Comparison of Convolutional Neural Network Models for Food Image Classification

Gözde ÖZSERT YİĞİT
Computer Engineering Department
Gaziantep University
Gaziantep, Turkey
gozsertyigit@gantep.edu.tr

Buse Melis ÖZYILDIRIM
Computer Engineering Department
Çukurova University
Adana, Turkey
mozyildirim@cu.edu.tr

***Abstract*— According to some estimates of World Health Organization (WHO) , in 2014, more than 1.9 billion adults aged 18 years and older were overweight. Overall, about 13% of the world's adult population (11% of men and 15% of women) were obese. 39% of adults aged 18 years and over (38% of men and 40% of women) were overweight. The worldwide prevalence of obesity more than doubled between 1980 and 2014. The purpose of this study is to design a convolutional neural network model and provide a food dataset collection to distinguish the nutrition groups which people take in daily life. For this aim, both two pretrained models Alexnet and Caffenet were finetuned and a similar structure was trained with dataset. Food images were generated from Food-11, FooDD, Food100 datasets and web archives. According to the test results, finetuned models provided better results than trained structure as expected. However, trained model can be improved by using more training examples and can be used as specific structure for classification of nutrition groups.**

Keywords*— *Deep learning, Convolutional Neural Network, Food Classification, Nutrition Categorization

I. INTRODUCTION

Nutrition is not to satisfy hunger, to calm the feeling of hunger or to eat and drink everything we want. Nutrition; is a behavior which is necessary to realize consciously to assure the contribution of nutritional elements necessary for the body to protect and develop the health and to increase the quality of life in sufficient amounts and at the appropriate moments. The human requires about 50 nutritional elements for his life. When these nutritional elements are not taken of sufficiently, poor nutrition occurs. Each of these elements is determined by how much it should be taken daily, for healthy grow and development of human and to live healthy and productive for a long time. When any of these nutrition is not consumed, or consumed insufficiently, growing and improving are prevented and health of human is damaged. If this nutrition are excessively consumed, as this excess nutrition taken is stored in the body in the form of fats (lipids), this is fatal for health. This situation is called Unbalanced Nutrition. In the prevention of unbalanced nutrition, it is very important to acquire the nutritional educational consciousness of a healthy diet. For a sufficient and well-balanced nutrition, it is necessary to consume nutrients in proportions recommended among 4 main groups of nutrients mentioned. The first one is the group of milk and this group of nutriments have to be consumed by all the age groups, in particular adult women, children and adolescents. The second group is group of meat-egg dry legumin. The third one is the group of vegetable and fruits and the last one is bread and group of cereals [1]. When consuming of these groups are insufficient and excessive, some diseases can be occurred like obesity and diabetics.

Nowadays obesity is among the major sanitary problems of the developed countries as well as developing countries. In general, obesity is the excess of the physical weight compared to physical height and that the fat mass of the body is more than non-fat mass of the body. In the daily life, the individuals (pregnant, the child who sucks, baby, child going to the school, young, old, labor, sportsman, people with cardiovascular diseases, diabetes, high blood pressure, disorder of the respiratory system, etc.) need daily energy changing according to the age, gender, profession, genetic specificities and health state.

To be able to live a healthy life, it is necessary to protect the balance between the input and the consumption of energy. The adipose tissue constitutes 15-18 % of the physical weight of the adult man and 20-25 % of the physical weight of the adult woman. When this proportion exceeds the rate of 25 % in men and the rate of 30% in women. As it is understood obesity is considered as a disease having fatal repercussions on the life quality and expectancy which appears due to the fact that the energetic input provided by the nutriments (calorie) is more than the energy consumed and with the fat accumulation of the excess of energy in the body (more than 20 %) [1].

Diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period [2]. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. As of 2015, an estimated 415 million people had diabetes worldwide [3], making up about 90% of the cases. This represents 8.3% of the adult population with equal rates in both women and men [4].

In this study, some food groups are classified by using deep Convolutional Neural Network (CNN) algorithm. The dataset is obtained from an archive in web area [5] and it includes real-world food image (Food-100). Food ID given in dataset are used as class labels and class number is determined as 100.

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

There exist some studies in literature [6, 7, 8, 9, 10]. These paper is differentiated from the studies in the literature with its classifier structure. Instead of using pretrained structures, in the study number of images in dataset is increased by cropping and transformations and new structure is created.

