COMPARATIVE EVALUATIONS OF CNN BASED NETWORKS FOR SKIN LESION CLASSIFICATION

Evgin Goceri and Ayse Akman Karakas
Akdeniz University, Engineering Faculty, Biomedical Engineering Department, Antalya, Turkey

ABSTRACT

The aim of this work is to classify Hemangioma, Rosacea and Acne Vulgaris diseases from digital colored photographs automatically. To determine the most appropriate deep neural network for this multi-class classification, network architectures have been examined. To perform a meaningful comparison of deep networks, they should be (i) implemented with the same parameters, (ii) applied with the same activation, loss and optimization functions, (iii) trained and tested with the same datasets, (iv) run on computers having the same hardware configurations. Therefore, in this work, five deep networks, which are applied widely in image classification, have been used to compare their performances by considering these factors. Those networks are VGG16, VGG19, GoogleNet, InceptionV3 and ResNet101. Comparative evaluations of the results obtained from these networks have been performed in terms of accuracy, precision and specificity. F1 score and Matthew's correlation coefficient values have also been computed. Experimental results indicated that ResNet101 architecture can classify images used in this study with higher accuracy (77.72%) than the others.

KEYWORDS

VGG, GoogleNet, ResNet, Inception, Skin Disease, Lesion Classification

1. INTRODUCTION

Traditional diagnosis of Hemangioma, Rosacea and Acne Vulgaris is mainly based on visual examinations of lesions by a dermatologist. Therefore, computer assisted diagnosis for these diseases with pattern recognition can be performed. By this way, the number of patients in dermatology can be reduced and subjective diagnosis problem can be eliminated.

Superiority of deep neural networks, particularly Convolutional Neural Networks (CNNs), in image classification (Lai, 2019; Korotcov et al., 2017; Lee et al. 2017) was the main motivation behind this study to use a deep network for this multi-class classification. However, there are several deep network architectures designed with different number of layers, different activation, loss and optimization functions, which have an important role in accuracy of a classifier.

To determine the most appropriate deep network, architectures should be examined and their performances should be compared. To perform a meaningful comparison, these networks should be (i) implemented with the same parameters, (ii) applied with the same activation, loss and optimization functions, (iii) trained and tested with the same datasets, (iv) run on computers having the same hardware configurations. Therefore, in this work, different CNN based deep networks have been implemented by considering the above factors and their image performances have been compared. These networks that are used widespread in image classification have been applied; VGG16, VGG19, GoogleNet, InceptionV3 and ResNet101.

Skin lesion images segmented from photographs have been used in this study. Figure 1 shows example photographs (Figure 1.a, b, c) and lesion images (Figure 1.d, e, f). Images have been obtained from several public databases (Web_1, 2020; Web_2, 2020; Web_3, 2020).

