Recognition of Badminton Action Using Convolutional Neural Network

Nur Azmina Rahmad[1], Nur Anis Jasmin Sufri[2], Muhammad Amir As'Ari[3], Aizreena Azaman[4]
[1,2]School of Medical Engineering and Health Sciences, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[3,4]Sport Innovation and Technology Center (SITC), Institute of Human Centered Engineering (IHCE), Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

Article Info

Article history:

Received Jan 17, 2019
Revised May 1, 2019
Accepted Nov 11, 2019

Keyword:

Action recognition
Convolutional Neural Network
Deep learning
Image classification

ABSTRACT

Deep learning approach has become a research interest in action recognition application due to its ability to surpass the performance of conventional machine learning approaches. Convolutional Neural Network (CNN) is among the widely used architecture in most action recognition works. There are various models exist in CNN but no research has been done to analyse which model has the best performance in recognizing actions for badminton. Hence, in this paper, we are comparing the performance of four different established pre-trained models of deep CNN in classifying the badminton match images to recognize the different actions done by the athlete. Four models used for comparison are AlexNet, GoogleNet, VggNet-16 and VggNet-19. This experimental work categorized images into two classes: hit and non-hit action. Firstly, each image frame was extracted from Yonex All England Man Single Match 2017 broadcast video. Then, the image frames were fed as the input to each classifier model for classification. Finally, the performance of each classifier model was evaluated by plotting its performance accuracy in the form of confusion matrix. The result shows that the GoogleNet model has the highest classification accuracy which is 87.5% compared to other models. In a conclusion, the pre-trained GoogleNet model is capable to be used in recognizing actions in badminton match which might be useful in badminton sport performance technology. The main contribution of this paper is that it provides an analysis of the performance of four different pre-trained deep CNN models in recognizing badminton actions which have not been done before by other researchers. Thus, the analysis will help in the future work to improve the existing deep learning models' architecture for a better performance in badminton action recognition.

Corresponding Author:

Nur Azmina Rahmad,
School of Medical Engineering and Health Sciences,
Universiti Teknologi Malaysia,
Johor Bahru, Malaysia.
Email: azminarahmad@gmail.com

1. INTRODUCTION

The computer vision field has been widely used in various applications such as video surveillance, human-computer interaction, robotics, object andaction recognition and sport analysis [1, 2]. Action recognition is a very challenging problem in computer vision field. There are two modalities in action recognition: 1) sensor-based modality and 2) video-based modality. In this new era of technology, where video transmissions are widely available online, video-based modality is increasingly used in recognizing the action.

There are three components in action recognition framework: 1) feature extraction, 2) action representation and 3) classification as illustrated in Figure 1.

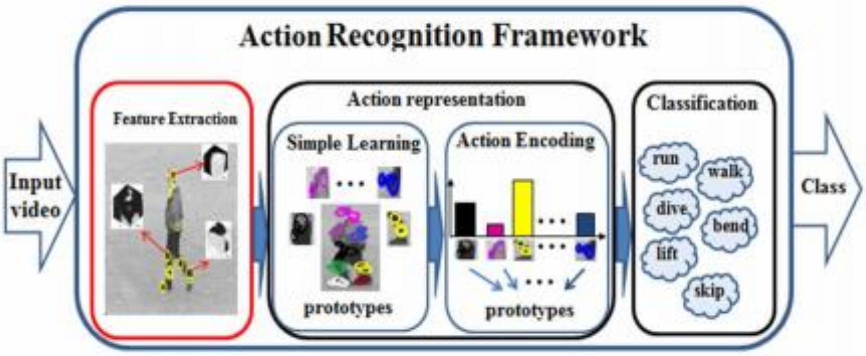


Figure 1. Action recognition framework [3]

Currently, deep learning approach has become a research interest in action recognition because handcrafted approach does not have the capability to extract high-level features due to certain limitation such as image or video noise and complexity [3-6]. However, deep learning works excellently in extracting high-level features directly from raw data as its architecture consists of hundreds of hidden layers.