In [6], new algorithms were proposed to analyze the food images captured by mobile devices (e.g., smartphone). The key technique innovation in this paper is the deep learning-based food image recognition algorithms. The proposed algorithms are based on Convolutional Neural Network (CNN). The experimental results of applying the proposed approach to two real-world datasets (UEC-256 and Food-101) have demonstrated the effectiveness of the solution.

In [7], the effectiveness of deep convolutional neural network (DCNN) is examined for food photo recognition task. To tackle the problem, best combination of DCNN-related techniques are sought such as pre-training with the large-scale ImageNet data, fine-tuning and activation features extracted from the pre-trained DCNN. The fine-tuned DCNN which was pre-trained with 2000 categories in the ImageNet including 1000 food-related categories was the best method, which achieved 78.77% as the top-1 accuracy for UECFOOD100 and 67.57% for UEC-FOOD256, both of which were the best results so far. Also, the food classifier employing the best combination of the DCNN techniques were applied to Twitter photo data. The great improvements on food photo mining in terms of both the number of food photos and accuracy were achieved. In addition to its high classification accuracy, DCNN was found very suitable for large-scale image data, since it takes only 0.03 seconds to classify one food photo with GPU.

Another study proposed to classify this dataset, used conventional and deep features together and classified with linear support vector machine (SVM). To extract deep features, pretrained Overfeat model was utilized in [8]. According to the reported results, this approach provides 72.26% success.

In [9], HOG and Fisher Vector coding of color features were used in SVM to classify dataset. In that study, the main was real time food recognition system on a smartphone. According to the experiment results, 79.2% classification rate was obtained.

In [10], experiments on food/non-food classification and food recognition were reported by using a GoogLeNet model based on deep convolutional neural network. The experiments were conducted on two image datasets created by their own, where the images were collected from existing image datasets, social media, and imaging devices such as smart phone and wearable cameras. Experimental results show a high accuracy of 99.2% on the food/non-food classification and 83.6% on the food category recognition.

The rest of the paper is structured as follows: Section 2 gives details about the dataset and proposed deep learning-based approach for this study. Section 3 presents results and discussions. The last section, Section 5, concludes the paper.

II. Material and Method

A. Material

The dataset that has been used for this paper is a combination of databases provided from archives in the web area. The dataset "Food-11" was created by researchers who proposed [10]. It consists 16643 images from well-known databases Food-101, UECFOOD-100 and UEC-FOOD-256 and from social media, then grouped into 11 categories. These categories were determined in accordance with the major types of food that people consume in daily life. Moreover, in this study to increase the number of data, "FooDD" dataset also was utilized [11]. In "FooDD" dataset, images were taken with different camera types and under different conditions and images were categorized into their type such as "Banana, Bread, Grape, etc."

In this study, images and categories provided in Food-11 dataset were used and in accordance with these categories images from FooDD dataset were labelled. For instance, images in the Grape category in FooDD were labelled as Vegetables/Fruits. Nevertheless, in some categories such as Dairy products, number of images was too small, to increase it additional images were also collected from web and Food100 dataset and by changing image properties new images were obtained. Hence, for each category in this dataset there exists at least 1500 images.

B. Method (Deep Convolutional Neural Network)

Deep convolution neural networks which was proposed in 1998 by Lecun has become an effective tool in solving pattern recognition problems [12,13]. Traditional convolution neural networks are occurred of several convolution and pooling layer coming in succession. Deep convolution neural network includes a multi-layered structure after several convolution and pooling layer too. Generally, as an activation function at the last layer of this multi-layer structure, logistic classification function is used in given (1).

$$Y_p[j] = e^{\frac{(Y_{p-1}[j])}{(\Sigma_{j=1}^{N} Y_{p-1}[j])}} \quad (1)$$

In Fig. 1, one of the well-known architecture Alexnet is shown [14]. Alexnet consists of five convolution layers with kernel sizes 11, 5, 3, 3, and 3, respectively. While the first and second convolution layers are followed by ReLU, normalization, and pooling layers, the other convolution layers are followed by ReLU and pooling layers except third and fourth convolution layers. There are not pooling layers between third and fourth convolution layers and between fourth and fifth convolution layers. Maximum operator is utilized in pooling layers. After these layers three fully connected layers followed by ReLU and dropout layers exist. At the last layer, softmax with loss function is used [14].

