A Deep Learning-Based Dessert Recognition System for Automated Dietary Assessment

Dimitrios-Marios Exarchou, Anastasios Alexiadis(✉), Andreas Triantafyllidis, Dimosthenis Ioannidis, Konstantinos Votis, and Dimitrios Tzovaras

Centre for Research and Technology Hellas, Information Technologies Institute (CERTH/ITI), 6km Charilaou -Thermi, Thessaloniki, Greece
{exarchou.dimitris,talex,atriand,djoannid,kvotis, Dimitrios.Tzovaras}@iti.gr

Abstract. Over the past few years, a significant part of scientific research has been focused on the assistance of patients who suffer from obesity or diabetes. Monitoring the food intake through self-report in diet control applications has been proven both time-consuming and non-practical and can be easily sidelined especially by children. In this paper, we propose the design and development of a novel system, which will assist obese or diabetic patients. We have implemented transfer learning as well as fine-tuning to different pre-trained CNN models to automatically distinguish dessert from non-dessert food images. For further training of these deep neural networks, a new dataset was constructed, which derived from the original Food-101 dataset. To be precise, 19 categories of desserts were used, which correspond to 19K images combined with 19K images of non-desserts. Google InceptionV3 architecture appeared to have the best performance, reaching a validation accuracy of 95.89%. To demonstrate feasibility of out platform and the independence of data biases, we constructed another data collection of food images, which was captured under challenging light and angle of capture conditions.

Keywords: CNN · Computer vision · Deep learning · Image recognition · Dessert recognition · Food Image · Image pre-processing · Diabetes · Obesity

1 Introduction

It is known that food intake is essential for the preservation of human life. The nutrients in food enable the cells in our bodies to perform their necessary functions. However, it is also proven to be an important risk factor for people who suffer from many common chronic non-communicable diseases such as obesity [1, 2], which comprises one alarming public health issue. In 2016, more than 1.9 billion adults and over 340 million adolescents were overweight according to the World Health Organization[1]. In recent years, junk food has become part of our everyday diet, especially for children [3–5].

[1] https://www.who.int/.

X. Gao et al. (Eds.): MobiHealth 2021, LNICST 440, pp. 60–70, 2022.
https://doi.org/10.1007/978-3-031-06368-8_4

Multimedia and commercial stimuli have led many young people to consume unhealthy edible products, unaware of their poor nutritional value. Serious life-threatening diseases, such as childhood obesity and diabetes emerge mainly in western countries, due to improper diet habits.

To address this situation, the nutrition industry supplied several diet plan proposals. At the same time, diet tracking applications have been developed, employing large databases of food images and their respective nutritional values. Initially, this issue was confronted by the development of applications for manual food logging and user's self-reports, which demonstrated warnings and dietary plans [6].

Nevertheless, those recording systems were often used incorrectly by the patients since they relied on manual input. They were prone to recall bias issues and could mislead, resulting in the exacerbation of patients' medical conditions. In addition, because of their complex usage, they were often easily sidelined. More specifically, children and adolescents face difficulties in self-reporting food intakes due to issues related to recall or monotonous text reporting.

This work aims to the implementation of a new technology, which will automatically identify foods containing sugar. The feasibility of these research efforts should assist in the maintenance of a healthier lifestyle by encouraging healthy food consumption at an early age. In contrast with previous research, this work probes the performance of an automated dessert recognition system under challenging image acquisition conditions.

2 Related Work

When focusing on this area, there are limitations and challenges to consider, such as the negligible intra-class differences within the food images. Data collections were soon created, playing the role of benchmarks. Food-5 K on the one hand and Food-11 on the other were used to solve the binary and the multiclass problem respectively. The preliminary approach employed hand-crafted features and the extraction of feature descriptors, such as SIFT, SURF, BRIEF, etc. [7]. However, this traditional method was outperformed by newer technological developments, such as deep learning techniques. The extraction of features is now performed by CNNs, which convolve the input image with specific kernels (also known as filters). Those visual features are reflected in the activation of internal neurons' layers. While the layers on the input side correspond to more syntactic information, the layers closer to the output convey more semantic information [8]. The two methods are illustrated in Fig. 1.

