Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 1, April 2019, pp. 333~339
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i1.pp333-339

Comparing bags of features, conventional convolutional neural network and alexnet for fruit recognition

Nik Noor Akmal Abdul Hamid, Rabiatul Adawiya Razali, Zaidah Ibrahim
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia

Article Info

Article history:

Received Jun 17, 2018
Revised Sep 30, 2018
Accepted Dec 6, 2018

Keywords:

Alexnet
Bag of features
CNN
Fruit recognition

ABSTRACT

This paper presents a comparative study between Bag of Features (BoF), Conventional Convolutional Neural Network (CNN) and Alexnet for fruit recognition. Automatic fruit recognition can minimize human intervention in their fruit harvesting operations, operation time and harvesting cost. On the other hand, this task is very challenging because of the similarities in shapes, colours and textures among various types of fruits. Thus, a robust technique that can produce good result is necessary. Due to the outstanding performance of deep learning like CNN and its pre-trained models like AlexNet in image recognition, this paper investigates the accuracy of conventional CNN, and Alexnet in recognizing thirty different types of fruits from a publicly available dataset. Besides that, the recognition performance of BoF is also examined since it is one of the machine learning techniques that achieves good result in object recognition. The experimental results indicate that all of these three techniques produce excellent recognition accuracy. Furthermore, conventional CNN achieves the fastest recognition result compared to BoF, and Alexnet.

Corresponding Author:

Nik Noor Akmal Abdul Hamid,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
Shah Alam, Selangor, Malaysia.
Email: niknoorakmal1994@gmail.com

1. INTRODUCTION

Fruit recognition is useful for automatic fruit harvesting that can reduce or minimize human intervention in their fruit harvesting operations and also the operation time and harvesting cost. Fruit recognition system plays an important role in automatically detecting and inspecting the fruits for harvesting within the fruit images. The implementation of fruit recognition application gives great value of products to the consumers [1]. Fruit recognition application is also useful for fruit disease detection and recognition. The detection and identification of fruit is based on human's naked eyes which is time consuming and costly. Besides, it can facilitate the control of fruit diseases as the disease can be avoided by appropriate sprinkling of pesticides through automatic fruit recognition process.

The performance of fruit recognition [1], speech recognition[2], visual object recognition [3], celebrity face recognition [3] and many other domains like genomics and drug discovery [3] has dramatically improves with the use of deep learning. Deep learning is a class of machine learning algorithms that uses multiple layers that contain nonlinear processing units. One of the techniques under deep learning is Convolutional Neural Networks (CNN) [4]. CNN provides successful results in areas of image recognition and classification. The input is an image used for recognition, and during convolutional process, the output of the image became activation maps. Convolutional layer acts as a filter towards the input in terms of sizes and padding for feature extraction. Pooling layer is operating as a reducer of the feature maps. At the end, the output layer acts as fully connected layer and perform the object classification.

Alexnet, a pre-trained CNN model, has produced very good results for the past few past years [5]. AlexNet is the winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It is designed by the SuperVision group, which consists of Geoffrey Hinton, Alex Krizhevsky, and Ilya Sutskever [6]. This model shows big impacts on image recognition and classification tasks as it produces outstanding performance. AlexNet achieved the top 5 errors from 26% to 15.3% in ILSVRC [6]. The network has more filters per layer with stacked convolutional layers consisting of 11x11, 5x5,3x3 convolutions, max pooling, dropout, data augmentation, ReLU activations, and SGD with momentum for face recognition [6][7]. ReLU activation is attached after every convolutional layer. AlexNet was trained for six days simultaneously on two Nvidia GeForce GTX 580 GPUs which is the reason why their network is split into two pipelines [8].

Bag of words (BoW) [9] has been used for document classification. Bag of Features (BoF) was introduced first by [10] for video retrieval followed by [11] for image categorization that inspired from the original text representation model. An image is represented as an unordered collection of visual words. BoF gives an extremely compact description of images as they are represented as histograms of local descriptors. The main idea is to obtain visual words (features) by quantizing the local descriptors of images in the dataset based on a visual vocabulary. The algorithm takes as an input the training data description and gives as an output a set of clusters. Each cluster is represented by one visual word. The image is now represented as a bag of visual words and a histogram can be built with a dimension equal to the visual vocabulary size, each bin will contain the visual word's frequency with respect to the image [12].