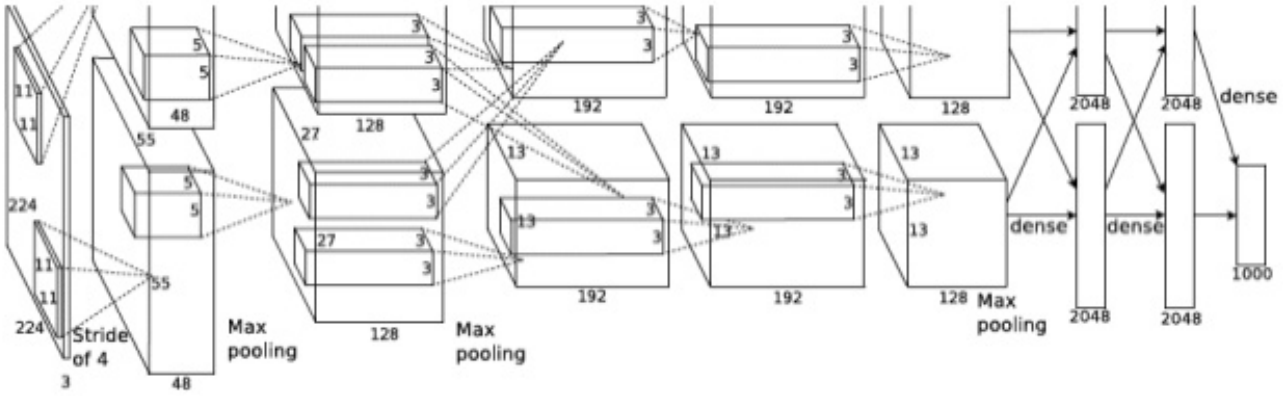

Fig. 1. Alexnet Structure

While convolutional layers provide obtaining common local features on training images, pooling layer creates a map as clustering of these features. This map provides features to take place in a different place than place in training images of features. Thus, independence of location is ensured. When the number of layers are increased, it can be observed that convolutional layer provides to obtain the features of independent transformation [15,16]. Convolution layer performs shifting operation of a kernel in a specific size or a specific filter on the input. Inputs are kernels are multiplied, bias value is added and is passed through an activation function such as the tangent hyperbolic sigmoid, linear rectification and logarithmic function. Generally, different features are obtained by applying more than filter on an image simultaneously. In (2), convolution process at pth layer is shown: Kp shows number of kernel at pth layer, n and m are indices to kernels exist in pth and (p-1)th layers. x and y shows height and weight of each kernel at pth layer, respectively, ws shows size of applied kernel, r and q are indices to each dimension of ws and w shows the applied kernel (filter), y_p shows the result of pth layer and y_{p-1} is the output values of previous layer [10].

$$
\begin{aligned}
&for\ (n = 0; n < Kp\ ; n++)\\
&for\ (m = 0; m < Kp\text{-}1; m++)\\
&\quad for(y = 0; y < \text{weight}; y++)\\
&\quad for(x = 0; x < \text{height}; x++)\\
&\quad\quad for(r = 0; r < ws; r++)\\
&\quad\quad for\ (q = 0; q < ws\ ; q++)\\
&y_p(n; x, y) \mathrel{+}= y_{p-1}\,(m, x + r, y + q) * w\,(m\ , n; r, q);
\end{aligned}
\tag{2}
$$

After convolution layer, it is passed through pooling layer. At this layer, shifting process is done on output of convolution layer by using specific kernel size. In every shifting, pooling is done by selecting maximum or average from feature group obtained from convolution. In 3, the sample of pooling layer is on t. layer where r and q are indices to each dimension of pooling kernel, s is the size of one dimension of kernel, x and y are indices to current nodes of feature map, y_t is the output value of tth layer, third line in the equation shows the maximum operation process [15].

$$
\begin{aligned}
&for(r = 0; r < s; r++)\\
&for\ (q = 0; q < s\ ; q++)\\
&y_t(x, y) = \max\ (\ y_t\ (x, y), y_{t-1}\ (x * s + r, y\ * s + q)\ )\ ;
\end{aligned}
\tag{3}
$$

In deep convolution neural network, training is provided by back propagation learning and usually tilt drop method is used for the error minimization [15-16]. The biggest advantage of deep convolution neural networks is providing deep impressions that are not affected from transformation about data. Being a large structure is disadvantage in terms of time complexity. Since the parallels programmable structure of the network permits application on GPU, this disadvantage can be eliminated by benefiting from today's graphic processor. Also, the parallel program feature also provides the implementation on the GPU as well as other enhancements.