Bossard et al., [9] introduced a publicly available food image dataset, Food-101[2], with 101 food categories. They also examined a weakly supervised method to mine discriminative components with Random Forests, reaching an accuracy of 50.76%. More recently, Şengür et al. examined a feature concatenation method [10], employing the last two fully connected layers of AlexNet and VGG-16. The reported accuracy on the binary problem was 99%, while on the Food-101 challenge they achieved 79.86%. Attokaren et al. [11] leveraged the pretrained Google InceptionV3 model in combination with a multi-crop evaluation technique and obtained 86.97% accuracy. Apart from the

[2] https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

general food recognition problem, the scientific community focused on the detection of unhealthy eating habits. Aiming to strengthen the motivation of children to adopt healthy diet habits, a recent work [12] proposed a social robot-based platform, based on camera images that are automatically captured by a commercially available social robot. The measured validation accuracy in a dataset of 53884 images was 99.68%.

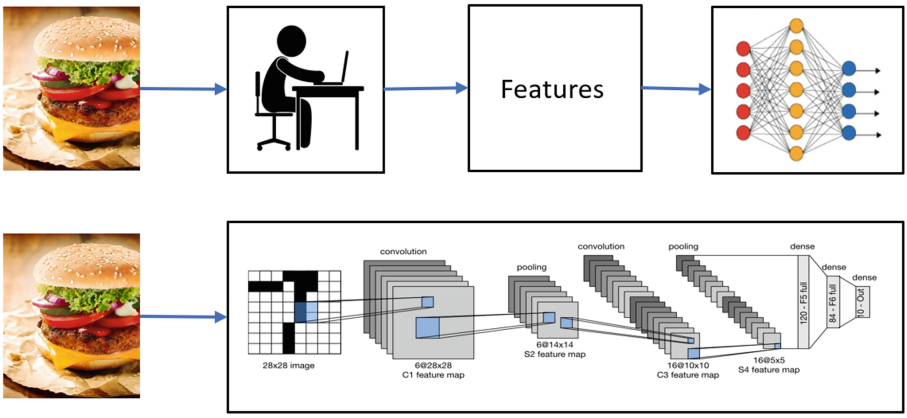

Fig. 1. (a) Traditional computer vision vs (b) Deep learning technique.

In this paper, we present the design and deployment of a dessert recognition system. The outcomes of this research are intended to contribute to the development of an end-to-end diet suggestion assistant that will address the threatening issue of childhood obesity. The structure of the paper is as follows; in the next section, we introduce the proposed methodology. Section 4 demonstrates the results and feasibility of our method. Finally, we discuss the conclusions in Sect. 5.

3 Proposed Method

In this section, we present the proposed workflow for system development. We describe our methods for dataset construction, data augmentation, transfer learning and the training process.

3.1 Dataset Construction

To the best of our knowledge, there is no previous research on the topic of automatic dessert recognition. Thus, the initial challenge was to create a dessert dataset. We gathered food images from the Food-101 dataset. Specifically, 19 categories (apple pie, baklava, beignets, carrot cake, cheesecake, chocolate cake, chocolate mousse, churros, crème brulee, cupcakes, donuts, frozen yogurt, ice cream, macarons, panna cotta, red velvet cake, strawberry shortcake, tiramisu and waffles) provided us with 19k images of desserts, while the remaining 82 categories were used to randomly sample without

replacement 19k images of non-dessert images. The workflow described above resulted in a dataset of 38k images and was subsequently split into training, validation and testing partitions. 80% of the dataset was used for the training set, while validation and testing sets corresponded to the 15% and 5% of the dataset, respectively. Some samples are illustrated in Fig. 2.

Fig. 2. Dessert images of desserts from 19 categories

3.2 Data Augmentation

To enrich the training samples and enhance the generalizability of the system, a data augmentation process was implemented. Hence, we used *torchvision*'s transforms to randomly flip the input image horizontally and vertically, to rotate it in a random angle from 0 to 45° and to adjust the color jitter. For the classification of unknown images, we used a multi-crop (10-crop) strategy, which produced 10 crops for each image (upper left, upper right, lower right, lower left, centre and their flipped versions). The images were finally normalized, according to the *ImageNet* standards.