The architecture of a pre-trained CNN model like AlexNet is fixed while we can design our own architecture for a conventional CNN model. When the conventional CNN model goes deeper in their convolution architecture, it can reach a lower identification error rate compared to the human's eyes. A conventional CNN is able to give a great solution in extracting the hierarchical representation of input data which it remains unchanged to conversion and scales [13]. The conventional CNN model produces great results for object recognition applications, thus it is suitable to examine the fruit classification problem. However, in computer vision, the fruit classification task provides challenges in image recognition because of the similar shapes, colors and textures among the various fruits. Thus, the main objective of this research is to investigate the recognition accuracy performance of BoF compared to conventional CNN and pre-trained CNN model which is Alexnet in recognizing fruit based on color images and grayscale images.

2. RESEARCH METHOD

The experiments for this research have been conducted using Matlab2018a using MacBook Pro with 512GB storage, the processor is 3.1GHz Intel Core i5 and memory is 8GB RAM. The Fruit dataset was obtained from ResearchGate [14]. This dataset contains 30 classes of fruits which are Apple Braeburn, Apple Golden 1, Apple Golden 2, Apple Golden 3, Apple Red 1,Apple Red 2,Apple Red 3, Apple Red Delicious, Apple Red Yellow, Apple Granny Smith, Apricot, Avocado, Avocado Ripe, Banana, Banana Red, Cactus Fruit, Cantaloupe 1, Cantaloupe 2, Carambola, Cherry 1, Cherry 2, Cherry Rainier, Clementine, Cocos, Dates, Granadilla, Grape Pink, Grape White, Grape White 2 and Grapefruit Pink. The dataset consists of 960 training images and 240 validation images where each class has exactly 40 images. The images were in various views for each class. The size of the images for each class is 100 by 100 pixels but all of the images were resized to 224x224 for this experiment. Figure 1 shows sample images from Fruit dataset for thirty classes.

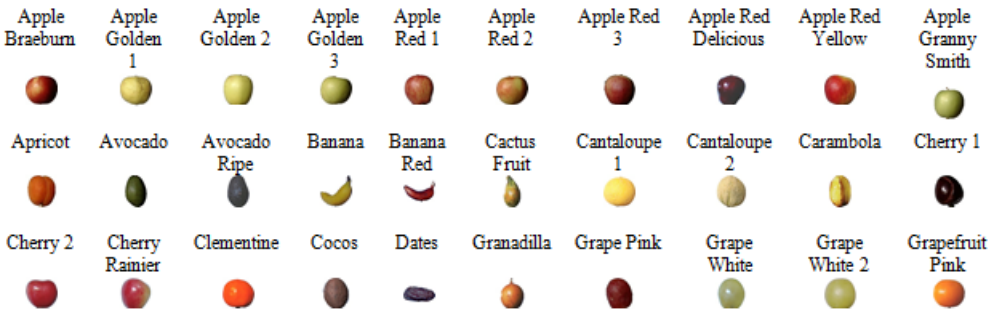

Figure 1. Sample Pictures from Fruit Dataset [14]

2.1. Bag of Features BoF

Inspired from the original text representation model, BoF was introduced for image categorization that was represented as an unordered collection of visual words [15]. As they are represented as histograms of local descriptors, BoF gives an extremely compact description of images. A local descriptor is used in image categorization and object recognition tasks and also to match similar object instances. Many methods for feature description can be employed. Thus, in this work, we target the result based on visual words accuracy. Activities to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results of search can be performed by using machine learning techniques such as BoF [3]. In order to obtain a BoF descriptor, we need to extract features from the image. The feature used is Speeded Up Robust Features (SURF). SURF descriptor is equal to common image transformations which are image rotation, scale changes, illumination and small changes in viewpoint. In addition, SURF is able to compute distinctive descriptors quickly [16]. The classifier used to classify the SURF is Support Vector Machine (SVM). SVM can be used for multiple kind of object recognition like fruit recognition, brain wave recognition and image classification of remote sensing [17]. A multiclass SVM was used to accommodate a multiclass problems but SVM actually was developed for binary classification [17]. Figure 2 shows the illustration of process for bag of features.