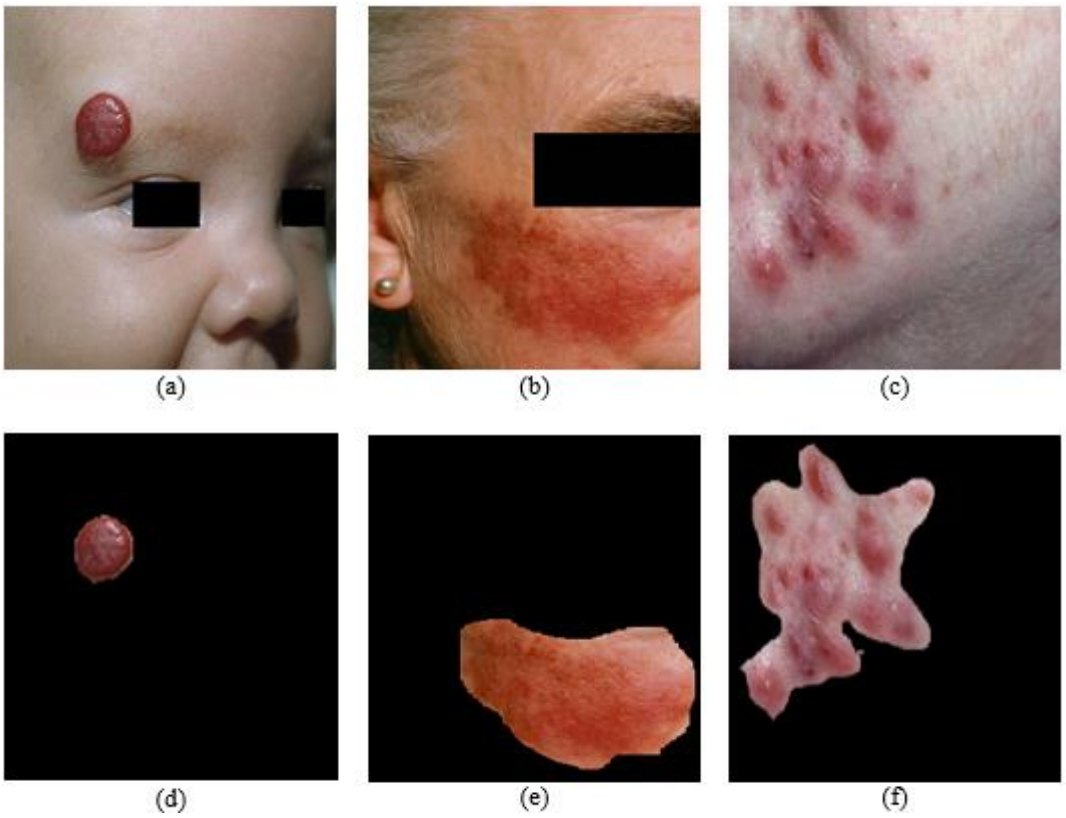

Figure 1. Hemangioma (a), Rosacea (b), Acne Vulgaris (c), Hemangioma lesion (d), Rosacea lesion (e), Acne Vulgaris lesion (f)

In this study, the image dataset has been constructed with equal number of images (101 images) from each disease class. The dataset has been divided into two groups as training and testing. The training (testing) group includes 60 (41) images from each class.

The chosen activation function in this study is ReLU (Rectified Linear Unit) and optimization function is Adaptive Moment Estimation (ADAM). To consider pixel-wise similarities, cross entropy has been used as a loss function. Initial learning ratio, epoch and iteration numbers are 0.0003, 6 and 108 respectively.

Implementations have been performed using Matlab (R2019b) on the same computer, which includes 16 GB RAM memory and Core i7-4930 K processor.

This paper has been structured as follows: Section 2 gives a short information about the network architectures applied in this work. Section 3 explains measurement metrics used to compare classification performances of those networks. Section 4 gives results, finally, Section 5 presents conclusions.

2. CNN BASED NETWORKS APPLIED IN THIS WORK

The deep neural network architectures applied widespread for image classification have been used in this work. Those networks are Visual Geometry Group (VGG) network, GoogleNet, InceptionV3 and ResNet101, which are pretrained on images from ImageNet (Web_4, 2020). In this section, these architectures are explained.

2.1 VGG16 and VGG19

VGG network was proposed by the Visual Geometry Group, Engineering Science Department at the University of Oxford (Simonyan et al., 2015). VGG16 is a 16-layer network. In this architecture, 224x224 colored images are passed through five blocks of convolution layers where each block is composed of increasing numbers of 3x3 filters. The stride is set to one. The blocks are separated using maximum pooling layers, which reduce volume size. A maximum pooling operation is applied over 22 windows with stride two. The five blocks of convolution layers are connected to three fully-connected layers. The top layer is called as soft-max layer to obtain class probabilities as outputs. Similar to VGG16 architecture, VGG19 network model has been used. The "16" and "19" stand for the number of weight layers in the network (Simonyan et al., 2015).

2.2 GoogleNet

GoogleNet architecture has 22 layers and uses Inception modules as the main structural contributions of the network. Therefore, GoogleNet is also known as InceptionV1. Inception module was developed by considering how an optimal local sparse structure in a CNN can be approximated using dense components (Szegedy et al., 2015). GoogleNet architecture uses 3 different filters, which are 1x1, 3x3 and 5x5 dimensional, for the inputs. The 1x1 convolution operation with 128 filters is applicable for dimension reduction. Also, to use less number of parameters, the fully-connected layers are replaced with the one including global-average-pooling after the last convolutional layer. Because the pooling operation calculates mean values of the channel values across the two-dimensional feature map (Szegedy et al., 2015).

2.3 InceptionV3

InceptionV3 is a further version of GoogleNet and was introduced in (Szegedy et al., 2016). Main differences between GoogleNet and InceptionV3 are presented in Table 1.

