FCT Foundation for Science and Technology, I.P., within the scope of the project Ref UIDB/05583/2020. Furthermore, we would like to thank the Research Centre in Digital Services (CISeD), the Polytechnic of Viseu for their support."

References

Al-Qizwini, M., Barjasteh, I., Al-Qassab, H., and Radha, H. (2017). Deep learning algorithm for autonomous driving using googlenet. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 89–96. IEEE.

Arora, S. and Bhatia, M. (2018). A robust approach for gender recognition using deep learning. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Baldassarre, F., Mor´ın, D. G., and Rod´es-Guirao, L. (2017). Deep koalarization: Image colorization using cnns and inception-resnet-v2. *arXiv preprint arXiv:1712.03400*.

Baxter, J. (2012). Food recognition using ingredient-level features. *Ηλεκτρονικο´]. Available: http://jaybaxter. net/6869 food project. pdf*.

Caldeira, M., Martins, P., Cec´ılio, J., and Furtado, P. (2019). Comparison study on convolution neural networks (cnns) vs. human visual system (hvs). In *International Conference: Beyond Databases, Architectures and Structures*, pages 111–125. Springer.

Cheng, X., Zhang, Y., Chen, Y., Wu, Y., and Yue, Y. (2017). Pest identification via deep residual learning in complex background. *Computers and Electronics in Agriculture*, 141:351–356.

Ciocca, G., Napoletano, P., and Schettini, R. (2017). Learning cnn-based features for retrieval of food images. In *International Conference on Image Analysis and Processing*, pages 426–434. Springer.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.

Doersch, C. and Zisserman, A. (2017). Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060.

Du, J. (2018). Understanding of object detection based on cnn family and yolo. In *Journal of Physics: Conference Series*, volume 1004, page 012029. IOP Publishing.

Hassannejad, H., Matrella, G., Ciampolini, P., De Munari, I., Mordonini, M., and Cagnoni, S. (2016). Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 41–49.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Kawano, Y. and Yanai, K. (2013). Real-time mobile food recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7.

Kawano, Y. and Yanai, K. (2014a). Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 589–593.

Kawano, Y. and Yanai, K. (2014b). Foodcam-256: a large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 761–762.

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

Nivrito, A., Wahed, M., Bin, R., et al. (2016). *Comparative analysis between Inceptionv3 and other learning systems using facial expressions detection*. PhD thesis, Brac University.

Pouladzadeh, P. and Shirmohammadi, S. (2017). Mobile multi-food recognition using deep learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(3s):1–21.

Sarwar, A., Mehmood, Z., Saba, T., Qazi, K. A., Adnan, A., and Jamal, H. (2019). A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. *Journal of Information Science*, 45(1):117–135.

Shen, L., Lin, Z., and Huang, Q. (2016). Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tang, P., Wang, H., and Kwong, S. (2017). G-ms2f: Googlenet based multi-stage feature fusion of deep cnn for scene recognition. *Neurocomputing*, 225:188–197.

Varga, D., Saupe, D., and Szira´nyi, T. (2018). Deeprn: A content preserving deep architecture for blind image quality assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Yanai, K. and 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)*, pages 1–6. IEEE.

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