III. Proposed Method

In this study, foods were classified by using deep convolutional neural networks. Data used in this study, were collected from several datasets and web area. The dataset includes 16950 training data from 11 categories. Since dataset collected from different sources, firstly, irrelevant areas were removed such as thumb in images from FooDD dataset. Two scaled sample images are shown in Fig. 2.

Fig. 2. Scaled sample images

As in frequently used convolutional neural network models (Alexnet, Googlenet), transformations were applied to images. Proposed structure was determined by try-fail approach. Different structures were tested on dataset. The structure described below performed best.

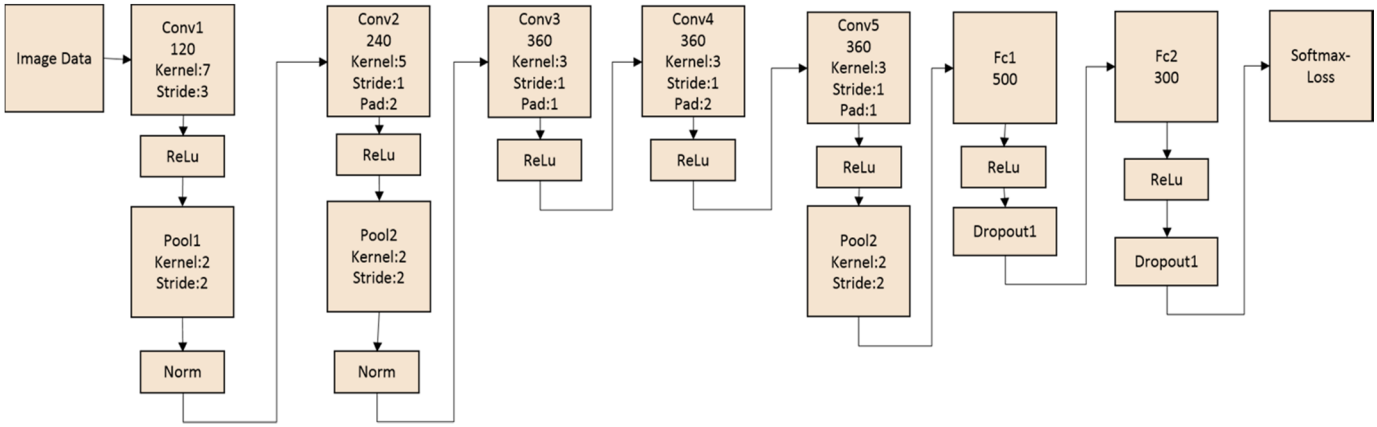

Fig. 3. Proposed Structure

Proposed structure consists of 5 convolutional and pooling layers, 2 normalization layers and 4 fully connected layers. Maximum selection method is used at pooling layer. ReLU function is utilized after each convolution. For convolution layers, kernel sizes were chosen as 7x7, 5x5, 5x5, 3x3 and 3x3 from shallow layers to deep layers. The features obtained from convolutional layers are transmitted to pooling layer. Stride for pooling layer is chosen as 2 and kernel sizes 2x2. Thus, the total number of neurons in the output of each pooling layer is reduced by 1/4. In normalization layer, local response normalization was utilized. For hidden layers of fully connected part, neuron sizes were determined as 500 and 300, respectively. Output layer neuron number is 11 to represent each category in our dataset. Softmax classifier is utilized in the structure. Proposed structure is shown in Fig. 3.

IV. RESULTS

Proposed architecture was implemented on collected dataset with Caffe framework which can run on GPU. Since weight and bias values were generated randomly, tests were carried out 5 times to eliminate random effect. Training parameters for proposed structure are given in Table 1. For testing, Food-11 evaluation images were used, it includes 3347 images.

TABLE I. TRAINING PARAMETERS

Parameters	**Values**
Learning Rate	0.3
Momentum	0.7
Maximum Iteration	40000

In this study, generated dataset was tested on pretrained models and the proposed structure. Since proposed structure is similar to Alexnet and Caffenet, these two models preferred for comparison. Caffenet and Alexnet were finetuned with images in our dataset instead of retraining. For fine tuning, the same parameter values given in Table 1 were used. Table 2 shows test results obtained from these structures.

