Evaluation of VGG Networks for Semantic Image Segmentation of Malaysian Meals

N Jamil[1], N A N Redzuan[2], M F Ismail[3], and W A W Ramli[4]
{njamil@gmail.com[1]}

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia[1], Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia[2], Faculty of Computer and Mathematical Sciences, and Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia[3]

Abstract. This paper evaluates VGG-16 and VGG-19 networks in performing semantic image segmentation of Malaysian meals. This is a preliminary investigation of using transfer learning models to recognize food objects in typical Malaysian meals. Most current works of food recognition system calculate the calories and nutritional content of a meal based on the food object recognition, regardless of the portion size. Our final aim is to develop a food recognition system that considers the portion size in calculating the calories and nutritional content. Therefore, semantic segmentation of the food objects in the meal is a very important stage. Our work also initiated the training datasets for Malaysian meals that will be made available to the public. Using a small training dataset and a basic configuration of the VGG network, our results show inconsistent findings of the performance of VGG-16 and VGG-19. These findings will serve as a fundamental guideline to improve the semantic segmentation of food images.

Keywords: food images, Malaysian meals.

1 Introduction

Image segmentation is an important step towards object recognition and information extraction [1]. It was and still is a challenging task to achieve an accurate and meaningful segmentation [2][3]. Semantic segmentation is a process of assigning every pixel to predefined classes and aims to solve structured pixel-wise labelling problems [4][5]. Semantic image segmentation has been widely used in many computer vision applications and remote sensing area. Some examples are content-based image retrieval, foreground/background extraction, face recognition, human pose estimation and scene categorization [5]. In the last decade, most semantic segmentation relied on hand-crafted features [6] and classifiers such as Random Forests [7], Boosting [8] or Support Vector Machines [9]. These hand-crafted features are not

INCITEST 2019, July 18, Bandung, Indonesia

DOI 10.4108/eai.18-7-2019.2287943

robust, affected by the variable lighting conditions, and are mostly complex and costly [10]. The amazing success of deep learning in image classification was brought into the semantic segmentation task. Several neural network architectures for semantic segmentation were introduced such as SegNet [11] or fully convolutional networks [12] [13]. Even though they used different datasets, all of them employed VGG [14] network, which is a very large model designed for multi-class classification. Therefore, we proposed the use of SegNet architecture and compared two VGG networks for the semantic segmentation of Malaysian meals images. This paper is organized as follows: Section 2 presents the related work followed by our proposed methodology in Section 3. In Section 4, we describe the implementation and the results. Finally, the conclusion is deliberated in Section 5.

2 Related work

The use of deep learning to recognize multiple food items in a meal is becoming common in the past few years due to greater awareness of health. In many food recognition systems, the recognition of each meal object is necessary to estimate the calorie count and the nutritional content of a meal. In [15], they employed a two-step deep neural network in which the first step determines the hidden nodes and the edge parameters. While in the second back-propagation step, the base and the weights are adjusted to achieve the desired classification results. They used their own Western food dataset comprising 7,000 images of 30 categories of single-item food as training set, and images containing multiple and mixed food items were used as the test set. Another related work done by [16] used 10 categories of most popular Japanese food images and Convolutional Neural Network (CNN) for the purpose of food detection and recognition through parameter optimization. Therefore, our work focused on collecting our own Malaysian meal datasets and employing SegNet architecture to perform semantic segmentation prior to our future work of developing a food recognition system. The next section describes the SegNet architecture.

1.1 SegNet Architecture

Segnet [11] is a deep fully convolutional neural network architecture for semantic pixel-wise segmentation that is trained using road scene image datasets. Semantic pixel-wise labelling is performed by labelling each pixel of an image to some classes or categories depending on the domain applications. Basically, SegNet architecture is mainly convolutional with encoder-decoder pairs that are used to produce sparse feature maps for classifications of different resolutions. There is no fully connected layer in SegNet. Therefore, the number of parameters is reduced from 134M to 14.7M. The final layer is a soft-max classifier that feeds in feature maps from the final decoder. The basic architecture of SegNet is shown in **Figure 1**.

