Indonesian Journal of Electrical Engineering and Computer Science
Vol. 12, No. 2, November 2018, pp. 476~481
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v12.i2.pp476-481

Celebrity Face Recognition using Deep Learning

Nur Ateqah Binti Mat Kasim[1], Nur Hidayah Binti Abd Rahman[2], Zaidah Ibrahim[3], Nur Nabilah Abu Mangshor[4]
[1,2,3]Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA (UiTM), Shah Alam, Selangor, Malaysia
[4]Faculty of Computer and Mathematical Sciences, UniversitiTeknologi MARA (UiTM), Campus Jasin, Melaka, Malaysia

Article Info

Article history:

Received May 29, 2018
Revised Jul 30, 2018
Accepted Aug 3, 2018

Keywords:

AlexNet
Convolutional neural network
Deep learning
Face recognition
GoogLeNet

ABSTRACT

Face recognition is one of the well studied problems by researchers in computer visions. Among the challenges of this task are the occurrence of different facial expressions like happy or sad, and different views of the images such as front and side views. This paper experiments a publicly available dataset that consists of 200,000 images of celebrity faces. Deep Learning technique is gaining its popularity in computer vision and this paper applies this technique for face recognition problem. One of the techniques under deep learning is Convolutional Neural Network (CNN). There is also pre-trained CNN models that are AlexNet and GoogLeNet, which produce excellent accuracy results. The experimental results indicate that AlexNet is better than basic CNN and GoogLeNet for face recognition.

Corresponding Author:

Nur Ateqah Binti Mat Kasim,
Faculty of Computer and Mathematical Sciences,
UniversitiTeknologi MARA (UiTM),
Shah Alam, Selangor, Malaysia.
Email: ateqahkasimofficial@gmail.com

1. INTRODUCTION

The great progress of automatic face recognition in recent years has made large-scale face identification possible for many practical applications [1]. This application is widely used when the images for the persons to be recognized are available beforehand, and an accurate recognizer is needed for a large and relatively fixed group of people. For example, most of the face recognition application is used for search engine [2], recognition for public figure in media industry, and video streaming companies for movie character annotation [3]. In this paper, we investigate the application of deep learning methods namely Convolution Neural Network (CNN) and two pre-trained CNN models that are AlexNet and GoogLeNet for face recognition due to their excellent accuracy performances in computer vision.

2. RELATED WORK

Previously, research in object recognition uses handcrafted features such as texture features for fall activity recognition [4] and leaf recognition [5]. Besides that, color features have also been applied for fruit recognition [6] where it involves identifying the significant feature and classifier to obtain good recognition results. However, currently, the object recognition research has progressed to Deep Learning (DL) where no handcrafted feature is required and yet the results produced are excellent.

Deep Learning (DL) was applied to solve many problems for the last few years. The problems range from computer vision to natural language processing. In many cases DL outperformed other existing techniques [7]. DL methods start by extracting a representation of the face image using local image

descriptors. Then they aggregate such local descriptors using pooling mechanism into an overall face descriptor. This work is concerned with deep architectures for face recognition. The defining characteristic of such methods is the use of a CNN feature extractor, a learnable function obtained by composing several linear and non-linear operators. DeepFace [8] is a representative system of this class. This method uses a deep CNN to classify faces using 4 million examples of training images spanning 4000 unique identities. The same CNN is applied to pair faces to obtain descriptors and the matched similarity images are performed using Euclidean distance [8]. The goal of training process is to maximise the distance between incongruous pairs and minimise the distance between congruous pairs of faces which is portraying the same identity. In addition to using very large amount of training data, DeepFace uses an ensemble of CNN as well as pre-processing phase where the face images are aligned to a canonical pose using a 3D model.

Another application that uses DL is automatic colorization of black and white image [9]. DL can be used to colour the image by using the objects and their context within the photograph. It acts much like a human operator. These capability leverages of the high quality and very large CNN trained for ImageNet and co-opted for the problem of image colorization. The approach involves the use of very large CNN and supervised layers that recreate the image with the addition of color features.

Besides that, DL can be utilized to add sounds to silent movies. In this task it will synthesize sounds to match a silent video [10]. The system is trained using 1000 examples of videos with sound of a drum stick striking different surfaces. A DL model associates with the video frames of pre-recorded sounds in order to select sounds to play that best matches with the scene [10]. The system was then evaluated using a turing-test like setup where humans had to determine which video had the real or the fake (synthesized) sounds [10]. It used both CNN and LSTM recurrent neural networks [10].

