A Deep Convolutional Neural Network for Food Detection and Recognition

Mohammed A. Subhi
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Selangor, Malaysia
mohd.a.subhi@gmail.com

Sawal Md. Ali
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Selangor, Malaysia
sawal@ukm.edu.my

***Abstract*— In this paper, we propose a new deep convolutional neural network (CNN) configuration to detect and recognize local food images. Various types of food with different color and texture reflect the fact that the food image recognition is considered a challenging task. However, deep learning has been widely used as an efficient image recognition method, and CNN is the contemporary approach for deep learning to be implemented. CNN has been optimized to the tasks of food detection and recognition with few modifications. We present a new dataset of the most consumed local Malaysian food items which was collected from publicly available Internet sources including but not limited to, image search engines. For evaluation of recognition performance, CNN achieved significantly higher accuracy than traditional approaches with manually extracted features. Additionally, it was found out that convolution masks show that the features of food color dominate the features map. For the process of food detection, CNN also exhibited considerably higher accuracy than other conventional methods.**

Keywords*— Convolutional neural network; Deep learning; Food detection; Food recognition

I. INTRODUCTION

Obesity has been known as the excess body fat due to the imbalance of food intake [1]. Proper health monitoring and effective management of the calorie intake is carried out through better reporting of food consumption. Conventionally, it has been achieved by manual self-reporting and observation. Though the method is statistically acceptable, yet it suffers from the issue of underreporting and hence, neither it accurately reveals the actual amount of calories, nor the real behavioral human eating patterns [2], [3].

Recently, the accuracy and effectiveness of food intake reporting systems have improved by applying pattern recognition and image processing techniques to automatically identify and detect food items [4]. In these systems, inclusive nutrition database information is used to produce a daily food intake report for individuals relying on the identification and recognition of food images. The analysis of food images is considered a challenging task due to many factors including the identification of multi food classes within a single plate or the variance of the food texture for the same type. Additionally, the number of overall food classes is not determined yet.

Based on image analysis, a conventional automatic vision-based dietary assessment system involves four fundamental steps starting with food detection, classification of food type, volume or weight assessment, and lastly nutritional information assessment [5]. Recently, the development in image processing and object detection, machine learning approaches, and specifically deep learning and its implementation of convolutional neural networks (CNN), has enhanced the image identification and recognition accuracy.

In the last few years, CNN has been widely used in food recognition applications, and it achieved better performance than the conventional machine learning methods. The authors in [6], have adapted the structure of AlexNet model as presented by the authors in [7] and constructed a deep convolutional neural net (CNN) using images acquired from Food-101 dataset. This approach achieved a top-1 accuracy of 56.4%. The authors in [8], have also deployed CNN for food recognition and identification, their work included 10 food classes and the outcomes exhibited the exceptional performance of CNNs in contrast with other conventional methods by achieving a 73.7% detection accuracy. The authors in [9] implemented CNN as the sole feature extractor applied on the UEC-FOOD-100 dataset [10] and achieved an accuracy of 72.3%, 100 classes of Japanese food were implemented from the dataset. The features map is a combination of the output of a pre-trained AlexNet model as well as manually crafted features. The authors in [11] have retrained the AlexNet model with two datasets. Their approach achieved a top-1 accuracy of 78.8% for UEC-FOOD-100 dataset and 67.6% for UEC-FOOD-256. The authors in [12] have proposed a new food dataset, the images were acquired from public Internet social media sources (Instagram) and presented a comparable CNN classification performance with other datasets. The authors achieved outstanding outcomes with a high accuracy of 99.1%. CNN have proven its superior performance when compared to previous works that use conventional machine learning approaches.

Research efforts have been dedicated towards many aspects of a food recognition system, yet a comprehensive solution to accurate food classification and recognition is still needed, bearing in mind the wide selection of food items and mixed food items in many dishes. Hence, it is particularly challenging to correctly identify every food item, due to the fact that many food items are indistinguishable in terms of shape or color and

978-1-5386-2471-5 ©2018 IEEE

some food characteristics are even hard to be recognized by simple examination. For example, lamb and beef meats look very much similar and it's visually challenging to differentiate. Additionally, in a real environment, mixed or prepared food dishes make the detection issue unfeasible to solve. Hence, we state that it would be a good practice to generally classify and identify food items to attempt to automatically approximate its dietary information. This may offer additional insights on daily food consumption habits and nutritional information.