CNN is one of a supervised classification technique. It has taken place in many recent works with its simple but precise architecture. CNN falls into the deep learning classifier category in which it eliminates the manual feature extraction in the machine learning pipeline. CNN model will automatically extract the features of the image before classifying it into respective class [4]. The pipeline is similar with Artificial Neural Network (ANN): input layer, hidden layer and output layer, but the hidden layer of CNN could consist up to hundreds of layers to improve its performance accuracy. There are several pre-trained CNN model with different network architecture that are available such as LeNet, AlexNet, VggNet and ResNet. To train the CNN architecture from scratch required many data and consumes a lot of time. However, another way to train the CNN in a short time and does not require so much data is through transfer learning which the existing pre-trained CNN model can be used.

There are few works have been done on implementing and analysing CNN in their studies [5-16]. Work in [5] evaluates the performance of two classifiers and two feature extractors in classification of Caltech 265 images. Two classifiers used in comparison study are Linear Support Vector Machine (SVM) and Quadratic SVM while two feature extractors used are Bag of Words (BoW) and pre-trained CNN. The study proved that the classification accuracy is the highest when the features were extracted from CNN.

In [6], the authors introduced an improved AlexNet model for scene classification, as AlexNet model is limited in image classification by decomposing the large convolutional kernel into two small convolutional kernels with reduced stride. 5*5 convolution is decomposed into two 3*3 convolution and 3*3 convolution is decomposed into a structure of 3*1 convolution then 1*3 convolution. The experiment was conducted on SUN397 and Places 2 datasets. In comparison with AlexNet and ZFNet model, the proposed improve AlexNet model has the highest accuracy.

Study in [7] compared two CNN models (GoogleNet and AlexNet) in classifying the different flowers using Visual Geometry Group’s 102 category flower dataset. The method was divided into image segmentation and classification. Image segmentation was used to remove the background from images. Their finding is that GoogleNet performs better than AlexNet in flowers categorization.

The purpose of this study is to implement and investigate the performance and capability of transfer learning method of different pre-trained CNN models in recognizing badminton action. At the end of this study, a suitable pre-trained CNN model will be proposed to automatically recognize the actions in badminton from broadcasted video. For an efficient sport performance analysis, the automated action recognition system in sport field will be very beneficial to coach. In Section 2, we provide an explanation of our methodology and design of experiment. In Section 3, we provide the results and briefly discuss the obtained results. Lastly, the conclusion and further work are stated in Section 4.

2. RESEARCH METHOD

Figure 2 shows the flow of the experimental work.

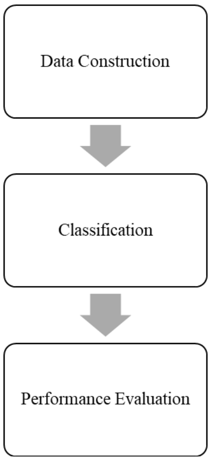


Figure 2. Block diagram of the experimental work

For dataset construction, firstly, the full duration broadcast video of Yonex All England Man Single Match 2017 with 720p resolution and frame rate 25 frames per second obtained from the Youtube database was extracted into still image frames. The purpose of using the still image frames in this study because we want to avoid the video's variable length problem. For instance, one video image might be 20 seconds while another is 50 seconds. This video extraction produced 138130 image frames. Then, we annotated each image frame into hit and non-hit action. Hit action refers to the action of players hitting the shuttlecock while non-hit action refers otherwise. Lastly, 80 image frames were selected randomly from the total image frames which consist of 40 images for hit action and 40 images for non-hit action. The extraction process was done using VirtualDub software. Figure 3 shows the example of image frames used in this experimental work for hit and non-hit action.

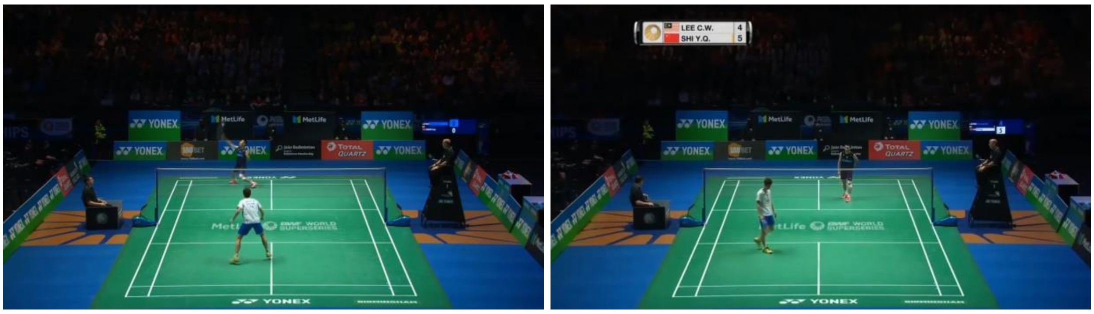


Figure 3. Image frames for hit and non-hit action

In classification, 80 image frames were divided into 64 training images (32 for hit and 32 for non-hit) to train the CNN and the remaining 16 image frames (8 for hit and 8 for non-hit) as testing images to test the CNN. Then, the image frames were fed to each model for classification.