DL also can be used to classify and detect text and objects in photographs [11]. State-of-the-art results have been achieved on benchmark examples of this problem using very large CNN. A breakthrough in this problem by Alex Krizhevsky et al. results on the ImageNet classification problem called AlexNet [11].

3. RESEARCH METHOD

3.1. The Dataset

Celebrity face dataset has been used for training where it stores at most 200,000 and 40 attributes [12]. Different face expressions, views and background are the sample of 40 attributes indicated in this dataset. Figure 1 shows the sample attributes includes in this dataset. There are different attributes in the datasets; gender is one of the examples of the attributes [12].

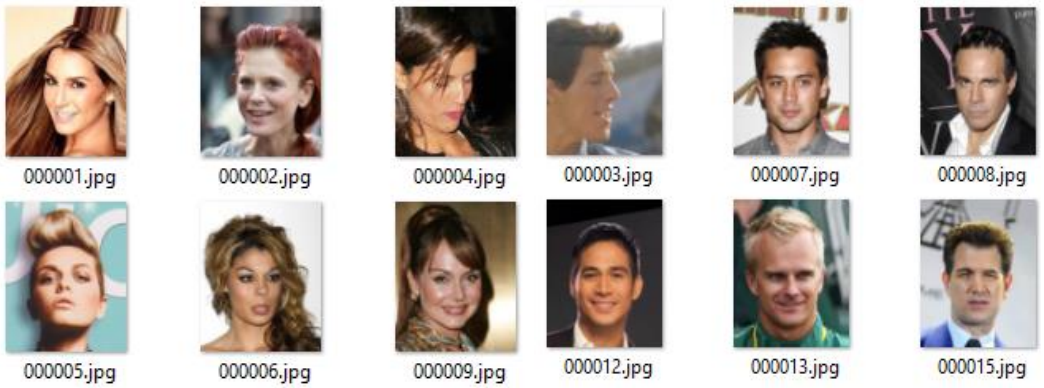

Figure 1. Celebrity face classified by gender [12]

3.2. Convolutional Neural Network

Convolutional Neural Networks (CNN) have taken the computer vision community by storm, it is significantly improving the state of the art in many ways in computer vision applications.The important ingredients for the success of such methods is the availability of large quantities of training data. However, in the world of face recognition, large scale public datasets have been lacking, and largely due to this factor, most of the recent advances in the community remain restricted to Internet giants such as Facebook and Google. For example, the most recent face recognition method by Google was trained using 200 million images and eight million unique identities.

The current CNN models for face recognition tend to be deeper and larger to fit large amount of the data from the public resources such as the Internet. According to Xiang Wu [13] the performance of CNN has greatly improved, for example, the accuracy on the challenging LFW benchmark has been improved from 97% to 99% [13]. This improvement is mainly due to the fact that CNN can learn a complex data distribution from the large-scale training dataset.

Several recent papers have also hypothesized that CNN develop an understanding about objects based on the training data, as such that they are even able to generate new images [14]. However human is very capable to recognize unfamiliar objects, by identifying their important features, mainly their shapes. They can also identify objects in various forms such as different scales, orientations, colours or brightness. Therefore, it remains to be seen how CNN compare to humans in terms of "semantic generalization". Figure 2 illustrates the architecture of a CNN. The input is an image used for recognition, during convolutional process, the output of the image became activation map. Convolutional layer acts as a filter towards the input in terms of sizes, padding, features and etc. Pooling layer is operating as a reducer for number of parameters. Both layer acted as features extraction to produce a generic features. At the end, the output layer act as fully connected layer. There are a few layers that lie on the output layer such as output generator layer for generating the loss while training the image [15].

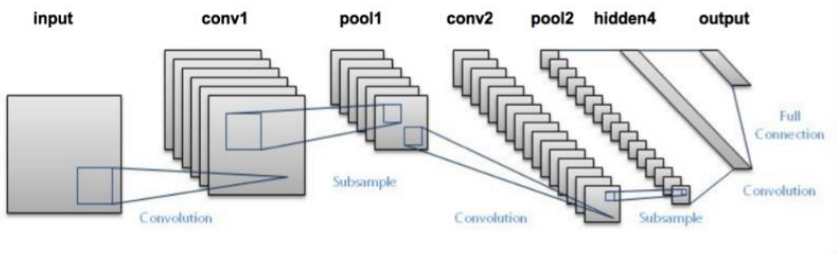

Figure 2. The image of CNN architecture [15]

