International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Facial Expressions Recognition using CNN

Archan Agrawal[1], *Prof A.A. Somani*[2]

[1]Research Scholar, Dept of E & Tc , MIT , Aurangabad , India
[2]Professor, Dept of E & Tc , MIT , Aurangabad , India

ABSTRACT

This paper presents the design, implementation, test, and evaluation of a Facial Expression Recognition (FER) system that applies a machine learning algorithm based on Convolutional Neural Networks (CNNs) with the aim of correctly classifying seven facial expressions (namely surprise, happiness, sadness, fear, anger, disgust, and neutral). By experimenting with pretrained AlexNet model and proposed CNN model on facial database Real-World Affective Face database (RAF-DB), the most suitable hyper parameters that yielded a good level of performance were obtained. Additionally, a deep understanding of the strengths and limitations of CNNs was gained. The training accuracy for the proposed CNN model is 72.34% which is very similar to the pretrained AlexNet CNN model which is 72.77%.

I) INTRODUCTION

Image Processing in artificial intelligence is a challenging research area in the field ofcomputer vision. Facial expressions are the fundamental way to expresses the human emotions .The applications of FER are in the field of crime control, health care, virtual education, e-commerce, entertainment, etc.Researchers are still facing problems to behave machine like humans. Latest researchers use the Deep Neural Network (DNN) Domains Especially Convolutional Neural Network (CNN). Inthis research we have implemented CNN for feature extraction and classification using Adam optimizers. We implemented facial expression recognition in real time via webcam. We also visualized the activations of hidden layers in convolutional neural network. MATLAB software is used for performing and analyzing this research.

II) BLOCK DIAGRAM

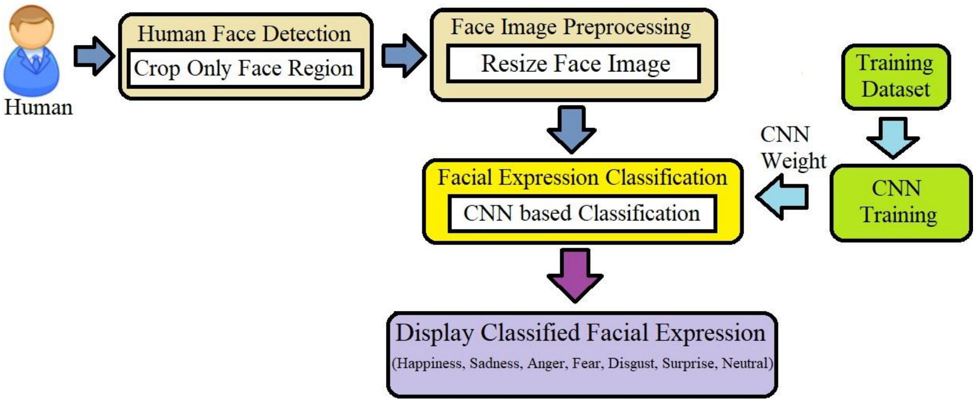


Figure 1: Basic Block Diagram of Facial Expression Recognition System.

Here the first step is to train the training dataset using Convolutional Neural Network. Save the trained weights of CNN in MATLAB. Within the CNN algorithm the trained or learned CNN weights will classify the testing dataset. Then, in real time machine acquires the human face by placing the bounding box on the face. In pre- processing step, it will resize the cropped face image similar to the size of images present in the dataset. After resizing it will classify the expression using learned CNN weights, and the output expression will be display on the real time screen. Viola-Jones algorithm is used to detect the face in real time.

III A) CONVOLUTIONAL LAYER

The convolutional layer is an important element of the network that transforms one volume of activation into another by convolving a small area with the input volume. This operation is called discrete convolution . Convolutional layers provide the network with two important features: local connectivity and parameters sharing.

Local connectivity is achieved when neurons are connected to a local region of the input volume and this local region is a hyper parameter called receptive field with dimensions$r \times r$andtheconnectionsoftheneurontothereceptivefielddimensions, but full along the entire depth of the input volume and Local connectivity drastically reduces the number of connections on a neural network, this not only diminishes the processing time, but also improves the performance of the model by preventing over fitting