The training of AlexNet and GoogleNet model took place on graphical processing unit (gpu) Nvidia GeForce GT 740 with computing capability 3.0 while the training of VggNet-16 and VggNet-19 model took place on central processing unit (cpu) Intel® Core ™ i7 processor with 3.40 GHz processor speed and 8 GB RAM. The tranining of deep learning algorithm took place on the Matlab 2018b software platform. For each model, the training parameters were set as shown in Table 1.

Table 1. Training parameters

Training options	AlexNet	GoogleNet	Vgg-16 Net	Vgg-19 Net
Training optimizer	Sgdm	Sgdm	Sgdm	Sgdm
Mini-batch size	5	5	1	1
Maximum epochs	10	10	10	10
Execution environment	gpu	gpu	cpu	cpu
Initial learning rate	0.0001	0.0001	0.0001	0.0001

AlexNet, GoogleNet and VggNet are the most popular and widely used CNN models. CNNs are used on vision-based dataset for image classification, object detection, image recognition and image segmentation. Table 2 summarises the details of each model. These models were trained to classify 1000 object categories. However, in this study, we fine-tuned these models with our dataset to classify only 5 action categories.

Table 2. The summary of CNN models

Network	Year	Layer	Salient feature	Parameters	Top5 accuracy
AlexNet	2012	8	Deeper	62M	84.70%
GoogleNet	2014	22	Wider parallel kernels	6.4M	93.30%
Vgg-16 Net	2014	16	Fixed size kernels	138M	92.30%
Vgg-19 Net	2014	19	Fixed size kernels	138M	92.30%

Lastly, the classification performance of each model was analysed in term of performance accuracy and visualised using the confusion matrix. As for the confusion matrix, the columns represent the result of the predicted class and the rows represent the actual class of the variables. Anything on the leading diagonal is a correct answer (green colour) for each different action while others (red colour) are the falsely classified action. As for confusion matrix's legend, 1 represents hit action and 2 represents non-hit action.

3. RESULTS AND ANALYSIS

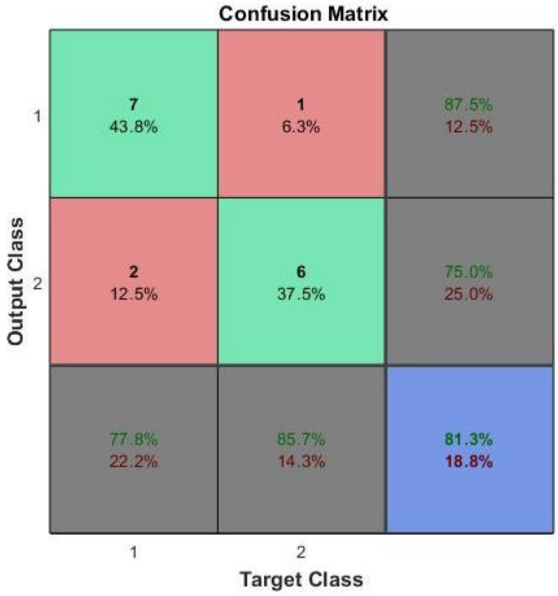
As mentioned earlier, the aim of this experimental work is to evaluate and compare the performance of four different pre-trained CNN models in recognizing the actions of badminton. Table 3 shows the performance accuracy of each model in recognizing the badminton actions, meanwhile Figures 4-7 illustrate the confusion matrix of AlexNet, GoogleNet, VggNet-16 and VggNet-19 model. The green boxes represent the number and percentage of correctly classified actions while red boxes represent the number and percentage of falsely classified actions. Blue box in diagonal line represents the percentage of the performance accuracy.