3.3. AlexNex

AlexNet achieved the top 5 errors from 26% to 15.3% in ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [10]. The network had similar architecture as LeNet by YannLeCun et al but was deeper [8]. It also has more filters per layer with stacked convolutional layers consisting of 11x11, 5x5,3x3, convolutions, max pooling, dropout, data augmentation, ReLU activations, SGD with momentum [10]. ReLU activation is attached after every convolutional and fully-connected layer. AlexNet was trained for 6 days simultaneously on two Nvidia Geforce GTX 580 GPUs which is the reason for why their network is split into two pipelines [16]. AlexNet method is designed by the SuperVision group, which is consisting of Geoffrey Hinton, Alex Krizhevsky, and IlyaSutskever [10]. Figure 3 illustrates the architecture of AlexNet.

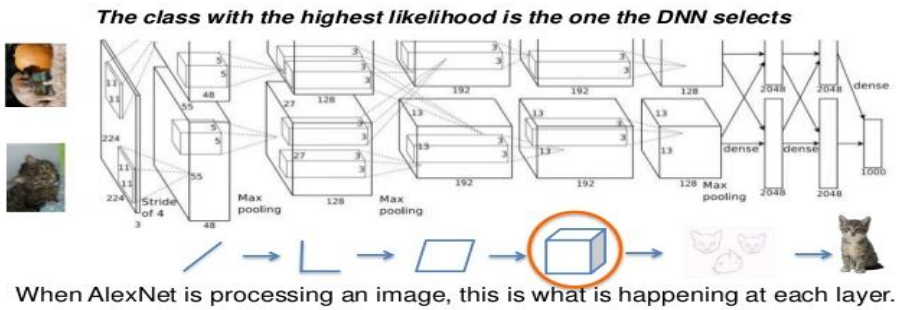

Figure 3. The image of AlexNet architecture [10]

3.4. GoogLeNet

GoogLeNet achieved the top-5 error rate of 6.67% according to "Deep sparse rectifier networks In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics" [17]. This was much closer and almost similar to human level performance. As it turns out, this was actually rather hard to

do and required some human training in order to beat GoogLeNets accuracy. After a few days of training, the human expert (Andrej Karpathy) was able to achieve a top-5 error rate of 5.1% for single model and 3.6% for ensemble. It used batch normalization, image distortions and RMSprop. This module is based on several very small convolutions in order to drastically reduce the number of parameters. GoogLeNet's architecture consist of 22 layer but reduced the number of parameters from 60 million (AlexNet) to 4 million. Figure 4 illustrates the architecture of GoogLeNet [17]. In GoogLeNet, there are 9 inception modules occured for considering the clustering and network within the network. During the inception modules, the module range being calculated and removing the fully connected layers. Meanwhile, pooling in the inception modules reduces the numbers of parameters involved. Besides that, shadow network and auxiliary classifier are added to provide better outputs. Furthermore, GoogLeNet has more layers due to the 9 inception modules that repeating the convolutional, pooling, softmax and concat processes [17], [18].

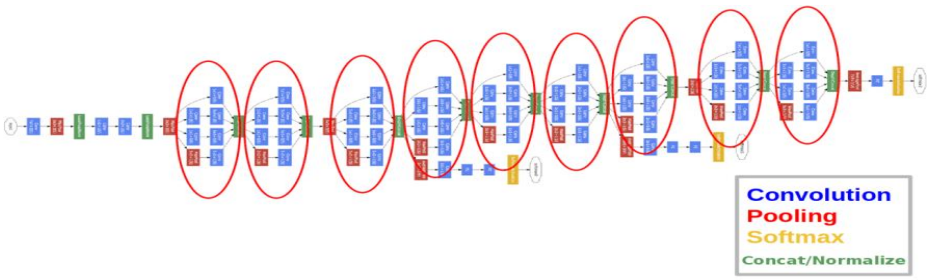

Figure 4. Image of GoogLeNet architecture [17]

4. RESULTS AND ANALYSIS

Matlab R2018b is use as the tool to train and test the dataset for this paper. The size of the image is define as 100x100x3 which means that the size of the image is 100 x 100 pixels and the value 3 indicates that the training image is a color image. The first convolution layer of CNN extracts the edges of the image presented. Figure 5 shows the results of the celebrity face recognition with validation accuracy of 99.72%. The elapsed time is 48 seconds to complete the process with maximum 232 iterations and 4 epochs. Furthermore, the average for iteration per epoch is 58. Meanwhile, the iteration frequency is 30 iterations and patience is 5 and the learning rate 0.01 with the schedule of learning rate is constant.