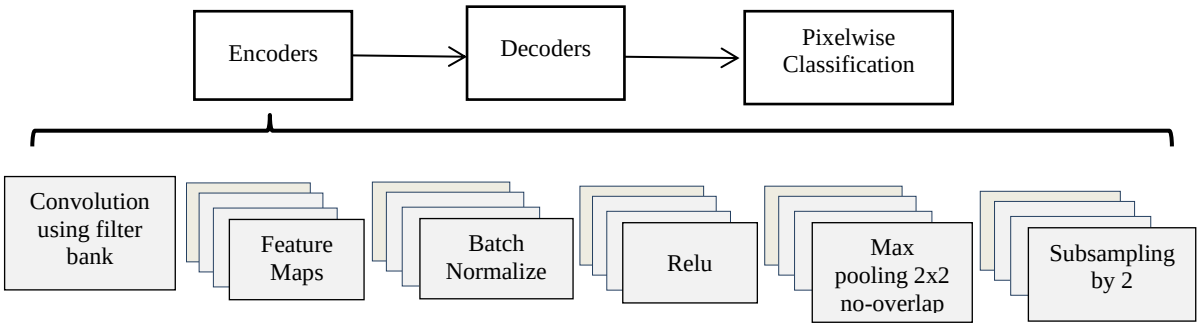

Figure 1. A simplified Segnet architecture [11] consisting of an encoder network and a corresponding decoder network. The final layer is a pixel wise classification layer that assigns pixels to the corresponding class.

The encoder network in SegNet is topologically identical to the convolutional layers in VGG16 [11]. The novelty of Segnet is in the subsampling stage, where the decoders use the max-pooling indices received from the corresponding encoder to perform non-linear upsampling of their input feature maps. These methods have shown increased classification accuracy while reducing the feature map size.

2.2 VGG Network

VGG network [14] is trained using more than a million images of ImageNet enabling it to classify images into 1000 object categories. Even though different layers of VGG network exist, the basic configuration of a VGG network comprises a stack of convolutional layers with 3x3 filters. Max-pooling is performed over a 2 × 2 pixel window, with stride 2. A stack of convolutional layers (which has a different depth in different architectures) is followed by three Fully-Connected (FC) layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the soft-max layer. All hidden layers are equipped with the rectification non-linearity.

3 Methodology

The overall process flow diagram of the semantic segmentation is shown in **Figure 2**:

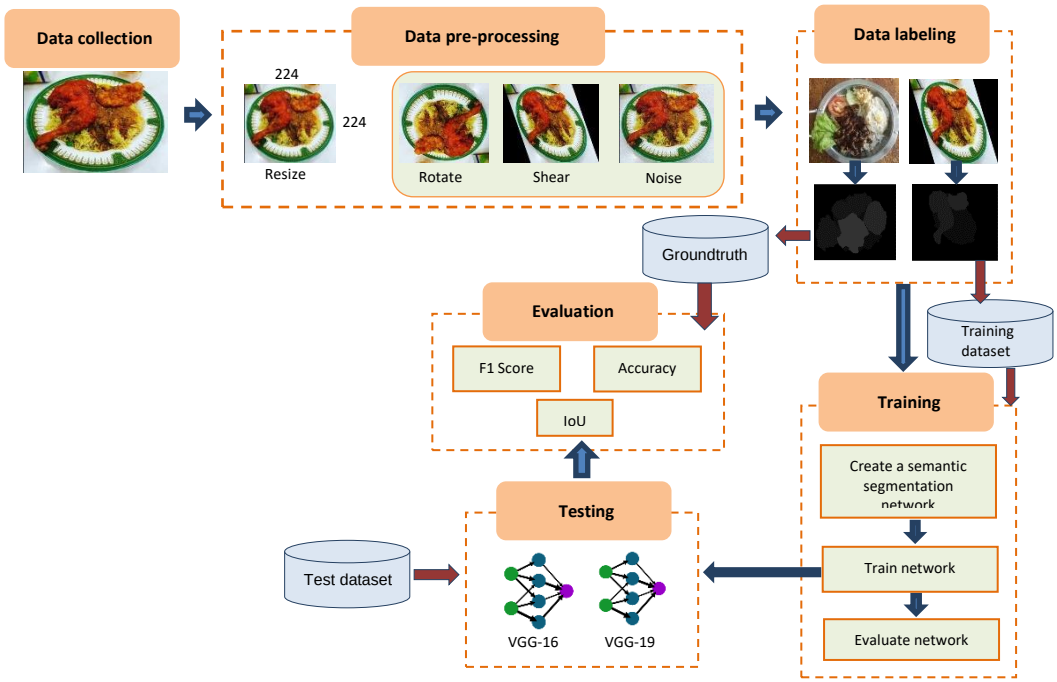

Fig. 2. Process flow diagram of the Semantic Segmentation of Malaysian Food

Data collection

The images used in this paper are food commonly consumed by Malaysians known as *Nasi Campur* comprising rice, poultry, seafood, and vegetables. Some images were captured using a phone camera with 1280 x 720 resolutions under controlled and uncontrolled lighting, while other images with lower dimensions were collected from blogs, social media and Internet sources. A total of 91 images were collected with some examples shown in **Figure 3.**

(a) Controlled lighting (b) Uncontrolled lighting

Fig. 3. Sample meal images taken under (a) controlled illumination and (b) uncontrolled illumination

3.1 Data augmentation

Since the images are gathered from different sources, they were resized to 224 x 224 resolutions to standardize with the requirement of the VGG network. The original images were also augmented to create a larger image dataset for training of the network. Data augmentation can also alleviate memorization of training data and assist the learning model's performance on data from outside the training dataset. In this work, two geometric transformations that are rotation and shearing, and salt & pepper noise are added to the images to create the augmented data. After augmentation, the total image datasets of 364 images were then divided into 292 (80%) training dataset and 72 (20%) test dataset. Figure 4 illustrates examples of several augmented images and table 1 shows the divisions of the augmented images into training and test datasets.

Original Rotation Shearing Noise

Fig. 4. Examples of meal images after different augmentation techniques

Table 1. Division of training and testing datasets

	Training dataset	Test dataset
Original image	73	18
Rotated image	73	18
Sheared image	73	18
Noisy image	73	18

3.2 Data labeling

After data augmentation, data labelling is done to the training datasets to train the VGG network in segmenting the region of interests. In our work, there are 8 categories of region of interests that are rice ('Nasi'), vegetables ('Sayur'), chicken ('Ayam'), meat ('Daging'), fish ('Ikan'), egg ('Telur'), prawns ('Udang') and cuttlefish ('Sotong'). Data labelling is also important to create the ground truth data for testing datasets. In figure 5a, the object egg, meat and vegetables are labelled and figure 5b shows the labelled fish, rice and vegetables objects.

(a) (b)

Fig. 5. (a) Data labelling of egg, meat and vegetable. (b) Data labelling of fish, rice and vegetable and coloured for display purposes

After data labelling, the distribution of the class labels was plotted to understand the overall class balance. As can be seen from figure 6, the classes were not balanced with classes such as 'Udang', 'Daging' and 'Sotong' showing that the number of pixels in these classes is very small. These rare classes may pose a challenge during learning of the VGG network resulting in poor semantic segmentation. Therefore, class balancing was done by using class weighting calculated using inverse frequency weighting. The higher the frequency of a class, the smaller weight is assigned to the class. Table 2 presents the weightage assigned to each class for class balancing.