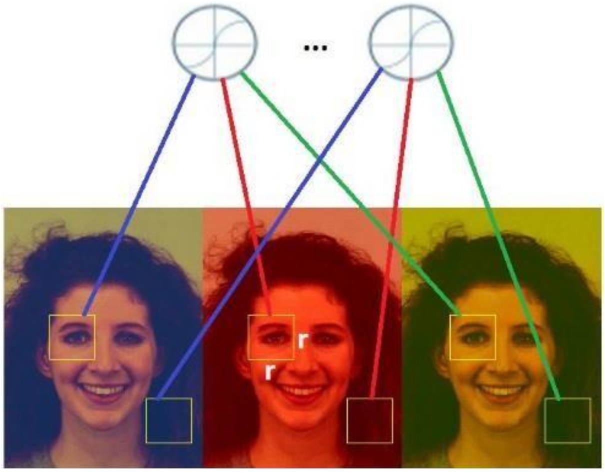


Figure 2 Receptive field and Local connectivity.

Parameter sharing occurs due to neurons within a group, called feature map, or activation map, that share the same parameters and cover different parts of the image by using different receptive fields. This provides the network with translation invariance which means that useful features that are learned in some portion of the image can be used everywhere else without independently learning those features.

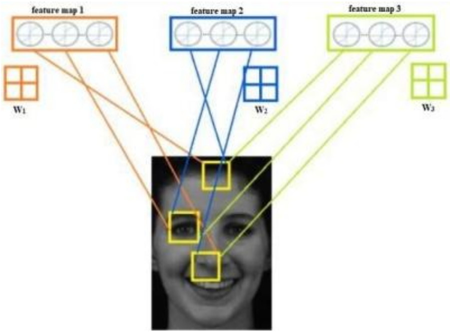


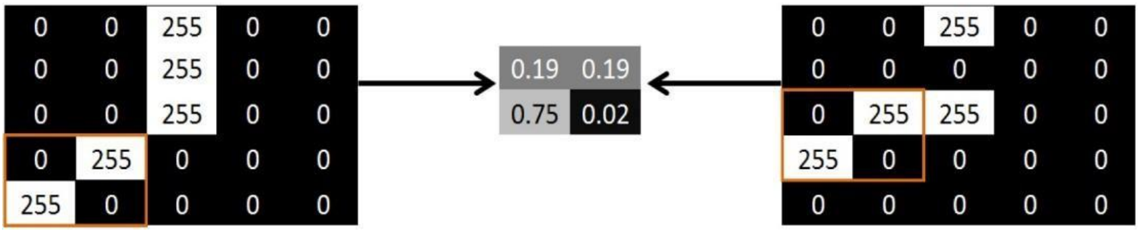
Figure 3 Parameter Sharing.

In each convolutional layer, there are filters that represent sets of weights. These filters, also known as kernels, are convolved with an input volume to generate an output volume.Eachconvolutiongeneratesanactivationmapwhichdetectsaspecifictypeof features. The number of these activation maps is proportional to the number of filters in the corresponding layer.

$$W_1 - F + 2P$$
$$Width = S + 1$$
$$H_1 - F + 2P$$
$$Height = S + 1$$
$$Depth = K$$

III B) POOLING/SUBSAMPLING LAYER

Thepooling layer reduces the spatial size of the input decreasing the number of parameters and the computation in the network as a consequence, the network gains local translation invariant is mtocontrol over fitting. The spatial size reductions achieved by taking a set of hidden units within a neighbourhood and aggregating their activations, using a pooling function (Maxpooling, average pooling, L2-normpooling or fractional maxpooling . It is worth nothing that the spatial size remains unchanged in the depth dimension.



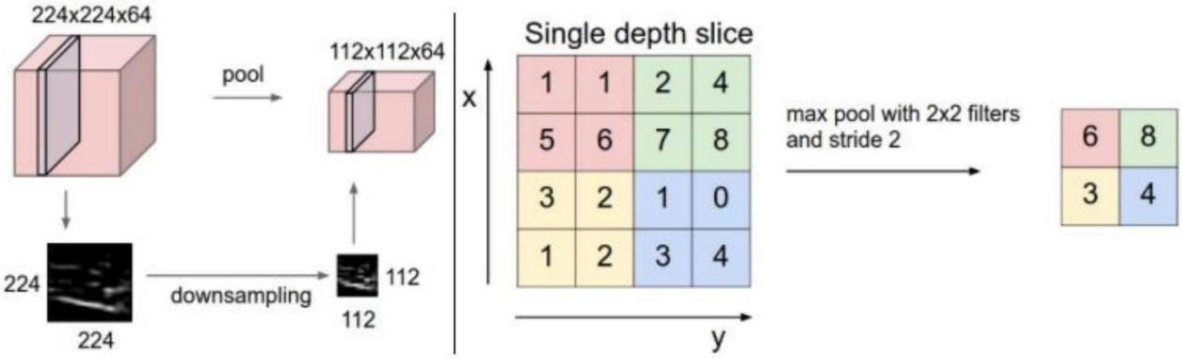
0	0	255	0	0
0	0	255	0	0
0	0	255	0	0
0	255	0	0	0
255	0	0	0	0

0.19	0.19
0.75	0.02

0	0	255	0	0
0	0	0	0	0
0	255	255	0	0
255	0	0	0	0
0	0	0	0	0

Figure 4 Local translation invariance and spatial size reduction.