Figure 5. Accuracy of training

4.1. AlexNet

AlexNet is one of the pre-trained CNN models. The scale of the size and color is defined as 227 x 227x3. The accuracy has achieved 100%. Figure 5 shows the result for AlexNet. The validation accuracy is 100% which means that it has achieved is very accurate. The elapsed time is 9 min 8 seconds for the process

to complete and achieve such accuracy with 6 epochs and 102 iterations. Hence, the iterations per epoch are 17. The learning rate schedule is constant with 0.0001. The training process achieves 100% accuracy since first epoch.

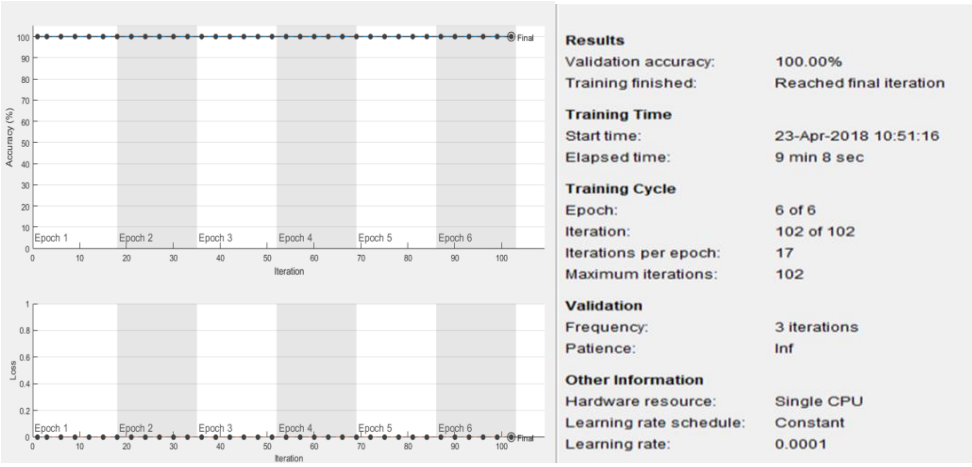

Figure 6. The result performance of AlexNet

4.2. GoogLeNet

GoogLeNet is another popular pre-trained CNN model that has been reported to produce very high accuracy [8]. Size image is define at 227x227x3. For this experiment, GoogLeNet receives the same accuracy as AlexNet which is 100%. Figure 6 shows the results of GoogLeNet. The validation accuracy is 100% equivalent to 1 after it reached the final iteration. The completion time is 14 minutes 47 secords with 6 epochs and maximum iteration is 102. Hence, the iterations per epoch are 17. As shown on the graph, the learning rate schedule is constant at 0.0001.

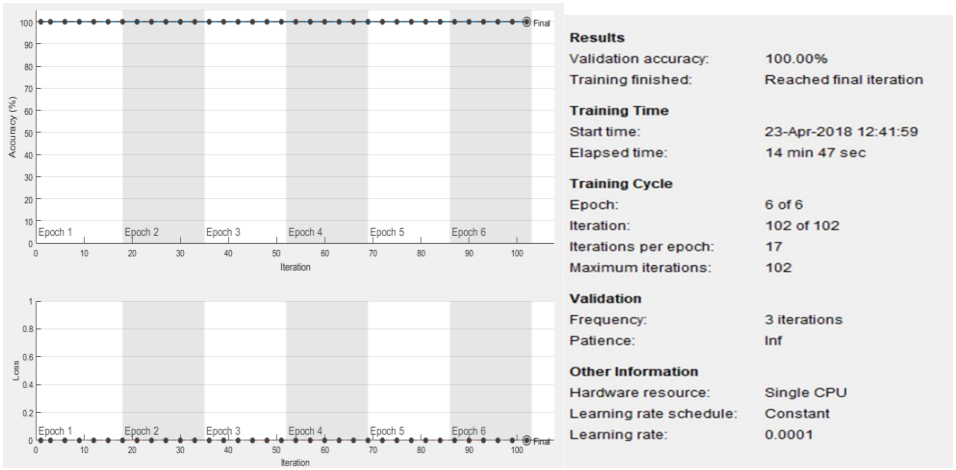

Figure 7. Result performance produced by GoogLeNet

5. CONCLUSION

Table 1 shows the promising results produced by CNN, AlexNet and GoogLeNet despite the differences in gender, face expressions, hair style, features, and background of the images in the dataset. Based on the results displayed in Table 1, we can see that AlexNet and GoogLeNet have better accuracy which is 100% compared to CNN which is 99.72%. It shows that the machine has perfectly recognized all celebrity images in the dataset using AlexNet and GoogLeNet. This is due to the training of millions of data.