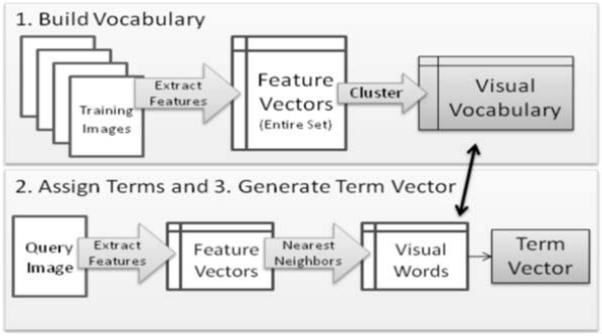

Figure 2. Illustration Process of Bag of Features [18]

2.2. Conventional Convolutional Neural Network (CNN)

The architecture of a conventional CNN consists of three layers which are convolve layer, pooling layer and Rectified Linear unit (ReLu) which is also known as a structured series of layers [1]. The conventional CNN's role is to track data similar with the conventional feedforward neural network. Each image is submitted through the layers until a loss function is achieved at the top layer [5]. The extraction of features from an image is performed by using filters and image patches that stride over the input image in the convolve layer. On the other hand, ReLu layer replaces all negative pixel values in the feature map with zero. In order to reduce the dimensionality, pooling layer is applied that allows the feature map to be down-sampled. Max pooling layer computes the maximum local of feature map. Then, neighboring pooling takes input from the feature maps that are shifted by more than one rows or columns. Figure 3 shows the layer of a conventional CNN [19].

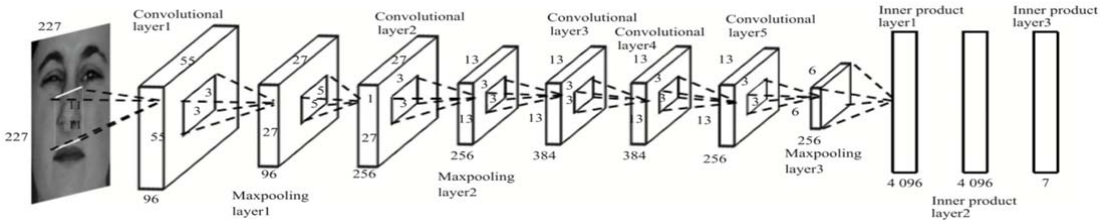

Figure 3. The layer of a conventional CNN [19].

2.3. Alexnet

The common pre-trained CNN model investigated in this paper is AlexNet that is the winner of the ILSVRC in 2012 [1][5]. AlexNet has more filter layers with stacked of convolution layers compared to the conventional CNN architecture where it is designed with deeper architecture. For this research, the fully-connected layers are fine-tuned to classify 30 different categories since the dataset consists of 30 different fruits. An illustration of the layer of AlexNet is shown in Figure 4.

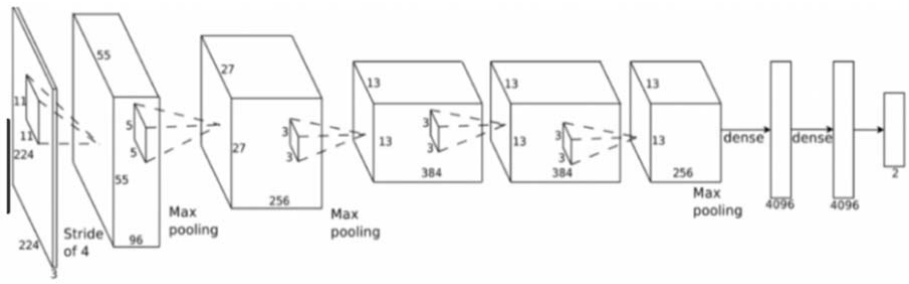

Figure 4. The layers of Alexnet [2]

3. RESULTS AND ANALYSIS

3.1. Bag of Features

The size of image in the input layer is 224x224x3 pixels for color images and 224x224x1 pixel for grayscale images. Speed up Robust Features (SURF) is extracted in BoF where it detects the best scale invariance. For this model, RGB images and grayscale images were tested and the result shows that the total processing time for grayscale image is faster than RGB images which is due to the less number of pixel representation but accuracy 1 is obtained by RGB images and not grayscale images. This is because the conversion process from colour image to grayscale image eliminates some data that may be useful in object recognition. Table 1 shows the different results produced by BoF for color and grayscale images.

