PAPER • **OPEN ACCESS**

Implementation of Convolutional Neural Networks to Recognize Images of Common Indonesian Food

To cite this article: I Putu Agus Eka Darma Udayana *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **846** 012023

IOP Conf. Series: Materials Science and Engineering **846** (2020) 012023 doi:10.1088/1757-899X/846/1/012023

Implementation of Convolutional Neural Networks to Recognize Images of Common Indonesian Food

[1]I Putu Agus Eka Darma Udayana, [2]Made Sudarma, [3]Putu Gede Surya Cipta Nugraha

[1,3]Department of Informatic Engineering STMIK STIKOM Indonesia, [2]Department of Electrical and Computer Engineering, Faculty of Engineering Udayana University

[1]agus.ekadarma@stiki-indonesia.ac.id, [2]sudarma@ee.unud.ac.id, [3]surya.cipta@stiki-indonesia.ac.id

Abstract. One of the applications of computer vision in the popular culture is food recognition which is popularized in the internet with a "hotdog and not hotdog" problem. Food recognition is also useful in many popular lifestyle apps such as calorie counter app or any diet related app. In this paper is proposed a CNN aided technique for recognizing food that is common in Indonesia. The technique is consist of 3 main phases, one is pre-processing normalize the data one is model formation and training which is known as the common binary classifier template, boosted with pooling and evaluated by cross-entropy technique in this paper the model used is the pre-trained model to test the testing data afterward, which is show a promising result with a relatively short training time. The experiments focused on how CNN can be used as a component to recognize food so that in the future it can be used to develop better calorie counter applications. In this experiment 10,000 data were used for training and 50 tests for each food category with a total of 500 food image data used for testing with the best accuracy reaching 88% for one of the categories.

1. Introduction

Apply human intelligence to computer systems is a very popular study conducted by researchers, for example research on computer vision. In the implementation of problems encountered in computer vision research is the classification of general objects, for example mimicking the ability of humans to interpret information from pieces of images. To achieve this people usually apply feature engineering to overcome this problem, but technically this approach will not give the best results. Because the scope of the problem is limited, they can be addressed or bound only to certain dataset sets without optimization for general solutions. This problem occurs because of different points of view in terms of scale, brightness, object deformation and so on. Academics have struggled with this problem for a long time; this problem usually falls into the scope of case studies of artificial neural networks, solutions inspired by the neural network of human ideas carried out until the era of deep learning. At that time in 1989, Y Lecun et al. managed to make a classification of postal image codes using a sophisticated neural bait model called the Convolution Neural Network (CNN) [1]. But hardware constraints came that made this research not continued, until early 2009 when Jurgen came up with the idea of the Repeated Neural Network (RNN) which helped him recognize human handwriting. Since then, along with the development of GPU based computational neural networks, in mid 2012 convolutional neural network experiments were able to rival humans in certain data sets [2]. In another study in 2016 by Suartika et al. managed to classify animals categorized as poultry using the CNN

method with an accuracy of 20% - 50%. In conclusion, it is said to be able to increase the value of system accuracy needed for optimal training data [3]. Deep learning is a hot topic in research, because of its ability to model, a series of very complex problems such as pictures and sounds. One method of registering for learning in image processing/image recognition is Convolutional Neural Network (CNN) which comes not without logical reasons, it turns out the main idea of CNN stems from efforts to imitate networks in the changing human visual cortex to make CNN effective for processing large image data. But because this method also has its own weaknesses, namely CNN training time which is considered very time consuming. But once again the GPU developed into a savior with the presence of the Versatile Graphics processing unit and the emergence of the limitations of cloud computing is not as big a problem as before. [3] Based on this background, in this study the authors developed a system that is able to recognize the image of popular foods in Indonesia using the CNN method. The use of the CNN method in this study is motivated by the ability of the CNN method in public objects and needs to be deeply identified. Besides the problems that arose in the past, namely the long training time can be solved by the existence of cloud computing. To test the system developed, the authors tested the accuracy of food image recognition. By using accuracy testing, it will be known how optimal the CNN method and training data are to get the accuracy of food image recognition [4].