There are two configurations commonly used: the overlapping and non-overlapping pooling. The former applies a 3×3 Maxpooling filter,while the latter applies a 2×2 Maxpooling filter, both with a stride of 2.



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Figure 5 Down sampling (left) and Maxpooling (right)

This convolutional layer operates independently on every depth slice of the input and resizes this input spatially using the maxpooling operation by selecting the maximum value within a local neighbourhood . In the figure above, the neighbourhood is formed by 4 elements.
The Maxpooling operation can be expressed using the following equation:

$$yijk = maxp,qxi,j+p,kq$$

Where $xijk$ is the value of the ith feature map at position, p is the vertical index in the local neighbourhood; q is the horizontal index in the local neighbourhood; and y is the result of this operation in the pooling layer
An alternative to the pooling layer is a convolutional layer with larger strides which would also reduce the spatial size. According to given the aggressive reduction in the size ofthe representation when the pooling layer is used, “the trend in the literature is towards discarding the pooling layer in modern ConvNets”.

III C) FULLY CONNECTED LAYER

After the previous layers, there can be any number of fully-connected layers. These layers are regular neural networks with neurons that have full connections to all activations in the previous layer

IV) PROPOSED CNN FOR FACIAL EXPRESSION RECOGNITION

The proposed CNN architecture is similar to AlexNet, but the batch normalization has been used heavily after each convolutional layer and contains only one fully- connected layer and one dropout layer with the probability 0.5 and the depth of the proposed CNN architecture is 5-layer depth with the fully-connected and SoftMax layer.

The Real-World Affective Faces (RAF) dataset used in this dissertation project having15339RGB images few of them are Gray Scale images with the dimensions227 x 227 for AlexNet and 100 x 100 dimension for proposed CNN, with 96 x 96 dpi, and 24-bit depth. Moreover, the purpose of their work was to recognize the same seven facial expressions recognisedin the dissertation.

The structure of the deep network shows that this model also increase the number of convolutional filters as the size of this images is reduced after each maxpooling layer.