In this paper, we propose a new structure of an enhanced food detection and recognition model using CNN. After a food image is acquired, the first challenge is to detect if a food item exists in the image contents. Then food category recognition is performed. A total of (5800) images distributed in (11) local Malaysian food categories are selected. Images were collected from Internet sources including Google image search with acceptable quality.

II. Convolutional Neural Networks

Convolutional neural networks (CNN) provide exceptional model architecture for image classification and recognition [5]. It is composed of a sequence of filters applied to the raw input image to robustly extract and learn image features. These features will be used by the model for classification purposes.

CNN's are typically a configuration of three types of layers. Convolutional layers, apply a number of convolution filters with specific size (e.g. 3x3) to the input image. For each section of the image, a set of mathematical operations is applied to produce a single value in the output. Consecutively, these layers usually apply an activation function to these outputs to introduce nonlinearities into the model. A commonly used activation function is Rectified Linear Units (ReLU) [6]. The second type of layers are pooling layers, which in turn downsample the resulting image produced by the convolutional layers to reduce the size of the feature map for a faster processing time. A widely used pooling algorithm is max pooling. It extracts sub-sections of the feature map (e.g., 2x2-pixels), finds their maximum value, and drops all the other values. Fully connected layers, perform classification on the extracted features after the downsampling process by the pooling layers. In this type of layer, every node is connected to every node in the previous layer.

The major advantage of using CNNs is that the process of feature extraction is done automatically using a stack of convolutional modules which incorporate a convolutional layer followed by a pooling layer. All the layer blocks between the first module and the last follows the same pattern, except the last layers block which is followed by one or more fully connected layers that performs the classification. The final fully connected layer (or called dense layer) has a single node for every trained class in the model (prediction classes) along with a softmax activation function that generates a value between 0–1 for every class (the sum of all classes values is equal to 1). An interpretation of these softmax values is how likely it is that the contents of the image lies under each target class. Conventional CNN contains several convolution and pooling layers, along with a fully connected layer to deliver the final result of the process. In image classification, each unit of the final layer indicates the class probability. To properly design a CNN, an adjustment to its hyperparameters has to be made including the number of hidden layers, kernel size, and the activation functions. In this paper, we propose a basic and an optimized version of some of these parameters.

III. Proposed CNN Architecture

All input images were resized to 220x220x3 before feeding them to the CNN model. CNN extracts feature from the images in an automatic fashion starting from larger amounts of pixels to smaller details like color and edges. To demonstrate the functionality of a basic CNN model, a three layers model as shown in Fig.1. The first layer, a convolutional layer (CONV1) that is composed of 32* 5x5 filters applied to the original image resulting in a block of 220x220x32 images followed by a max pooling layer of [2x2] with stride 2. This will resize the image block height and width to half maintaining the number of images in the block to 110x110x32. Followed by another convolutional layer of 64* [5x5] filters, namely (Conv2) which results in a block of 110x110x64 that will be fed into another max pooling layer [2x2] which again will resize the height and width to half [55x55x64]. The next layers are fully connected layers, by flatting the features vector into 193600 features that go into a ReLu layer of 1024 and finally the fully connected layer of 11 classes as outputs.