3.3 Transfer Learning

The proposed method is based on the fine-tuning of deep pre-trained CNNs. In this paper we compare four different prestigious architectures: Google InceptionV3 [13], Resnet101 [14], VGG16 [15] and MobileNet [16] to address the problem of binary classification. The InceptionV3's architecture is shown in Fig. 3. For the implementation of the method, both *Tensorflow* and *PyTorch* frameworks were used. The output of the models was modified to the following sequential layers:

- Average Pooling 2D with output size = (1,1)
- Dropout (with different probability depending on the different architectures)

- Linear layer with 1 node, L2 regularization = 0.0005, sigmoid activation function and Xavier uniform initializer [17].

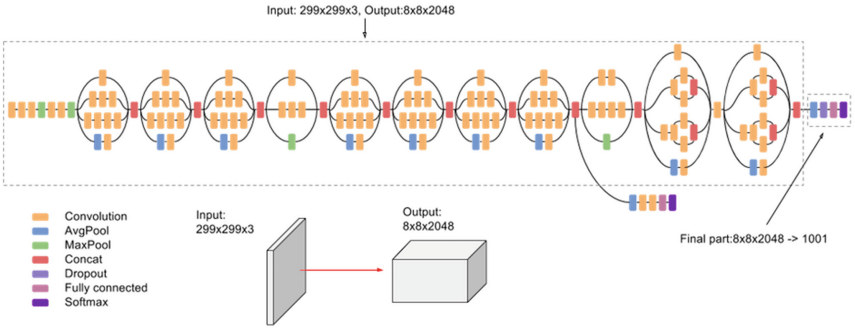

Fig. 3. InceptionV3 architecture (https://cloud.google.com/tpu/docs/inceptionv3-advanced)

3.4 Training Process

Three different strategies were considered for the fine tuning of the models described above. Firstly, the layers of the models were immediately unfrozen, while on the second case we examined a more gradual learning. Specifically, the second method included an initial training of the classifier layer, followed by the training of the entire model. Nonetheless, the impact of the interpolation of an intermediate training stage on a part of the total model's parameters was also investigated.

The optimization algorithm used was the Stochastic Gradient Descent, which is an optimizer with a good performance over large data-sets. The loss function used across all models was binary cross entropy:

$$J(w) = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log\left(1-\hat{y}_n\right)\right] \tag{1}$$

The hyperparameters used in each training scenario were generally different; however, the initial learning rate was 0.008 and after a certain number of epochs it was decreased by half for each training stage. The momentum was set to 0.9. The hyperparameters of the InceptionV3 model, which achieved the highest accuracy in our evaluation set, are the following:

- Initial learning rate = 0.008
- Momentum = 0.9
- Decreasing learning rate by half every 10 epochs.
- Trained to binary cross-entropy loss 0.0910 after 27 epochs.

4 Results

This section describes and compares different results found in our experiments. The obtained results for various models on the original Food-101 dataset are tabulated in Table 1. To generalize our conclusions, we examine the effectiveness of our system for manually captured images in Table 2.

4.1 Food-101 Evaluation

Our system development utilized *Tensorflow* and *PyTorch*, two open-source machine learning platforms. The training process was executed on Google Colaboratory[3], leveraging 12 GB of RAM and an NVIDIA Tesla K80. Different deep learning architectures and training strategies were simulated. The three different strategies described above are indicated as *unfrozen (1 stage)*, *2 stages* and *3 stages*. Classification accuracy and binary cross entropy loss were used as evaluation metrics and they are illustrated in the following table:

Table 1. Training metrics

Model	*Stages*	*Dropout*	*Train Loss*	*Val. Loss*	*Train Acc.*	*Val. Acc.*
InceptionV3	*unfrozen*	*0.5*	*0.0910*	*0.1164*	*96.30%*	*95.89%*
InceptionV3	*3 stages*	*0.5*	*0.0777*	*0.1230*	*96.83%*	*96.16%*
InceptionV3	*2 stages*	*0.75*	*0.0733*	*0.1616*	*97.12%*	*95.46%*
ResNet101	*unfrozen*	*0.6*	*0.1451*	*0.1269*	*93.99%*	*95.68%*
ResNet101	*2 stages*	*0.75*	*0.0776*	*0.2379*	*96.91%*	*94.12%*
VGG16	*unfrozen*	*0.5*	*0.4383*	*0.3475*	*79.10%*	*85.77%*
MobileNet	*unfrozen*	*0.2*	*0.1097*	*0.1309*	*95.63%*	*95.42%*

For a more detailed interpretation of the results, we extracted the confusion matrices and the ROC curves for the validation and training sets of the aforementioned dataset. As has already been discussed, we employed a voting ten-crop strategy to achieve a higher evaluation performance. The measured accuracies were 95.89% and 95.79%, for the validation and the test set, respectively. Those evaluation metrics for the unfrozen InceptionV3 model can be seen in Fig. 4 and Fig. 5.