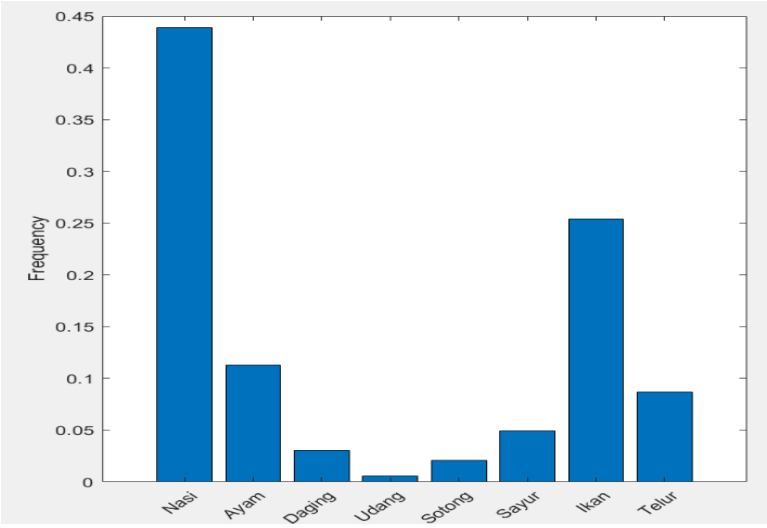

Fig 6. Distribution of the Malaysia food image classes in training dataset

Table 2. Class weightage

Class	Nasi	Ayam	Daging	Udang	Sotong	Sayur	Ikan	Telur
Weightage	0.1838	0.7435	2.4446	11.4039	3.0487	1.5268	0.2859	0.4336

3.3 Training the VGG network

The SegNet architecture trains using a pre-trained VGG network. In this paper, we compare two VGG network that is VGG-16 comprising 41 layers and VGG-19 consisting of 47 layers. Since, the network is used for semantic segmentation, the fully connected layer of VGG network is replaced by the pixelwise classification layer, softmax. The main purpose of training is to reduce (minimize) the loss function's value with respect to the model's parameters. In this study, the base learning rate is set at 0.001. As can be seen in table 3, the loss value of VGG-16 network reduces consistently from the first iteration until the 200th iteration. Similarly, the VGG-19 also shows consistent reduction of the loss function until the 300th iteration. The iterations are stopped once the loss value began to increase. However, the accuracy of the mini-batch ranges from a low 10% to approximately 26% for both VGG networks. Even though the accuracy is not encouraging, the VGG networks behave rather consistent based on the reduction of loss at each iteration. After the networks are trained, they are then used to perform semantic segmentation using the test dataset created earlier.

3.4 Evaluation of semantic segmentation

The performance of semantic segmentation of Malaysia meals was done using a set of evaluation metrics. The dataset metrics allow a high-level overview of VGG network performance. For a more detail understanding of the performance, inspection of per-class metrics was also done. The dataset metrics are described in **Table 4**.

Table 3. Evaluation of training the VGG networks

Epoch	Iteration	Base learning rate	Mini-batch accuracy	Mini-batch loss	Mini-batch accuracy	Mini-batch loss
			VGG-16		VGG-19	
1	1	0.001	10.18%	2.2847	10.81%	2.2952
1	50	0.001	10.51%	2.2028	11.43%	2.2549
2	100	0.001	16.67%	2.0204	11.79%	2.1353
3	150	0.001	20.29%	1.9890	13.32%	2.0339
3	200	0.001	16.19%	1.9869	17.11%	2.0004
4	250	0.001	24.17%	1.8476	19.70%	1.9014
5	300	0.001	25.04%	1.8367	22.42%	1.8937
5	350	0.001	19.72%	1.9216	18.96%	1.9796
6	400	0.001	21.90%	1.9324	22.33%	1.9319
6	438	0.001	26.48%	1.8609	21.99%	1.9769

Table 4. Description of the dataset metrics.

Metric	Description
Accuracy	Calculating the ration of correctly classified pixels for each class
Mean accuracy	The average accuracy of all classes in all images
Global accuracy	A quick estimation of the percentage of correctly classified pixels. Calculates the ratio of correctly classified pixels, regardless of class, to the total number of pixels.
Boundary F-1 score	Measure on how well the predicted boundary of each class aligns with the true
Mean Intersection over union (IoU)	Measures the number of pixels common between the ground truth labels and the predicted labels. It counts the amount of overlap pixel per class
Weighted IoU	Average IoU of each class, weighted by the number of pixels in that class. Used to reduce the impact of errors in the small classes on the aggregate quality score

4 Results and Discussions

Performance evaluations of the trained VGG-16 and VGG-19 networks for semantic segmentation of Malaysia meals were done using the testing datasets. The overall network performances are tabulated in table 5. The results showed that both VGG networks scored global accuracy of approximately 18% and mean accuracy of 38.38% for VGG-16 and 36.87% for VGG-19, respectively. Since the classes of our training dataset are disproportionate, Weighted IoU is used to measure the pixel overlaps. As can be seen, the Weighted IoU results for both networks are close indicating similar performances. Since, an overview performance is an inappropriate evaluation measures when the classes are imbalanced, per-class evaluations were done for fairer evaluations. The results of per-class evaluations are shown in table 6.