Table 1. Accuracy performance of BoF

Input	Image Input Size	Accuracy	Total Time
RGB Image	224x224x3	1	9 min 3s
Grayscale Image	224x224x1	0.98	5 min 47s

3.2. Conventional Convolutional Neural Network (CNN)

For conventional CNN, the dataset is tested with image size 224x224x3. In CNN, there are two main layers that play important roles in analysing the dataset which are convolve layer and maxpooling layer. Experiments on different values for both of these layers are performed to determine the best accuracy and the results are shown in Table 2. Based on Table 2, Fruits dataset is tested twelve times to see the accuracy of the recognition result based on different convolve layer and learning rate. For RGB images, a single convolve layer with (3,16) and learning rate 0.001 achieve accuracy of 1 and the total processing time is 3 minutes and 10 seconds. For double convolve layers with (5,20) and (3,20), the accuracy is 0.9967 with total processing time of 6 minutes and 5 seconds. Meanwhile for grayscale image, a single convolve layer with (5,20) and learning rate 0.001 shows the result of accuracy is 0.9933 and the total processing time for the experiment is 2 minutes and 18 seconds. For double convolve layer for grayscale image with (5,20) and (3,20) and learning rate 0.0001, the total processing time is 5 minutes and 58 seconds.

Table 2. Accuracy performance of conventional CNN

Input	No of layers	Convolve Layer	Stride	Epoch, Learning Rate	Accuracy	Total Time
RGB Image	Single Layer	3,16	3	5, 0.0001	1	4 min 27s
		5,20	2	5, 0.0001	1	6 min 08s
		3,16	3	5, 0.001	1	3 min 10s
		5,20	2	5, 0.001	1	5 min 35s
	Double Layer	5,20 3,20	3 3	5,0.0001	0.9967	6 min 5s
		9,40 3,20	3 3	5,0.0001	1	13 min 6s
Grayscale Image	Single Layer	3,16	3	5, 0.0001	1	4 min 27s
		5,20	2	5, 0.0001	1	3 min 05s
		3,16	3	5, 0.001	0.9.33	2 min 47s
		5,20	2	5, 0.001	0.9933	2 min 18s
	Double Layer	5,20 3,20	3 2	5,0.0001	1	5 min 58s
		9,40 3,20	3 2	5,0.0001	1	11 min 28s

3.3. Alexnet

In order to complete this experiment, the images are resized to 224x224x3 pixels. For the experiment with Alexnet, only color images are tested. Based on Table 3. Fruits dataset was tested three times to investigate the effect of different learning rates to the accuracy. It shows that accuracy 1 is obtained with 0.0001 learning rate and the total processing time to complete the experiment is 22 minutes and 2 seconds.

Table 3. Accuracy performance of Alexnet

No of Test	Image Input Size	Learning Rate	Accuracy	Total Time
1	224x224x3	0.0001	1	22 min 2s
2	224x224x3	0.001	0.5708	22 min 3s
3	224x224x3	0.0005	0.9542	23 min 02s

Table 4 shows the summary of fruit recognition performance using BoF, conventional CNN and Alexnet. By referring to Table 4, we can see that all the three models produce great accuracy which is 1 except for conventional CNN with grayscale image which is 0.99. The total time for Alexnet is the longest compared to the other two models due to the number of layers that it has which is more than the conventional CNN.

Table 4. Fruit recognition performance of BoF, Conventional CNN and Alexnet

Mode	Machine Learning		Deep Learning		
	Bag of Features		Conventional CNN		Alexnet
	RGB	Grayscale	RGB	Grayscale	RGB
Input Size	224x224x3	224x224x1	224x224x3	224x224x1	224x224x3
Accuracy	1	1	1	0.99	1
Total Time	9 min 3s	5 min 47s	3 min 10s	2 min 18s	22 min 2s

4. CONCLUSION

This paper has presented the different accuracy performance of BoF, conventional CNN and Alexnet for fruit recognition based on Fruit dataset. We analyze the performance of fruit recognition based on colour and grayscale images. Based on the results of the experiments, we can see that BoF with SURF and SVM still produce excellent results as CNN. Even though the overall training and testing time of BoF is longer compared to conventional CNN but it is still faster than AlexNet. This shows the robustness of BoF in

recognizing fruits with different shapes, colour and texture. For future work, we will do more experiments on other datasets with other machine learning and deep learning techniques.

ACKNOWLEDGEMENTS

The authors would like to thank Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, for sponsoring this research.