Table 3. Accuracy table

Model	Performance accuracy (%)
AlexNet	81.3
GoogleNet	87.5
Vgg-16 Net	50.0
Vgg-19 Net	50.0

The equation (1) below is the formula used to obtain the percentage accuracy. The total number of correctly classified actions refers to the sum of all the correct predicted classes in diagonal as illustrated in the confusion matrix while the total number of test samples refers to the total number of test samples used which is 16.

$$\%\ \text{accuracy} = \frac{Total\ number\ of\ correctly\ classified\ actions}{Total\ number\ of\ test\ samples}\ x\ 100\% \qquad (1)$$



	1	2	
1	7 43.8%	1 6.3%	87.5% 12.5%
2	2 12.5%	6 37.5%	75.0% 25.0%
	77.8% 22.2%	85.7% 14.3%	81.3% 18.8%

Figure 4. Confusion matrix for AlexNet model

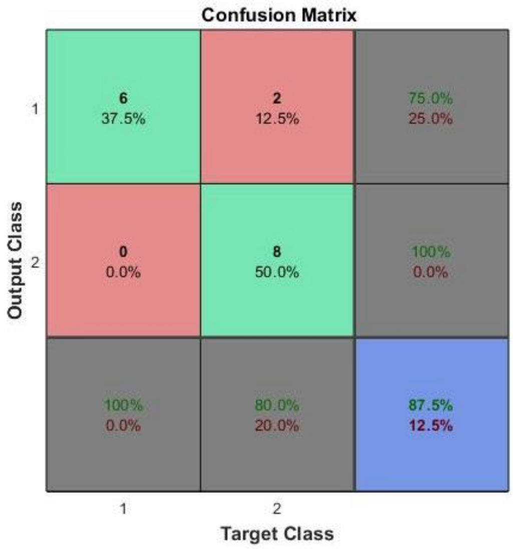


Figure 5. Confusion matrix for GoogleNet model

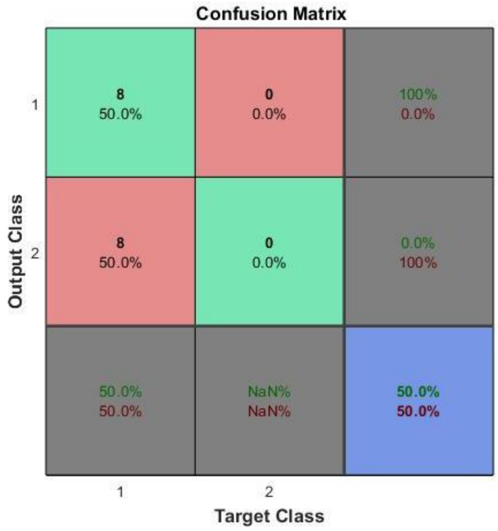


Figure 6. Confusion matrix for VggNet-16 model

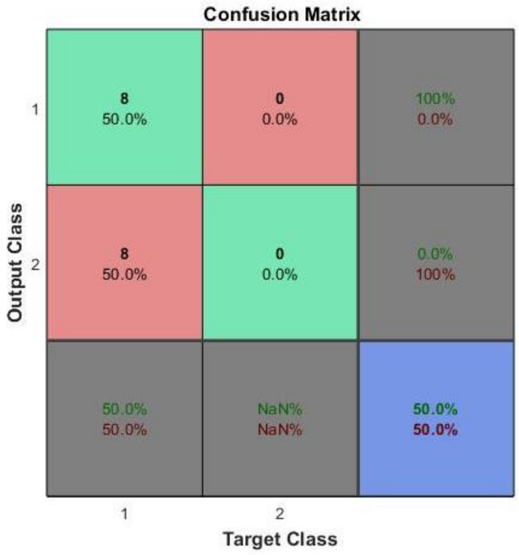


Figure 7. Confusion matrix for VggNet-19 model

Significantly, GoogleNet model has the highest accuracy compared to other model in which only two hit actions were falsely classified as non-hit action. There is only a slight difference in accuracy percentage between AlexNet and GoogleNet model (81.3% and 87.5% respectively) as shown in Figure 8.