2. Fundamental Theory

2.1. Convolutional Neural Network

Convolutional neural network or CNN is a level up of a very popular artificial neural network technique called multilayer perceptron which is designed to tackle 2 dimensional data problem like images, CNN itself is falls into a category of Deep neural network because of the added depth of the network. the CNN is popular among researcher who tackle image recognition problem, the problem which MLP fails to solve effectively due to the fact that MLP doesn't store the information of image spatial data and assume that every pixel is an independent feature which limits the ability of MLP to solve the image recognition problem.

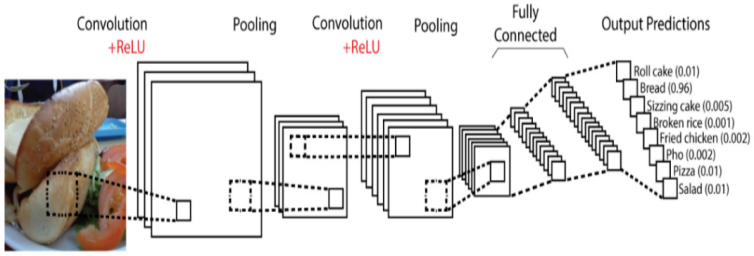

Figure 1. CNN Method Process

CNN is a term coined orignally by a researcher Kunishiko Fukushima from NHK broadcasting science research laboratory Kinuta Setagaya in Tokyo Japan with the name of Neocognitron [5]. The concept which Kunishiko coined later was perfected by Lichen a researcher from AT&T Bell Laboratories in New Jersey USA, Lecunn implement a full-pledged CNN model he named Le net that he trained to recognize a written digit [1]. Flies a couple of years a man named Alex Krizhevsky with his implementation of CNN is able to win an ImageNet Large Scale Visual Recognition Challenge 2012. A contest that later propels the popularity of deep learning to the mainstream scene as a powerful solution. more powerful than SVM (Support Vector Machine) a popular method at that time.

2.2. Architech of Convolutional Neural Network

Artificial Neural Network consist of various layers and a bunch of neurons for each layer. Both of that thing would not be possible if we stick to the old procedural programming paradigm as it would be different one case to another [6]. In the case of a multilayer perceptron, a network without a hidden layer could map a linear equation whatever the variable is, compared to a network with one or two hidden layers that could map almost of the equation with a simple data. But with the case with more complex data multilayer perceptron have a limit, the limit of MLP would be problem that needs more than 3 layers although there is a method to decided a number of neuron on each layer to get an optimal

result many scholars would not recommend a uses of more than 2 layer cause the overfitting problem and the diminishing return of the network capabilities to solve problem. With the revolution of deep learning, it can backup an MLP in its most weak are treating complex data. There is a function that can transform input data to data that would be easier to feed into a neural network, which is the main revolutionary aspect of deep learning. The layer that transforms the data before its trained to be a classifier class. That thing alone makes an effective neural network above 3 layers possible. But because of the flexibility of this transforming layer which is mainly used for feature extraction, there is no basic rule on behalf of how many layers that we should use in a deep learning network.

2.2.1. Convolution Layer. Convolution layer does all the convolution operation in the network that is coming from the previous layer that layer is the main process that is a fundamental of a convolutional neural network a convolution meaning a recurrent process to apply a function to an output of another function in the case of image processing it means to apply a kernel to every possible offset (as illustrated down below) [7].

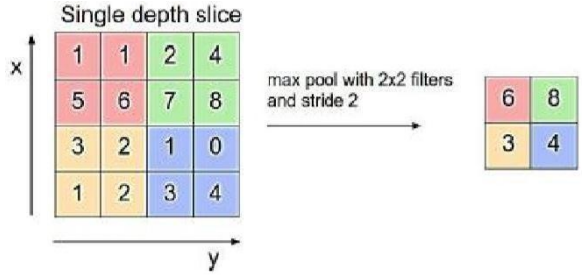

Figure 2. Convolution Operation

2.2.2. Subsampling Layer. Subsampling is a process of size reduction of an image. In the realm of image processing subsampling could be also to get more invariance of position and feature of an image in almost of all CNN. Subsampling is most common technique is a max pooling max pooling is to divide every output of the convolution layer to various smaller grid and then get the maximum of each grid to get the reduced format of the input image (which is illustrated in the image down below) [8].