[3] https://colab.research.google.com/notebooks/intro.ipynb?utm_source=scs-index.

4.2 Real Conditions Evaluation

To evaluate the performance of our model in real food image captures, with challenging light conditions and capture angle, we also created a data collection of 214 real-conditions images. Some examples are depicted in Fig. 6. To evaluate the real-conditions performance, we examined the ten-crop strategy in comparison with the simple prediction method. The highest classification accuracy was obtained for the InceptionV3 unfrozen trained model, while ResNet101 performed also up to the mark. The performances of the different architectures are shown in Table 2. Some examples of predictions are illustrated in Fig. 7.

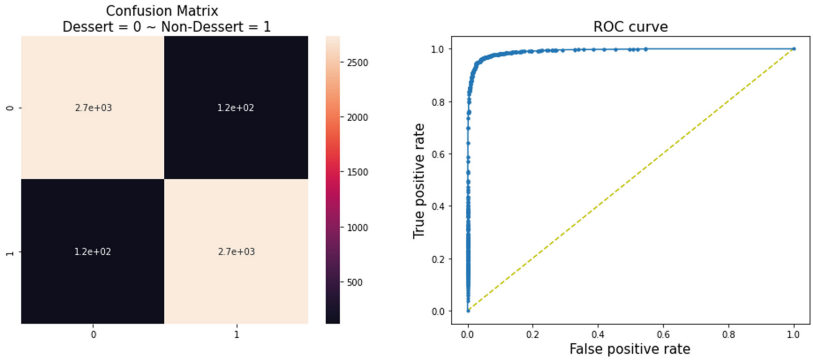

Fig. 4. Confusion matrix and ROC curve for the validation set

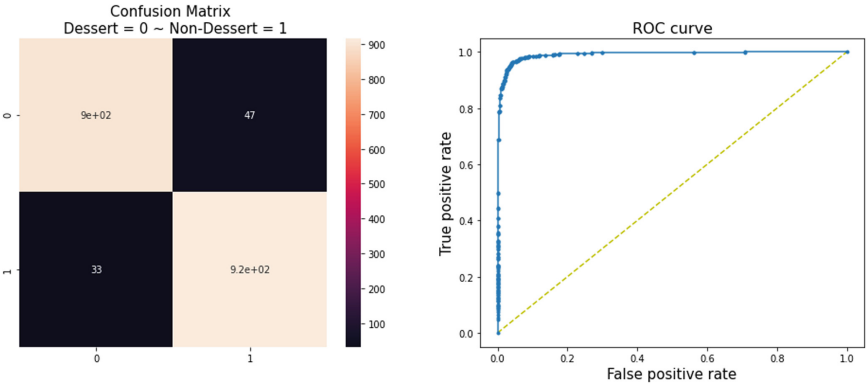

Fig. 5. Confusion matrix and ROC curve for the test set

Fig. 6. Real captures examples

Table 2. Real-conditions captures performance

Model	*Training*	*Evaluation Acc. without ten-crop*	*Evaluation Acc. with ten-crop*
InceptionV3	*unfrozen*	*94.86%*	*95.79%*
InceptionV3	*3 stages*	*89.25%*	*92.52%*
ResNet101	*unfrozen*	*89.25%*	*92.99%*
VGG16	*unfrozen*	*85.05%*	*82.24%*
MobileNet	*unfrozen*	*90.65%*	*89.25%*

4.3 Computational Complexity

In the real scene, such systems are usually deployed on mobile devices, which have limited storage and battery capacity. Thus, computational complexity is as important as accuracy of recognition. Computational complexity depends on both the total number of parameters and the prediction time. The prediction time of an image includes the time of image transformation, as well as the time of forward propagation in the model. These metrics for the various models are shown in Figs. 8 and 9, respectively. The lightest and fastest model is MobileNet, while the slowest and most complex is VGG16. In our analysis we chose InceptionV3. This model is twice as slow as MobileNet and requires five times more space to store. Therefore, in applications that require speed or low memory consumption MobileNet could be selected.