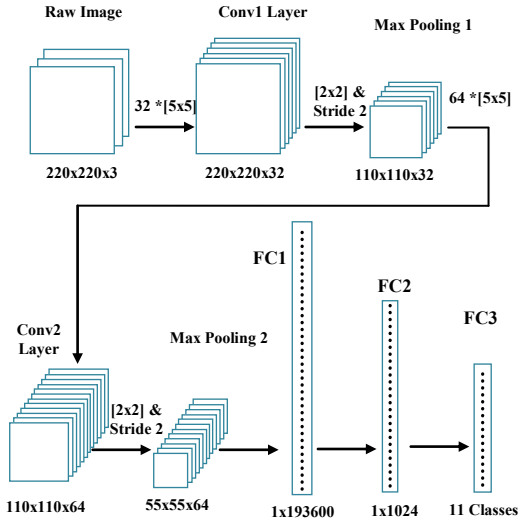

Fig. 1. Basic CNN Layer Structure

For more accurate model, we propose the use of a more complex model which consists of multi convolutional layers blocks before the final fully connected layer. Based on the structure suggested in [13], to increase the accuracy of model classification a deeper network is required. However, as a tradeoff depth affects the performance and the time required to obtain some results. We implemented a 24 layers model with 21 convolutional layers and 3 fully connected layers as shown in Fig. 2. In each convolutional layer stride is fixed to 1 pixel while the spatial padding of convolutional layer input is performed in such a way that it preserves the dimensions after convolution, for example the padding for 3 × 3 convolutional layers is 1 while for 5 x 5 layers the padding is 2. Spatial

pooling is applied through five max-pooling layers which follow some of the convolutional layers but not all with a 2 × 2 pixel window and stride 2.

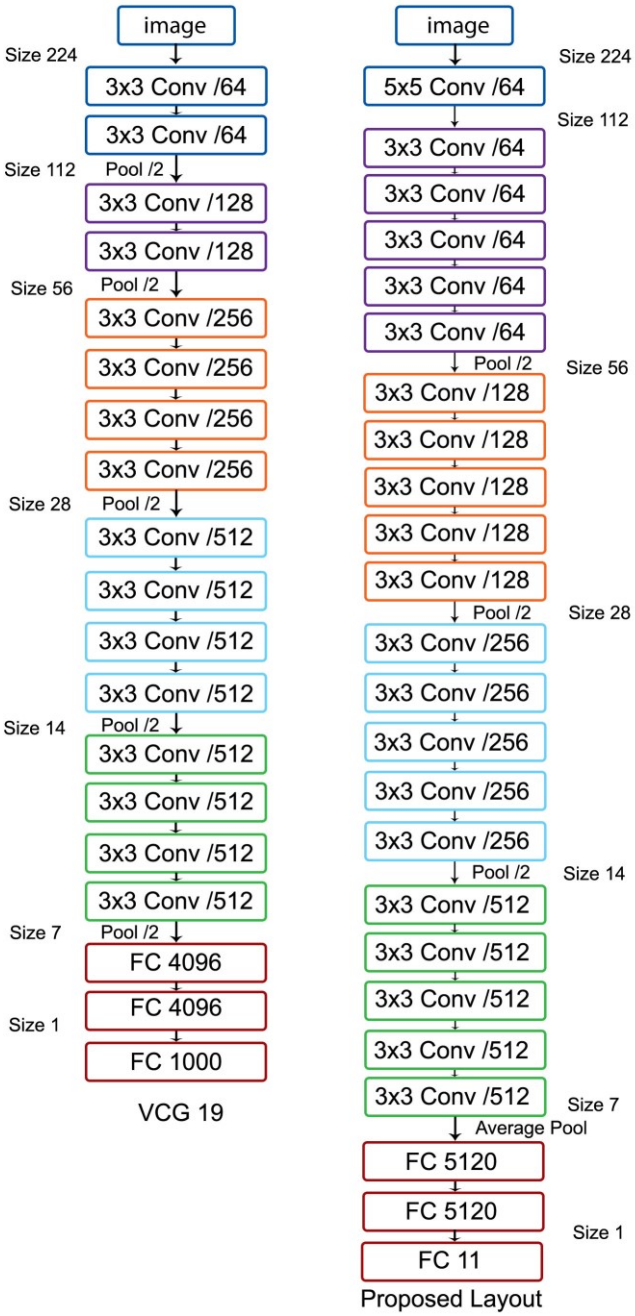

Fig. 2. Proposed CNN structure versus VCG 19

Since food classification is concerned with all the details that a food item may have such as the shape, color and texture, larger kernel size at the beginning of the layers ensures that shape features are maintained in the learning process, while smaller kernel size preserves the fine details of the food objects.