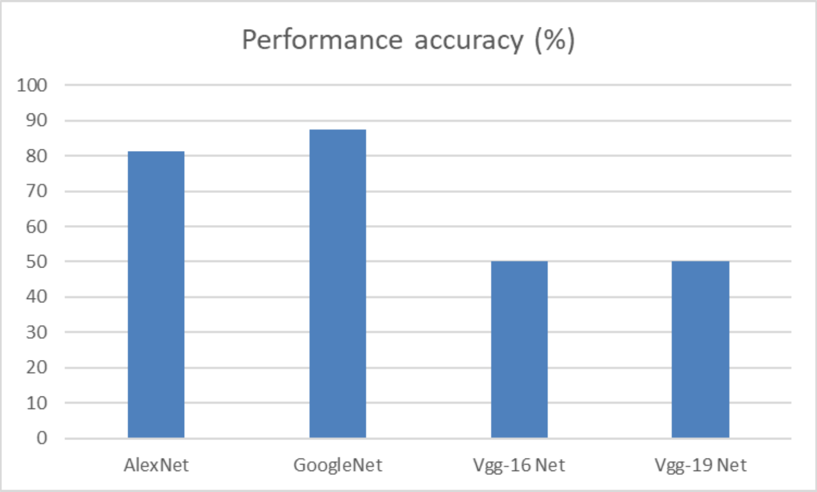
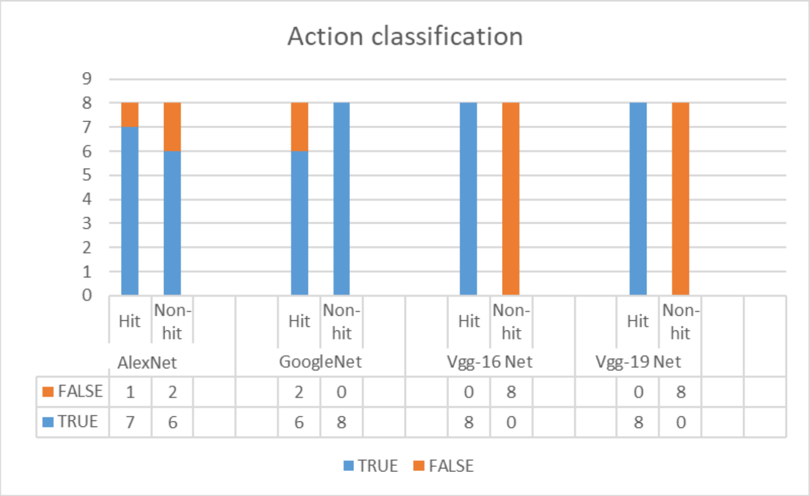


Figure 8. Bar chart of performance accuracy

This supports the previous study by [7], where GoogleNet has a better performance in flower categorization but not to a great extend. Whereas, both VggNet models have been left behind with only 50.0% of accuracy in which all non-hit actions were falsely classified as hit action as shown in Figure 9.



	AlexNet		GoogleNet		Vgg-16 Net		Vgg-19 Net	
	Hit	Non-hit	Hit	Non-hit	Hit	Non-hit	Hit	Non-hit
🟧 FALSE	1	2	2	0	0	8	0	8
🟦 TRUE	7	6	6	8	8	0	8	0

Figure 9. Bar chart of the number of data based on true or false classification

These four models have different architecture. As described in [17], AlexNet model consists of 8 learned layers- 5 convolutional layers and 3 fully-connected layers with 60 million parameters. However, GoogleNet model has 22 learned layers with number of parameters that have been reduced to 4 million by inception module [18]. According to [19], VggNet-16 and VggNet-19 model has 16 and 19 learned layers respectively with 140 million parameters. Therefore, the results strongly support the claim of previous studies that the deepest network has the highest accuracy. For this reason, GoogleNet model has the highest accuracy compared to AlexNet and VggNet model because it has the deepest layer. But, the results also show that Alexnet model performs better than VggNet model even though VggNet model has a deeper layer. This is because VggNet model has 140 million parameters compared to AlexNet model that only has 60 million parameters. As stated in [20], small amount of parameter variation can achieve significant growth in performance.