2.2.3. Fully Connection Layer. Fully connected layer is a layer that we usually find in a multi layer perceptron application that one layer that has a main goal of transforming a dimension data so that the data could be classified as a linear data. Each neuron of a convolution layer needs to be transformed to a one dimensional data at first before it can be faded to a full they connected layer [2]. But it should be implemented in the last layer of the network to avoid a data loss which makes this layer is not as often used as the other kind of layer.

2.2.4. Activation Layer. Activation function refer to a function that transform data to N+1 dimension that makes the data can be cutted in the hyperlane space that makes the classification of the data point possible in the realm of CNN the activation function that are commonly used is a sigmoid function[2]. A sigmoid function transform an input of x to an either 1 of 2 value 0 and 1 with a form of distribution like presented in figure 3.

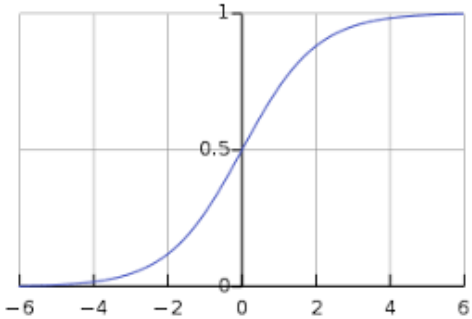

Figure 3. Sigmoid Function

A sigmoid function is basically best to classify a binary problem because its output tendency to be between zero and and one.

3. Methodology

3.1. Image Preprocessing

In the process of preprocessing food images are processed by wrapping and cropping. The wrapping process aims to determine the edge of the input image, after determining the maximum edge of the input image the cropping process will then be carried out so that the system can focus on recognizing the main objects of the food image.

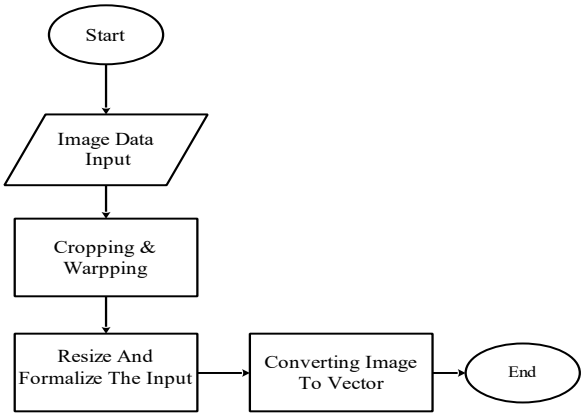

Figure 4. Flowcahart Preprocessing CNN

Figure 5. Ilustration Wrapping And Cropping

In accordance with figure 5, the image processing process begins by changing the size of any food image size to 128 x 128 pixel and then changing it to a gray scale for the training process.

3.2. Training Phase

The training process is used to obtain high accuracy from food image classification. This training phase consists of the feedforward process and the backpropagation process. To start the feedforward process, it takes the number and size of layers to be formed, the size of the subsampling and vector images. The feedforward process in vector images will go through a process of convolution and Max Pooling to reduce the size of the image and multiply the neurons.