IV. Food Dataset

We propose a new dataset that includes local Malaysian food items/ dishes. 3300 food images are distributed over 11 categories (Nasi Lemak, Nasi Goreng, Fried Noodles, Curry Puffs, Cucumber, Tomatoes, Chili Pepper, White Rice, Fried Chicken, Boiled Eggs, and Fried Eggs) with 300 images for each category as shown in Fig 3. The images were collected from publicly available Internet sources including but not limited to, image search engines. Food image were collected with respect to variations to pose, rotation, color and shape complexity to improve the identification accuracy when more than single food item is included in the image.

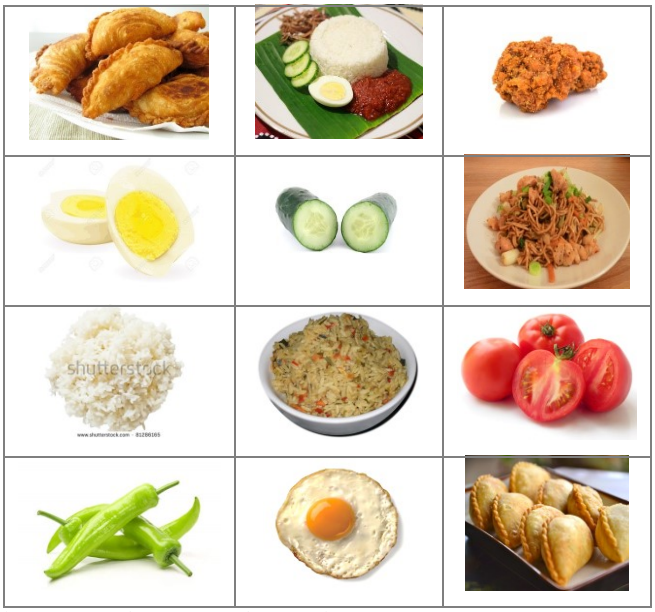

Fig. 3. Proposed Dataset Food Categories

V. Conclusions

In this paper, we have implemented a basic CNN approach to detect and recognize food items. We have presented a new dataset for local Malaysian food which contains (11) food categories with (5800) images. Two datasets were used for performance evaluation of our proposed approach, Food-101 were utilized for food/non-food classification and our proposed local Malaysian food dataset, which was used for the categorization of food items.

Additionally, we have implemented very deep convolutional networks (24 weight layers) for food image classification. Large kernel size at the beginning of the layers ensures that shape features are maintained in the learning process. It was shown that it is beneficial for classification accuracy to have this depth. The results confirm the significance of network depth in training visual representations.

Acknowledgment

This project is supported by Universiti Kebangsaan Malaysia under the research grant (DIP-2016-008).

REFERENCES

[1] W. H. Dietz *et al.*, "Management of obesity: improvement of health-care training and systems for prevention and care," *The Lancet*, vol. 385, no. 9986, pp. 2521–2533, 2015.

[2] Karthick K, S. Ramasamy, and P. Ramanathan, "Need for development of dietary assessment device for free living environment-a survey," 2015, pp. 1–7.

[3] W. Wu and J. Yang, "Fast food recognition from videos of eating for calorie estimation," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 1210–1213.

[4] L. Yang, J. Yang, N. Zheng, and H. Cheng, "Layered object categorization," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1–4.

[5] G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato, "Retrieval and classification of food images," *Comput. Biol. Med.*, vol. 77, pp. 23–39, 2016.

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

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1085–1088.

[9] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 2014, pp. 589–593.

[10] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, 2012, pp. 25–30.

[11] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *European Conference on Computer Vision*, 2014, pp. 3–17.

[12] H. Kagaya and K. Aizawa, "Highly accurate food/non-food image classification based on a deep convolutional neural network," in *International Conference on Image Analysis and Processing*, 2015, pp. 350–357.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv Prepr. ArXiv14091556*, 2014.