Overall, it can be inferred that GoogleNet model can perform better in recognizing action in bádminton. We also aware that our study may have two limitations. The first is GPU memory and the second is training time. Since the GPU is out of memory to train both VggNet models, we trained the models using the CPU, but take a longer time to complete the training process. Not only that, we found out that the machine used to train the model affects the performance accuracy of the model. The performance accuracy of VggNets drop significantly to 50%, even though these models should perform better than AlexNet. It is compulsory that the same machine should be used as a limited function of CPU may greatly affect the results. It is plausible that a number of limitations might could have influenced the results obtained.

4. CONCLUSION

Sport performance analysis is an important branch in sport practice. In order to analyse the performance of athletes using notational analysis approach, the sport analyst will manually recognize the action before doing the analysis. At this stage, this study provides an analysis on the performance of deep learning models in recognizing badminton action. It can contribute to the automatic action recognition using the most simple and non-time consuming transfer learning method which has not been done before. In the future, the experiment can be improved by classifying more action in badminton instead of classifying the action into hit and non-hit. Moreover, we believed that this study is the starting point in developing more advance deep learning architecture for automated badminton action recognition.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Universiti Teknologi Malaysia for the support of the project through Zamalah Scholarship and the Minister of Higher Education (MOHE) for supporting this research work under Research Grant No. Q.J130000.2545.19H88.

REFERENCES

[1] Chaquet, J. M., Carmona, E. J. and Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding. 2013; 17: 633-659.

[2] Borges, P. V. K., Conci, N. and Cavallaro, A. Video-based human behavior understanding: a survey. IEEE Transactions on Circuits and Systems for Video Technology. 2013; 23: 1993-2008.

[3] Koohzadi, M. and Charkari, N.M. Survey on deep learning methods in human action recognition. IET Computer Vision. 2017; 11: 623-632.

[4] Md Saufi, M., Zamanhuri, M. A., Norasiah, M. and Ibrahim, Z. A. Deep Learning for Roman Handwritten Character Recognition. Indonesian Journal of Electrical Engineering and Computer Science. 2018; 12(2): 455-460.

[5] Mat Kasim, N. A., Abd Rahman, N.H., Ibrahim, Z. and Abu Mangshor, N. N. Celebrity Face Recognition using Deep Learning. Indonesian Journal of Electrical Engineering and Computer Science. 2018; 2012(2): 476-481.

[6] Muhammad, N. A., Ab Nasir, A., Ibrahim, Z., Sabri, N .Evaluation of CNN, Alexnet and GoogleNet for Fruit Recognition. Indonesian Journal of Electrical Engineering and Computer Science. 2018; 12(2): 468-475.

[7] O'Shea, K. and Nash, R. An Introduction to Convolutional Neural Network. CoRR. 2015; abs/1511.08458: 1-11.

[8] Karpathy, A., et al., Large-Scale Video Classification with Convolutional Neural Networks. 2014. p. 1725-1732.

[9] Ullah, A., et al., Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. IEEE Access, 2018. 6: p. 1155-1166.

[10] Wang, L., Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.

[11] Simonyan, K. and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. in Neural Information Processing Systems Conference. 2014. Montreal,Canada: Neural Information Processing Systems Foundation, Inc.

[12] Tsunoda, T., et al. Football Action Recognition Using Hierarchical LSTM. in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017.

[13] Tora, M.R., J. Chen, and J.J. Little. Classification of Puck Possession Events in Ice Hockey. in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017.

[14] Christian, S., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Dumitru Erhan, D., Vanhoucke, V. and Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE CVPR 2015, 2015; 1-12.

[15] Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. CoRR. 2014; abs/1409.1556: 1-14.

[16] Pasi, K.G. and Naik, S.R. Effect of parameter variations on accuracy of convolutional neural network. In Computing, Analytics and Security Trends (CAST), International Conference, IEEE. 2017; 398-403.

[17] Abdullah and Hasan, M. S. An application of pre-trained CNN for image classification. In Computer and Information Technology (ICCIT). 2017; 1-6.

[18] Xiao, L. and Yan, Q. Scene classification with improved Alexnet model. In 2017 12th International Conference on Intelligent Systems and Knowledge Engineeering (ISKE). 2017; 1-6.

[19] Gurnani, A., Mavani, V., Gajjar, V. and Khandhediya, Y. Flower categorization using deep convolutional neural networks. CoRR. 2017; abs/1708.03763: 1-4.

[20] Krizhevsky, A., Sutskever, I. and Hinton, G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012; 25: 1-9.