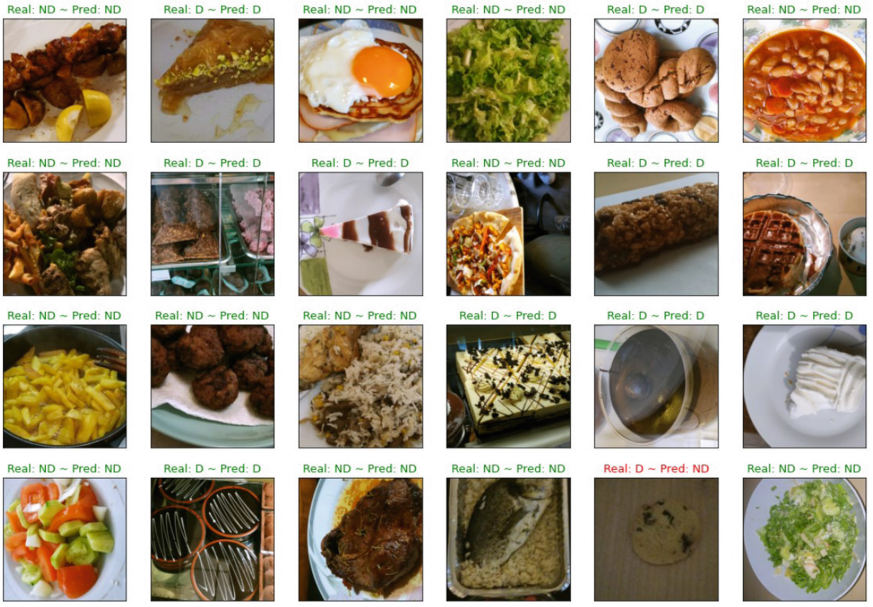

Fig. 7. Real captures predictions

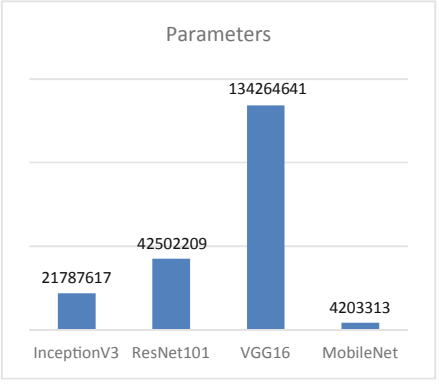

Fig. 8. Parameters of models

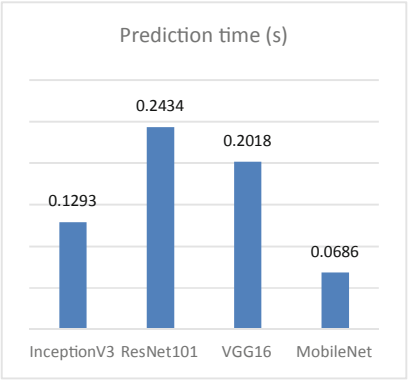

Fig. 9. Prediction time of models

5 Discussion

In the past decade the interest of the scientific community in computer vision was ignited by the challenging task of food recognition. As machine learning approaches require a large range of data, popular large datasets for food classification were presented and the necessity for even richer datasets with specific attributes was emphasized. In this paper, we aimed to solve the urgent problem of recognizing foods containing sugar by leveraging deep learning techniques. We presented the design of a system for automated diet tracking, which was relied on pretrained CNNs. Our model constitutes an autonomous

computer vision system aimed at the dietary assessment of obese or diabetic patients. Our objective was to create a system that can be used requiring minimal user interactions, so that it can be used independently from the user's technological literacy.

For the fine tuning of our model, we introduced a new data collection which derived from the original Food-101 dataset. This data collection was split into training, validation and testing sets and the reported evaluation accuracy was 95.89% and 95.79% for the validation and test set, respectively. To demonstrate the adequacy of our system we measured its performance in real conditions. Upon examining the results of the experiments, it is worth mentioning that only InceptionV3 that followed the unfrozen training strategy, managed to cope with this difficult challenge. This shows that with proper training data there are no limitations on the conditions under which a photograph is captured. However, there are restrictions on the identification of types of foods, with which the model has not previously interacted during the training phase.