Table 1. Main differences between GoogleNet and InceptionV3 architectures

Network	Convolution Filters	Pooling Operation	Layers	Feature Map Downsizing	Inception Module	Auxiliary Classifier
GoogleNet	1x1filter, 3x3 filter and 5x5 filter is used	Average value is computed from feature maps and global average pooling is applied	22	Max. pooling	Convolutions using 1x1, 3x3 and 5x5 filters, and max pooling using 3x3 filter is performed together	2
InceptionV3	Two 3x3 filters are used instead of 5x5 filter	Pooling is performed with an appropriate grid size reduction. Convolution and pooling operations are applied to construct feature maps.	42	Grid size reduction	Convolutions are divided into smaller and asymmetric processes. So, a novel inception module is constructed	1

2.4 ResNet101

ResNet architecture was proposed to solve the degradation problem that occurs in deep networks. The problem affects accuracy of deep networks, which saturate during convergence and degrade rapidly (He et al., 2016). Like GoogleNet, a global average pooling is applied before the classification layer in

ResNet. Unlike sequential neural network architectures such as VGG, ResNet includes residual building blocks. Every block in the ResNet architecture accomplishes an identity mapping and produces element-wise addition of the features obtained from convolutional layers. A major novelty in this architecture is to apply a batch normalization step and skip connections to train deeper network structures. ResNet101 is deeper than previous networks.

3. RESULTS

Classification performance have been evaluated in terms of accuracy, precision, specificity, F1 score and also Matthew's Correlation Coefficient (MCC)) (Web_5, 2020). These parameters have been computed using the following formulas;

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$\mathrm{F1}_{score} = 2\left(\frac{PPV * TPR}{PPV + TPR}\right) = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

where TN (True Negative), FN (False Negative), TP (True Positive) and FP (False Positive) terms for a disease, let's say Psoriasis, have the following meaning: TP: Psoriasis images are classified as Psoriasis; FP: non-Psoriasis images that are classified as Psoriasis; FN: Psoriasis images that are not classified as Psoriasis; TN: non-Psoriasis images that are not classified as Psoriasis.

F1 score, which is also known as Dice similarity measure, is computed by considering both precision and sensitivity. It is the harmonic mean of these two parameter values. In (5), the term PPV (Positive Predictive Value) is obtained by TP/(TP+FP) and the term TPR (True Positive Rate) is obtained by TP/P, which is TP/(TP+FN). To obtain F1 score, PPV and TPR values are taken into account equally. However, TN values are not used. Therefore, we also calculated MCC, which is another important metric used in machine learning to analyze quality of classifiers. Table 2 presents numerical values obtained by these six-evaluation metrics.

Table 2. Quantitative values obtained by five network architectures for multiclass lesion classification

Network	Accuracy (%)	Specificity (%)	Precision (%)	F1 score (%)	MCC (%)
VGG16	57.72	78.89	58.11	52.63	37.47
GoogleNet	66.17	90.16	67.41	65.18	58.37
VGG19	69.32	91.41	69.01	68.43	62.47
InceptionV3	72.86	93.02	71.53	71.30	65.00
ResNet101	77.72	94.76	79.81	77.82	72.61

In Figure 2, the lines with light-blue and black color show the training and testing accuracy values obtained by ResNet101, which classifies images with the highest accuracy. The line with dark-blue color shows the training accuracy values after normalization.