TABLE II. TEST RESULTS

Model	**Accuracy (%)**
Caffenet [17]	80.51
Alexnet [18]	82.07
Proposed Structure	70.12

According to the test results, fine-tuned models provided better results as expected due to their generalization ability. However, proposed structure may be improved by using more food images for each category, provide better results than general purpose models.

V. CONCLUSION

In this study, development of application specific convolutional neural network model was aimed. With this aim, it is thought that models will be trained faster and will not require excessive amount of data to perform accurate. To achieve this aim a convolutional neural network structure proposed and food image database was collected. Database includes images from some well-known food image databases and some images from web archives and transformation on these. While proposed structure was trained, well-known models Alexnet and Caffenet also finetuned with collected dataset. Test results show that finetuned models provide better results than proposed one as expected. However, performance of trained structure was also acceptable and if the number of images is increased proposed method may be improved.

Consequently, this study may be considered as first step for developing pretrained application specific convolutional neural network model. The idea behind this aim is to be able to train models with less data but provide better results.

REFERENCES

[1] The department of obesity, diabets and metabolic diseases, the ministry health of turkey, public health instituion. Access 10 March 2017.

[2] "About diabetes". World Health Organization. Archived from the original on 31 March 2014. Retrieved 4 April 2014.

[3] "Update 2015". IDF. International Diabetes Federation. p. 13. Retrieved 21 Mar 2016.

[4] Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, Shibuya K, Salomon JA, Abdalla S, Aboyans V, et al., "Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010.". Lancet. 380 (9859): 2163–96, Dec 15, 2012.

[5] "UEC FOOD 100: 100 kind of food dataset", Last modification, 14 March, 2017, http://foodcam.mobi/dataset.html.

[6] Liu, C., Cao, Y., Luo, Y., Chen, G., Vokkarane, V. and Ma, Y., "DeepFood: Deep Learning-based Food Image Recognition for Computer-aided Dietary Assessment", ICOST 2016, International Conference on Smart Homes and Health Telematics, vol. 9677, pp. 37-48, 21 May, 2016.

[7] Yanai, K. and Kawano Y., "Food image recognition using deep convolutional network with pre-training and fine-tuning", IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2015.

[8] Kawano, Y. and Yanai, K., "Food Image Recognition with Deep Convolutional Features, ACM UbiComp Workshop on Cooking and Eating Activities, 2014.

[9] Kawano, Y. and Yanai, K., "FoodCam: A real time food recognition system on a smartphone", Multimedia Tools and Applications, 74,14, pp. 5263-5287, 2015.

[10] Single, A., Yuan, L. and Ebrahimi, T., "Food/Non-food Image Classification and Food Categorization using Pre-Trained GoogLeNet Model", Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management (MADIMA'2016), pp. 3-11, 16 October, 2016.

[11] P. Pouladzadeh, A. Yassine, and S. Shirmohammadi, "FooDD: Food Detection Dataset for Calorie Measurement Using Food Images", in New Trends in Image Analysis and Processing - ICIAP 2015 Workshops, V. Murino, E. Puppo, D. Sona, M. Cristani, and C. Sansone, Lecture Notes in Computer Science, Springer, Volume 9281, 2015, ISBN: 978-3-319-23221-8, pp 441-448. DOI: 10.1007/978-3-319-23222-5_54

[12] J. Kim ve V. Pavlovic, "Discovering Characteristic Landmarks on Ancient Coins using Convolutional Networks", CoRR abs/1506.09174, 2015.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient based learning applied to document recognition". In Proceedings of the IEEE, pp. 2278– 2324, 1998.

[14] Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[15] Ginzburg, B., "Deep Learning Summer Workshop", Ver. 06. http://courses.cs.tau.ac.il/Caffe_workshop/Bootcamp/pdf_lectures/Lecture%202%20Caffe%20- %20getting%20started.pdf, 2014.

[16] Nogueira, R. F., "Fingerprint Liveness Detection Using Convolutional Neural Networks", IEEE Transactions on Information Forensics and Security, vol. 11(6), pp. 1206-1213, 2016.

[17] https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet, (Accessed 11.03.2017)

[18] https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet (Accessed 11.03.2017)