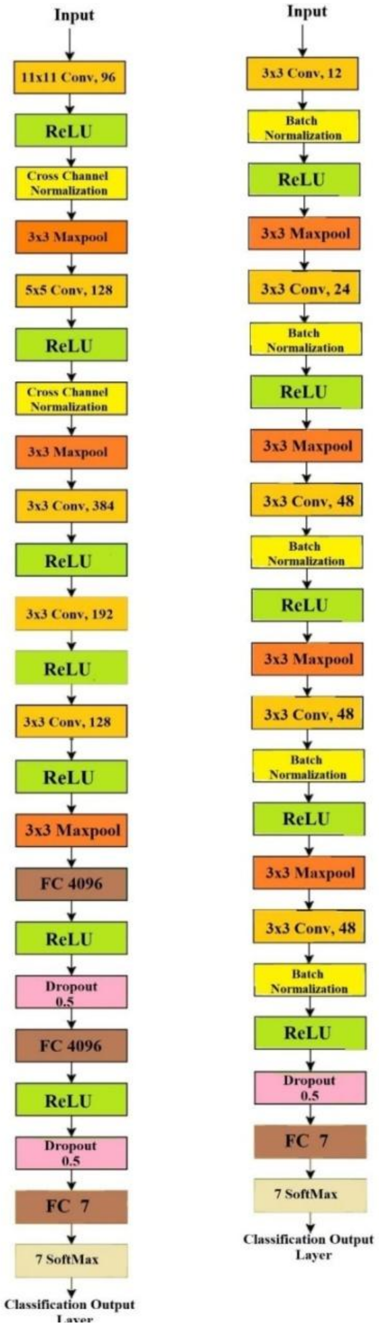
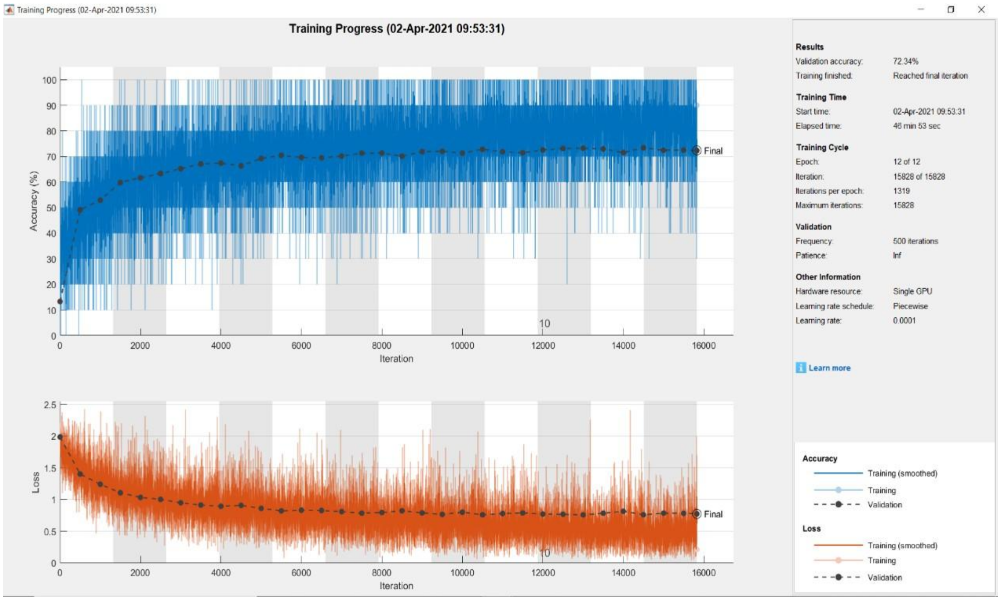


Figure 6 Architecture: AlexNet and Proposed CNN for Facial Recognition.



Results
Validation accuracy: 72.34%
Training finished: Reached final iteration

Training Time
Start time: 02-Apr-2021 09:53:31
Elapsed time: 46 min 53 sec

Training Cycle
Epoch: 12 of 12
Iteration: 15828 of 15828
Iterations per epoch: 1319
Maximum iterations: 15828

Validation
Frequency: 500 iterations
Patience: Inf

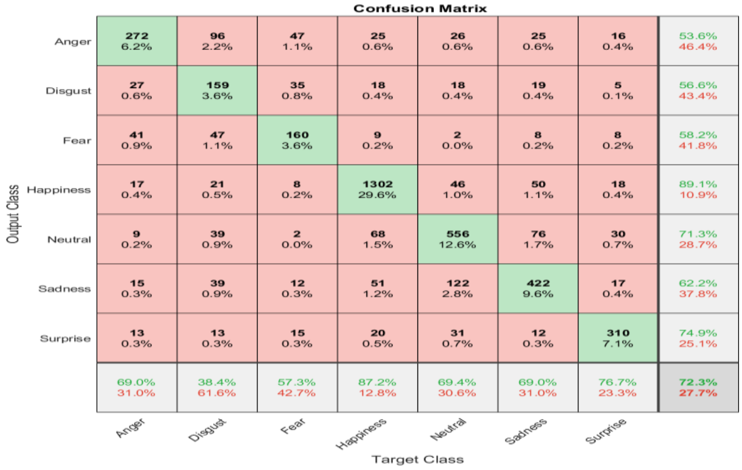
Other Information
Hardware resource: Single GPU
Learning rate schedule: Piecewise
Learning rate: 0.0001