Meanwhile, the speed of processing CNN recorded the fastest to complete training compared to AlexNet and GoogLeNet. This is due to the number of layers in these models. The more number of layers, the more time it takes to produce the results. CNN completes the execution or converges after 48 seconds while AlexNet achieved 100% accuracy in 9 minutes and 8 seconds. GoogLeNet requires 14 minutes and 47 seconds to converge or complete the execution.

In selecting which module to use, various factors need to be considered such as the availability of large amount of training data, the amount of time that can be spared for the training process and the number of errors that can be accepted. Future research includes the improvement of the CNN architecture and experiment on other pre-trained CNN models.

Table 1. Comparisons between CNN, AlexNetGoogLeNet

Parameters	CNN Scratch	AlexNet	GoogLeNet
Validation Accuracy	99.72%	100.00%	100.00%
Elapsed Time	48 secords	9 minutes 8 seconds	14 minutes 47 seconds
Number of Epoch	4	6	6
Number of Iteration	232	102	102
Validation Frequency	30	3	3

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] J. Hashemi, Q. Qiu and G. Sapiro, "Intelligent Synthesis Driven Model Calibration: Framework and Face Recognition Application", International Conference on Computer Vision (ICCV) 2017.
[2] R. Desai and B. Sonawane, "Gist, HOG, and DWT-based Content-based Image Retrieval for Facial Images", International.
[3] X. Qin, Y. Zhou, Z. He, Y. Wang and Z. Tang, "A Faster R-CNN based Method for Comic Characters Face Detection", 14th IAPR International Confeence on Document Analysis and Recognition (ICDAR) 2017.
[4] W. Ong Vui Jiunn, N. Sabri and Z. Ibrahim, "Image-based Human Fall Recognition using Gaussian Mixture Model and Support Vector Machine", International Journal of Control Theory and Applications, vol. 9, number 44, 2016.
[5] Z. Ibrahim, N. Sabri and N. N. Mohd Manghor, "Leaf Recognition Using Texture Features for Herbal Plant Identification', International Journal of Electrical Engineering and Computer Science (IJEECS), Vol. 9, No. 1 2018, pp.152-156.
[6] N. Sabri and Z. Ibrahim, "Palm Oil Fresh Fruit Bunch Ripeness Grading Identification using Color Features", Journal of Fundamental and Applied Science, 2017, 9(4S), pp. 563-579.
[7] Hadad Y (2018) Amazing Application of Deep Learning
[8] A. Kortylewski, B. Egger and A. Schneider, "Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems", Computer Vision and Pattern Recognition (CVPR) 2018.
[9] D. Varga, C. A. Szabo and T. Sziranyi, "Automatic Cartoon Colorization based on Convolutional Neural Network", 15th International Workhop on Content-Based Multimedia Indexing, June 2017, pp. 19-21.
[10] Brownlee, J. (2016, July 29). 8 Inspirational Applications of Deep Learning.
[11] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Networks", Advances in Neural Information Processing Systems 25, 2012.
[12] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From Facial Parts Responses to Face Detection: A Deep Learning Approach", in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
[13] Wu X, He R, Sun Z, Tan T, "A Light CNN for Deep Face Representation with Noisy Labels", IEEE Transactions on Information Forensics and Security (2018).
[14] Hosseini H, Xiao B, Jaiswal M, Poovendran R On the Limitation of Convolutional Neural Networks in Recognizing Negative Image.
[15] Gupta, D., Jain, K., Jain, A., & Analytics Vidhya Content Team. (2017, June 29). Architecture of Convolutional Neural Networks (CNNs) demystified.
[16] Gao H (2017) A Walk-through AlexNet.
[17] Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier networks In Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. JMLR W&CP Volume (Vol. 15, pp. 315-323).
[18] Arora, S., Bhaskara, A., Ge, R., & Ma, T. Provable bounds for learning some deep representations. ICML 2014.