Table 5. An overview of the semantic segmentation performance of VGG networks

	Mean accuracy	Global accuracy	MeanIoU	WeightedIoU	MeanBfScore
VGG-16	0.38377	0.18757	0.077887	0.22665	0.094905
VGG-19	0.36868	0.18464	0.076096	0.21979	0.093169

Table 6. Per-class evaluations of the VGG networks for semantic segmentation

Class	Accuracy	IoU	MeanBf Score	Accuracy	IoU	MeanBf Score
	VGG-16			VGG-19		
Nasi	0.67602	0.43392	0.20276	0.64999	0.44498	0.20891
Ayam	0.28072	0.052508	0.093352	0.14027	0.050255	0.10106
Daging	0.18223	0.01475	0.061082	0.092835	0.010913	0.067322
Udang	0.062765	0.0018051	0.055842	0.12129	0.0018053	0.053865
Sotong	0.050531	0.0027385	0.056003	0.080105	0.0.0033761	0.060748
Sayur	0.09968	0.02202	0.076417	0.19199	0.020723	0.071754
Ikan	0.12815	0.066829	0.10458	0.18085	0.076566	0.09969
Telur	0.020432	0.014198	0.062125	0.01982	0.014481	0.058828
Average	0.187566	0.076096	0.08902	0.184644	0.088532	0.090272

As expected, ‘Nasi’ class has the highest accuracy rate of 67.6% for VGG-16 and 64.99% for VGG-19. The average accuracy rates of both networks are also similar at approximately 18%. For VGG-16, the best accuracy rate of ‘Nasi’ class seems to indicate that since this class has the highest pixel distribution compared to other classes, the segmentation’s accuracy is the highest. However, this assumption is not true for ‘Ikan’ class that has the second highest pixel distribution but only achieved an accuracy rate of 12.8% compared to ‘Ayam’ and ‘Daging’ classes at 28% and 18.2%, respectively. The same scenario is found for VGG-19 network. For ‘Nasi’, ‘Ayam’, and ‘Daging’ classes, VGG-16 performed considerably better than VGG-19. However, VGG-19 network scored a higher accuracy rate for ‘Udang’, ‘Ikan’ and ‘Telur’ classes. There is no consistent evidence of which class is better segmented and which network segments the dataset better.

We further investigated the semantic segmentation’s performance using confusion matrix. Table 7 and 8 show the confusion matrix for VGG-16 and VGG-19, respectively. For VGG-16, the pixels of ‘Nasi’ class are mostly assigned to class ‘Sotong’. Meanwhile, VGG-19’s results showed that the pixels of ‘Nasi’ class are assigned to ‘Ayam’ class. On the other hand, the pixels of the ‘Sayur’ class are mostly misclassified as pixels of ‘Telur’ class using VGG-16. However, most of these pixels are wrongly assigned to ‘Sotong’ class for VGG-19. There is no conclusive evidence to state a misclassification of pixels to any specific classes based on the results of table 7 and 8. There is also an inconclusive finding of whether VGG-16 performs better than VGG-19, vice-versa.

Table 7. Confusion matrix of VGG-16

	Nasi	Ayam	Daging	Udang	Sotong	Sayur	Ikan	Telur
Nasi	67.6	10.59	5.386	3.102	2.427	4.046	5.65	1.2
Ayam	20.99	28.07	14.65	6.576	5.208	9.597	12.96	1.951
Daging	11.92	33.05	18.22	6.497	5.26	10.15	13.31	1.591
Udang	23.74	27.19	14.53	6.276	4.739	9.148	12.73	1.648
Sotong	34.36	20.6	10.88	6.496	5.053	8.156	11.67	2.785
Sayur	15.54	29.93	15.4	7.144	6.32	9.968	13.53	2.166
Ikan	20.03	28.61	14.97	6.612	5.369	9.581	12.81	2.022
Telur	18.42	29.11	14.89	6.751	5.265	10.21	13.31	2.043