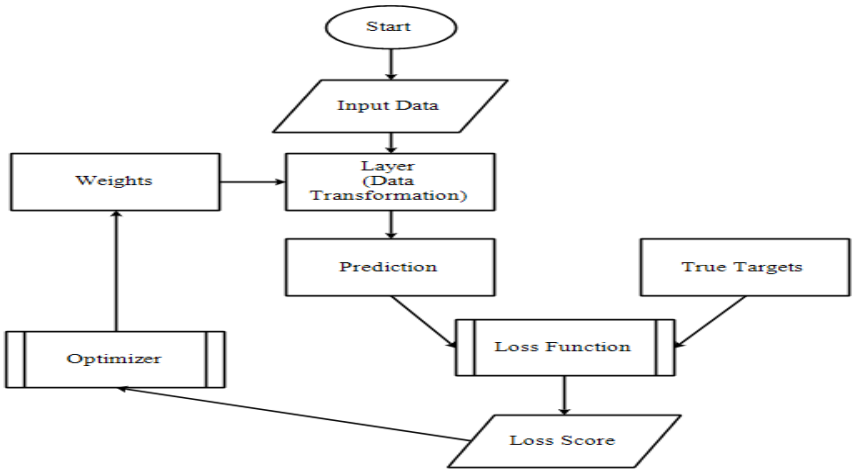

Figure 6. Flowchar Training Process CNN

20/20 [==========] - 41s 2s/step - loss: 0.0975 - acc: 0.9750 - val_loss:

0.0428 - val_acc: 0.9920
Epoch 796/800
20/20 [==============================] - 40s 2s/step - loss: 0.0760 - acc: 0.9740 - val_loss: 0.0531 - val_acc: 0.9837
Epoch 797/800
20/20 [==============================] - 40s 2s/step - loss: 0.0894 - acc: 0.9650 - val_loss: 0.0604 - val_acc: 0.9820
Epoch 798/800
20/20 [==============================] - 41s 2s/step - loss: 0.1327 - acc: 0.9600 - val_loss: 0.0736 - val_acc: 0.9740
Epoch 799/800
20/20 [==============================] - 40s 2s/step - loss: 0.0848 - acc: 0.9710 - val_loss: 0.0920 - val_acc: 0.9620
Epoch 800/800
20/20 [==============================] - 41s 2s/step - loss: 0.1137 - acc: 0.9630 - val_loss: 0.1151 - val_acc: 0.9660

Figure 7. Contoh Hasil Proses Training

In the training process, the author uses data as much as 10,000 food image data. The following is an example of the results of the training conducted by the author. As shown in Figure 8 the training process in this study is set up to 800 epochs with a loss of 0.1137 and 0.9630 acc using 1,000 food image data for each class. The accuracy of 0.9630 shows the system has been able to classify food types with an accuracy of 96.30% with misclassification of 11.37% in the case study data training. To test the system's ability to classify images it is necessary to test the system using food images outside of the training data.

4. Result

4.1. Implementation

The implementation of Convolutional Neural Networks (CNN) method in the classification of food image types in this study uses the python programming language. In its application, the CNN method uses several steps to predict the process of food image recognition. In the initial stage, food images will be included in the Convolution process with sizes Conv2D (32, (3, 3)), Conv2D (64, (3, 3)), Conv2D (128, (3, 3)) and input form (128, 128, 3). To reduce the dimensions of each food input image feature but still retain the most important information from the image, the Pooling process is used. In this study the pooling method used is Max Pooling with the pool_size = (2, 2) movement, followed by the 'Flattening' and 'Fully Connected' process to classify the features obtained in the previous process into various classes.

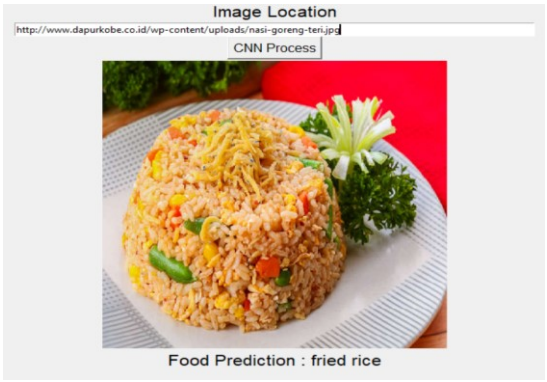

Figure 8. Food Image Recognition Results (Fried Rice)