REFERENCES

[1] N. A. Muhammad, A. A. Nasir, Z. Ibrahim, and N. Sabri, "Evaluation of CNN , Alexnet and GoogleNet for Fruit Recognition," *IJEECS*, vol. 12, no. 2, pp. 468–475, 2018.
[2] R. D. Safiyah, Z. A. Rahim, S. Syafiq, Z. Ibrahim, and N. Sabri, "Performance Evaluation for Vision-Based Vehicle Classification Using Convolutional Neural Network," *IJET*, vol. 7, pp. 86–90, 2018.
[3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[4] J. Ubbens, M. Cieslak, P. Prusinkiewicz, and I. Stavness, "The use of plant models in deep learning: An application to leaf counting in rosette plants," *Plant Methods*, vol. 14, no. 1, pp. 1–10, 2018.
[5] N. Sabri, Z. Abdul Aziz, Z. Ibrahim, M.a.R. Akmal Rasydan and A.H. Abd Ghani, "Comparing Convolution Neural Network Models for Leaf Recognition,", *International Journal of Engineering and Technology (IJET)*, vol. 7, pp. 141–144, 2018.
[6] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems," 2017.
[7] N. Ateqah, B. Mat, N. Hidayah, B. Abd, and Z. Ibrahim, "Celebrity Face Recognition using Deep Learning,", *IJEECS*, vol. 12, no. 2, pp. 476–481, 2018.
[8] I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks.", *In Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012
[9] K. S. George and S. Joseph, "Text Classification by Augmenting Bag of Words (bow) Representation with Co-Occurrence Feature", *IOSR Journal of Computer Engineering (IOSR-JCE),* Vol 16, No 1, pp 34-38., 2014.
[10] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," *Towar. Categ. Object Recognit.*, no. Iccv, pp. 1470–1477, 2003.
[11] C. S. Venegas-Barrera and J. Manjarrez, "Visual Categorization with Bags of Keypoints," *Rev. Mex. Biodivers.*, vol. 82, no. 1, pp. 179–191, 2011.
[12] C. Hiba, Z. Hamid, and A. Omar, "Bag of Features Model Using the New Approaches : A Comprehensive Study", *International Journal Advances Computer Science and Applications*, vol. 1, no. 7, pp. 226–234, 2016.
[13] Z. Ibrahim, N. Sabri and D. Isa, "Palm Oil Fresh Fruit Bunch Ripeness Grading Recognition Using Convolutional Neural Network", *Journal of Telecommunication, Electronic & Computer Engineering*, Vol 9, No. 3-2, 2018, pp. 109-113.
[14] Mureșan, Horea and Oltean, Mihai. "Fruit recognition from images using deep learning". *Acta Universitatis Sapientiae, Informatica.* 10. 26-42. 10.2478/ausi-2018-0002, 2018.
[15] E. Okafor, P. Pawara, F. Karaaba, and O. Surinta, "Comparative Study Between Deep Learning and Bag of Visual Words for Wild-Animal Recognition.", *In IEEE Symposium Series for Computational Intelligence (SSCI)*, pp. 1-8, 2016.
[16] S. Hwang, "Bag-of-visual-words approach based on SURF features to polyp detection in wireless capsule endoscopy videos," *Proc. 2011 Int. Conf. Image Process. Comput. Vision, Pattern Recognition, IPCV* 2011, vol. 2, no. i, pp. 941–944, 2011.
[17] Z. Ibrahim, N. Sabri, and N. N. A. Mangshor, "Leaf Recognition using Texture Features for Herbal Plant Identification". *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(1), 152-156,2018
[18] S. O'Hara and B.A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval". arXiv preprint arXiv:1101.3354, 2011.
[19] C. Zhang, P. Wang, , K. Chen, and J. K. Kämäräinen, "Identity-aware convolutional neural networks for facial expression recognition". *Journal of Systems Engineering and Electronics*, *28*(4), 784-792, 2017.

BIOGRAPHIES OF AUTHORS

Nik Noor Akmal Abdul Hamid is a Master's student at Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia where she is continuing her education in the field of Computer Science in Web Technology. Her area of interests is image processing, data science and big data, and electronic commerce.

Rabiatul Adawiya Razali is a Master's student at Universiti Teknologi MARA, Shah Alam, Selangor in the field of Computer Science in Web Technology. Her area of interest is image processing, database and knowledge-base.

Zaidah Ibrahim is an Associate Professor at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. She has been teaching courses related to Artificial Intelligence for over ten years. She is actively involved in research and publication under Digital Image, Audio and Speech Technology (DIAST) research interest group that include text and object recognition.