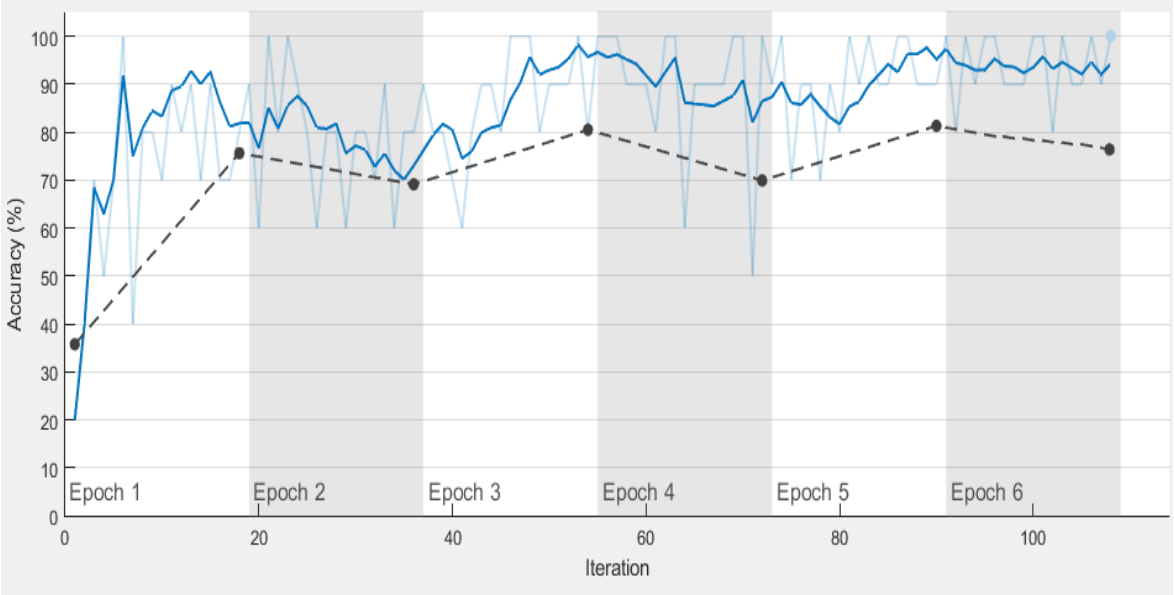

Figure 2. Accuracy values according to epoch obtained by ResNet101 network

In Figure 3, the lines with light-red and black color show the loss values obtained by ResNet101in the training and testing stages. The line with dark-red color shows normalized loss values obtained in the training stage.

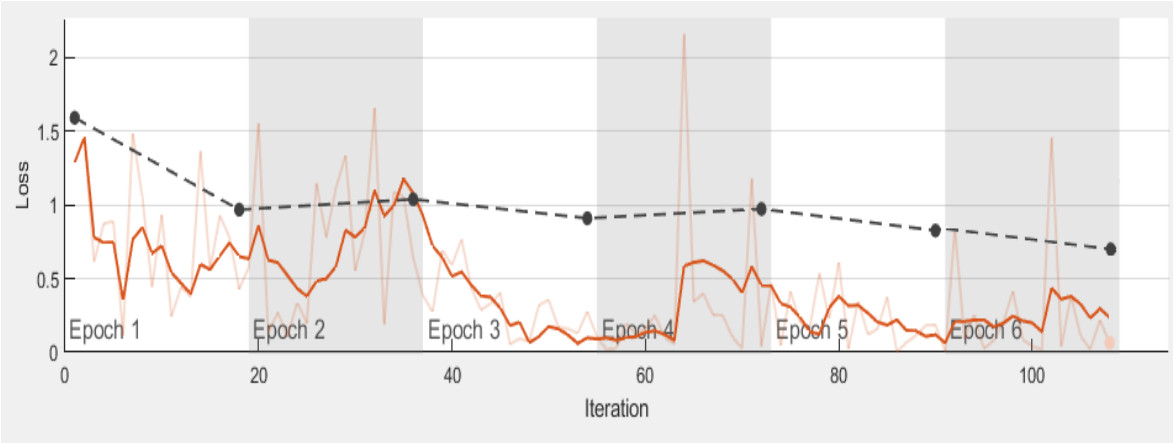

Figure 3. Loss values according to epoch obtained by ResNet101 network

4. CONCLUSION

In this study, classification performances of five deep network structures have been evaluated to classify three skin diseases. Comparative evaluations have been performed using accuracy, specificity, precision, F1 metric and MCC. Quantitative values indicated that ResNet101 can classify images with higher accuracy than the other networks.

ACKNOWLEDGEMENT

This work has been supported by The Scientific and Technological Research Council of Turkey (TUBITAK - 118E777).

REFERENCES

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 770–778.

Korotcov, A., Tkachenko, V., Russo, D.P., Ekins, S., 2017. Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceutics*, Vol.14, No.12, pp.4462-4475

Lai, Y., 2019. A Comparison of Traditional Machine Learning and Deep Learning in Image Recognition. *Journal of Physics: Conference Series, 3rd International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE2019)*, 9–11 August 2019, Guizhou, China, Vol. 1314, No.012148, pp.1-9

Lee, J.G., Jun, S., Cho, Y.W., Lee, H., Kim, G.B., Seo, J.B., Kim, N., 2017. Deep Learning in Medical Imaging: General Overview. *Korean Journal of Radiology*, Vol. 18, No. 4, pp.570-584.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large Scale Image Recognition. *International Conference on Learning Representations*, San Diego, USA, pp.1-15

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going Deeper with Convolutions. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., 2016. Rethinking the Inception Architecture for Computer Vision, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 2818-2826.

Web_1, 2020. Dermatoweb, http://www.dermatoweb.net (Access date: April 24, 2020)

Web_2, 2020. DermNet New Zealand. http://www.dermnetnz.org (Access date: April 24, 2020)

Web_3, 2020. DermQuest Image Library. https://www.dermquest.com/image-library (Access date: April 24, 2020)

Web_4, 2020. ImageNet. http://www.image-net.org (Access date: April 20, 2020)

Web_5, 2020. Matthew's Correlation Coefficient. https://en.wikipedia.org/wiki/Matthews_correlation_coefficient (Access date: May 1, 2020)