Learn more

Accuracy
Training (smoothed)
Training
Validation

Loss
Training (smoothed)
Training
Validation

Figure 7 training progress for proposed CNN model

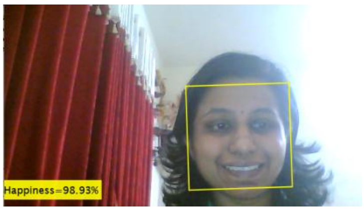

Confusion Matrix
Output Class
Target Class

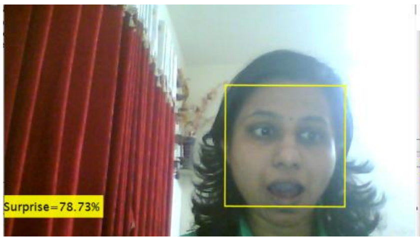
	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	
Anger	272 6.2%	96 2.2%	47 1.1%	25 0.6%	26 0.6%	25 0.6%	16 0.4%	53.6% 46.4%
Disgust	27 0.6%	159 3.6%	35 0.8%	18 0.4%	18 0.4%	19 0.4%	5 0.1%	56.6% 43.4%
Fear	41 0.9%	47 1.1%	160 3.6%	9 0.2%	2 0.0%	8 0.2%	8 0.2%	58.2% 41.8%
Happiness	17 0.4%	21 0.5%	8 0.2%	1302 29.6%	46 1.0%	50 1.1%	18 0.4%	89.1% 10.9%
Neutral	9 0.2%	39 0.9%	2 0.0%	68 1.5%	556 12.6%	76 1.7%	30 0.7%	71.3% 28.7%
Sadness	15 0.3%	39 0.9%	12 0.3%	51 1.2%	122 2.8%	422 9.6%	17 0.4%	62.2% 37.8%
Surprise	13 0.3%	13 0.3%	15 0.3%	20 0.5%	31 0.7%	12 0.3%	310 7.1%	74.9% 25.1%
	69.0% 31.0%	38.4% 61.6%	57.3% 42.7%	87.2% 12.8%	69.4% 30.6%	69.0% 31.0%	76.7% 23.3%	72.3% 27.7%

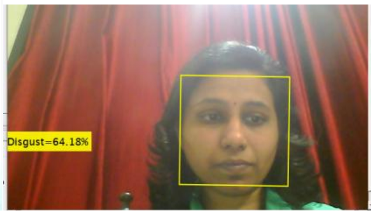
Figure 8 : Confusion matrix for the Proposed CNN model.

Figure 7 shows the training progress for proposed CNN model with the training accuracy 72.34%. Required training time is 46 min53sec, total number ofepochsused are 12 and it takes 15828 total number of iterations with 1319 iteration per epoch. Validation frequency used is 500.

V) RESULTS







To test the model in real time, we have vision cascade object detector system which uses Viola-Jones algorithm.Viola-Jones algorithm detects the facial features. We have put small theory that feed the image from webcam, preprocessed the image, then feed cropped face image to classify expression. Preprocessing step includes, cropping the face overlapped by bounding box then change the resolution of cropped face image to resolution of the images in dataset. It correctly predicts the expression when all parameters like light intensity, pose of head, distance of face from webcam should be right position. It correctly predicts the neutral, happiness and surprise expressions. Anger and fear expressions sometimes tend to mix but most of times predicts correctly. Disgust expressionis rarely predicted. Most of the time sadness expression goes wrong. While real time testing it is observed that there is no delay to detect the face and classify it. Figure 6.5, shows the real time testing model samples for correct classification of happiness expression with 77.87%, fear expression with 49.28% and surprise expression with 87.23% using MATLAB2018a.

REFERENCES

[1] Chibelushi, C. and Bourel, F. (2016). Facial Expression Recognition: A Brief TutorialOverview.

[2] Hinton, G. (2012). *Neural Networks for Machine Learning*. [online] Coursera. Available at: https://www.coursera.org/course/neuralnets [Accessed 8 Mar.2016].

[3] Y.Lv,Z. Fengand C.Xu,"Facial expression recognition via deep learning,"Smart Computing (SMARTCOMP) ,2014 International Conferenceon, Hong Kong,2014, pp. 303-308. doi:10.1109/SMARTCOMP.2014.7043872.

[4] Bishop,C.(2006).*Pattern recognition and machine learning*.NewYork: Springer, pp.227 - 249 and 256 -272.

[5] Cs231n.github.io. (2016). *CS231n Convolutional Neural Networks for Visual Recognition*. [online] Available at:http://cs231n.github.io/convolutional-networks/ [Accessed 15 Apr.2016].

[6] Murphy, K. (2012). *Machine learning*. Cambridge, Mass.: MIT Press, pp. 563 – 579.

[7] Cs231n.github.io. (2016). *CS231n Convolutional Neural Networks for Visual Recognition*. [online] Available at: http://cs231n.github.io/neural-networks-1/ [Accessed 26 Jul.2016].

[8] Poczos, B.and Singh, A. (2016). *Introduction to Machine Learning. Deep Learning*.CMU.

[9] Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A. and Bengio, Y. (2013). Maxout Networks. *JMLR WCP* 28, pp.1319-1327. arXiv:1302.4389[stat.M]