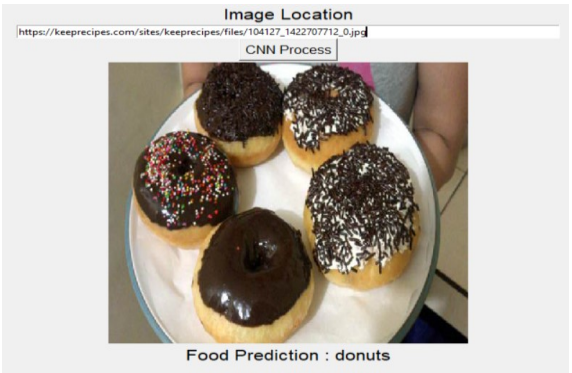

Figure 9. Food Image Recognition Results (Donuts)

In this step, to be able to classify food images the user must enter food images in the system and then the system tries to classify the types of food inputs from the user. Figure 8 and 9 are simulations of food type classification using the CNN method which results in the classification of fried rice and donuts.

4.2. Testing

In the research that has been carried out the application of the Convolutional Neural Networks (CNN) method has been able to classify food images that are inputted to the system. In its implementation to be able to classify food image types quite well, with the prediction of system errors in classifying food images by 11.37% in the image of training data. In testing the accuracy of the system this time the author will focus testing on 10 food image classes using 50 new food images for each test class outside the training data images with the provisions of using sigmoid activation for each type of food. The following are the results of tests that have been done by the author:

Table 1. Result of Accuracy Testing

Food Name	True		False		Total of Image Testing
	Amount of Image	Percentage (%)	Amount of Image	Percentage (%)	
Pizza	42	84	8	16	50
Steak	41	82	9	18	50
Sushi	44	88	6	12	50
Omelette	41	82	9	18	50
Noodles	49	98	1	2	50
Ice Cream	44	88	6	12	50
Hamburger	41	82	9	18	50
Fried Rice	48	96	2	4	50
Donat	48	96	2	4	50
Ceke	42	84	8	16	50
Min	41	82	1	2	50
Max	49	98	9	18	50
Average	44	88	6	12	50

Based on the test results of the classification of food images in table 1 using 50 food image test data outside the food image used in the training data with sigmoid activation produces good test results. In the test results, the system can classify food with an average error is 12% and average acc is 88%.

5. Conclution

Preprocessing methods and classification methods using Convolutional Neural Networks are reliable enough to determine the truth of the classification of object images. This is proven by the accuracy of 88%. The difference in input image size does not significantly affect the results of the system accuracy, because the input image has previously been resized in the preprocessing process. Based on the results of research conducted by the CNN method is very reliable in dealing with changes in input parameters. The more optimal the use of training data, the higher the accuracy of the system will result.

Acknowledgment

The author would like to thank the research assistant team and the ISETcon editorial team who provided input for writing this paper.

References

[1] LeCun Y 1990 *Handwritten Digit Recognition with a Back Propagation Network.*
[2] Coates A, Lee H and Ng A Y 2011 *An Analysis of Single-Layer Networks in Unsupervised Feature Learning* Qiche Gongcheng/Automotive Engineering Vol. **24** Issue 2.
[3] Suartika E P I P, Wijaya A W and Soelaiman R 2016 *Klasifikasi Citra Menggunakan Convolutional Neural Network (Cnn) pada Caltech 101* Jurnal Teknik ITS Vol **5** Issue 1.
[4] Udayana, I P A E D, Sudarma M, Kumara I N S 2017 *Balinese Latin Text Becomes Aksara Bali Using Rule Base Method* International Journal of Research in IT, Management and Engineering Vol **07**, Issue 5.
[5] Karpathy A *CS231n Convolutional Neural Network for Visual Recognition* [Stanford University: cs231m.github.io].
[6] Zufar M and Setiyono B 2016 *Convolutional Neural Networks untuk Pengenalan Wajah Secara Real - Time* J. Sains Dan Seni Its Vol. **5** Issue 2.
[7] Abroyan N 2017 *Convolutional And Recurrent Neural Networks For Real-Time Data Classification* Seventh Int. Conf. Innov. Comput. Technol., no. Intech, pp. 42–45.
[8] Nurfita R and Ariyanto G 2018 *Implementasi Deep Learning Berbasis Tensorflow Untuk Pengenalan Sidik Jari* Emitor: Jurnal Teknik Elektro Vol. **18** Issue 01.