Future work would involve some optimization on hyperparameters and model components such as which layers to freeze during transfer learning. Due to limited computing resources and time constraints, the choice of the model architecture was made empirically with respect to the measured performance. Nonetheless, a grid-search for the optimization of hyper parameters search would have been more efficient. Furthermore, we contemplate reinforcing the capabilities of our system to recognize different food categories. An automatic calories estimator would provide a crucial assistance in the fight against obesity. Finally, the usability of the system will significantly improve if we integrate it into mobile devices, creating an Android application.

In conclusion, with this work we laid the foundations for the creation of more specific datasets around food and the nutritional value of the individual ingredients. The scope of this research includes the mobilization of technological advances in the direction of combating the scourge of the unhealthy diet. For this reason, the work presented in this paper is a step towards engagement with healthy dietary habits.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement «GATEKEEPER/857223 Smart Living Homes – Whole Interventions Demonstrator for People at Health and Social Risks» (KOH.021064).

References

1. Shields, M., Tremblay, M.S., Connor Gerber, S., Janssen, I.: Abdominal obesity and cardiovascular disease risk factors within body mass index categories. Heal. Rep. **23**(2), 7–15 (2012)
2. Vucetic, I., Stains, J.P.: Obesity and cancer risk: evidence, mechanisms, and recommendations. Ann. N. Y. Acad. Sci. **1271**(1), 37–43 (2012). https://doi.org/10.1111/j.1749-6632.2012.06750.x
3. Tate, E.B., et al.: mHealth approaches to child obesity prevention: successes, unique challenges, and next directions. Transl. Behav. Med. **3**(4), 406–415 (2013). https://doi.org/10.1007/s13142-013-0222-3
4. Smith, A.J., Skow, A., Bodurtha, J., Kinra, S.: Health information technology in screening and treatment of child obesity: a systematic review. Pediatrics **131**(3), e894–e902 (2013). https://doi.org/10.1542/peds.2012-2011

5. Lau, P.W.C., Lau, E.Y., Wong, D.P., Ransdell, L.: A systematic review of information and communication technology–based interventions for promoting physical activity behavior change in children and adolescents. J. Med. Internet Res. **13**(3), e1533 (2011). https://doi.org/10.2196/jmir.1533
6. Abril, E.P.: Tracking myself: assessing the contribution of mobile technologies for self-trackers of weight, diet, or exercise. J. Health Commun. **21**(6), 638–646 (2016). https://doi.org/10.1080/10810730.2016.1153756
7. O'Mahony, N., et al.: Deep learning vs. Traditional computer vision. In: Arai, K., Kapoor, S. (eds.) Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1, pp. 128–144. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-17795-9_10
8. Amato, G., Bolettieri, P., de Lira, V.M., Muntean, C.I., Perego, R., Renso, C.: Social media image recognition for food trend analysis. In: SIGIR 2017 Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017)
9. 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) Computer Vision - ECCV 2014. ECCV 2014 Lecture Notes in Computer Science, vol 8694. Springer, Cham, pp 446–461 (2014). https://doi.org/10.1007/978-3-319-10599-4_29
10. Şengür, A. Akbulut, Y. Budakm, Ü.: Food image classification with deep features. In: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP)
11. Attokaren, D., Fernandes, I., Sriram, A., Murthy, Y.V., Koolagudi, S.: Food classification from images using convolutional neural networks. In: Proceeding of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, 5–8 Nov 2017
12. Alexiadis, A., Triantafyllidis, A., Elmas, D., Gerovasilis, G., Votis, K., Tzovaras, D.: A social robot-based platform towards automated diet tracking. In: Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 11–14 (2020)
13. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 2818–2826 (2016) https://doi.org/10.1109/CVPR.2016.308
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE Conf. CVPR **2016**, 770–778 (2016)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Published as a Conference Paper at ICLR (2015)
16. Howard, A.G.: MobileNets: efficient convolutional neural networks for mobile vision applications (2017). https://arxiv.org/abs/1704.04861
17. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. J. Mach. Learn. Res. Proc. Track. **9**, 249–256 (2010)