Table 8. Confusion matrix of VGG-19

	Nasi	Ayam	Daging	Udang	Sotong	Sayur	Ikan	Telur
Nasi	65	7.217	5.622	4.119	3.068	7.321	6.505	1.149
Ayam	21.31	14.03	11.57	9.995	7.176	18.15	15.75	2.027
Daging	20.31	11.31	9.284	12.49	7.143	20.48	17.63	1.358
Udang	19.03	12.9	9.98	12.13	7.485	20.6	16.22	1.648
Sotong	13.05	12.54	10.45	12.82	8.011	22.88	18.87	1.379
Sayur	15.55	14.98	13.11	10.56	7.437	19.2	17	2.172
Ikan	15.34	13.1	10.88	12.16	7.743	21.03	18.08	1.671
Telur	16.72	13.99	11.88	10.93	7.414	19.66	17.42	1.982

4 Conclusion

The aim of this paper is to evaluate the performance of VGG-16 and VGG-19 for semantic segmentation of the Malaysian meals. Based on the results presented in the earlier sections, we find that both VGG networks showed potential in semantic segmentation based on the consistent performance of the loss function. However, there is no conclusive evidence to indicate that one network is better than the other. The accuracy results of the semantic segmentation are also considerably low. These may be caused by many factors that are not within the scope of this paper. Further investigations to fine-tune the network should be done such as the class weightage computations, the base learning rate, batch sizes, the number of layers and regularization

techniques. In terms of training datasets, we should also consider using single object datasets with variations of food cuttings and sizes instead of using multiple objects in a meal.

References

[1] Wei X, Guo Y, Gao X, Yan M and Sun X : A new semantic segmentation model for remote sensing images *IGARSS* p 2–5.(2017).

[2] Ong HT and Ma KK : Semantic image segmentation using oriented pattern analysis. *Proc IEEE 8th International Conference on Information, Communications & Signal Processing* 2011 Dec 13p 1-4. (2011).

[3] Liu T and Stathaki T : Enhanced pedestrian detection using deep learning based semantic image segmentation. *Proc. 2017 22nd International Conference on Digital Signal Processing (DSP)* 2017 Aug 23 p 1-5. (2017).

[4] Marmanis D, Schindler K, Wegner JD, Galliani S, Datcu M and Stilla U : Classification with an edge : Improving semantic image segmentation with boundary detection. *ISPRSJournal of Photogrammetry and Remote Sensing* **135** 158–172. (2018).

[5] Wang LL and Yung NH : Apr 1 Hybrid graphical model for semantic image segmentation.*Journal of Visual Communication and Image Representation* **28**:83-96. (2015).

[6] Chen LC, Papandreou G, Kokkinos I, Murphy K and Yuille AL Apr 1 Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(4):834-48.(2018).

[7] Shotton J, Johnson M and Cipolla R : Jun 23 Semantic texton forests for image categorization and segmentation. *Proc 2008 IEEE Conference on Computer Vision and Pattern Recognition* p 1-8. (2008).

[8] Tu Z and Bai X : Auto-context and its application to high-level vision tasks and 3d brain image segmentation *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(10):1744-57. (2010).

[9] Fulkerson B, Vedaldi A and Soatto S: Class segmentation and object localization with superpixel neighborhoods *Proc. 2009 IEEE 12th International Conference on Computer Vision* p 670-677. (2009).

[10] Kestur R, Meduri A and Narasipura O. : MangoNet : A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard. *Engineering Applications of Artificial Intelligence* **77***:* 59–69. (2018).

[11] Badrinarayanan V, Kendall A and Cipolla R : Segnet: A deep convolutional encoder-decoder architecture for image segmentation *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(12):2481-95.(2017).

[12] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* p 3431-3440, (2015).

[13] Volpi M and Tuia D : Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*.**144**:48-60(2018).

[14] Simonyan K and Zisserman A. : Sep 4 Very deep convolutional networks for large-scale image recognition. (*Preprint* arXiv preprint arXiv:1409.1556), (2014).

[15] Pouladzadeh P and Shirmohammadi S. : Aug 10 Mobile multi-food recognition using deep learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **13**(3s):36, (2017).
[16] Kagaya H, Aizawa K and Ogawa M : Food detection and recognition using convolutional neural network. *Proc. of the 22nd ACM International Conference on Multimedia* 2014 Nov 3 p 1085-1088 (2014).