International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Facial Emotion Detection using Convolutional Neural Network

Rohit Jadhav[1], Jayesh Bhuke[2], Nita Patil[3]

[1,2]Student, Department of Computer Engineering, Datta Meghe College of Engineering, Airoli, Navi Mumbai
[3]Assistant Professor, Department of Computer Engineering, Datta Meghe College of Engineering, Airoli Navi Mumbai

---

Abstract - *Humans can easily identify emotions using their senses where as computer vision seeks to imitate human vision by analysing the digital image as input. For humans to detect an emotion will not be a difficult job to perform. Detecting emotion through voice for example detecting 'stress' in a voice by setting parameters in areas like tone, pitch, pace, volume etc in case of digital images detecting emotion just by analysing image is novel way. In this paper, we are trying to design a convolutional neural network model which can classify the input image into 7 different emotions. The respected emotions we are going to classify are angry, disgust, fear, happy, sad, surprise and neutral. In order to classify these emotions, we are implementing Convolutional Neural Networks (CNNs) that can efficiently and accurately elucidate semantic information coming from the faces in an automated manner. We also apply some data augmentation techniques in order to intercept overfitting and underfitting problems. To evaluated the proposed model, FERC-2013 dataset from kaggle competition is used to evaluate the designed CNN model. The Results of this model shows that it works better if it has more set of images to learn. The proposed model achieves the accuracy rate 65.34.*

Key Words: Convolutional Neural Networks; overfitting; underfitting.

1. INTRODUCTION

In recent years, the interaction between humans and computers has been constantly evolving, achieving a clear goal of natural interaction. The most expressive way for humans to express emotions is through facial expressions. Humans have little or no effort to detect and interpret facial and facial expressions in the scene. Still, developing an automated system to accomplish this task is still quite difficult. There are several related issues: the classification of expressions (eg, in the emotional category), detecting image segments as faces and extracting facial expression information. A system that accurately performs these operations in the real world will be an important step in achieving similar human interactions between people and machines.

Human communication conveys important information not only about intent but also about desires and emotions as well. In particular, the importance of automatically recognizing emotions from human speech and other communication cues has grown with increasing role of spoken language and gesture interfaces in human-computer interactions and computer mediated applications[8].

When machines are able to appreciate their surroundings, some sort of machine perception has been developed [1]. Humans use their senses to gain insights about their environment [5][3]. Nowadays, machines have several ways to capture their suitable algorithms allow to generate machine perception. In the last years, the use of Deep Learning algorithms has been proven to be very successful in this regard [6] [4] [2]. For instance, Jeremy Howard showed on his Brussels 2014 TEDx's talk [7] how computers trained using deep learning techniques were able to achieve some amazing tasks. These tasks include the ability to learn Chinese language, to recognize objects in images and to help on medical diagnosis.

In this paper, we are proposing deep learning methods and techniques. Convolution Neural Networks has been an effective and common way to solve the problems like facial emotion detection. CNN works better than Recurrent Neural Networks (RNN) because CNN will learn to recognize components of an image(e.g. lines, curves, etc.), but in case of RNN, it will similarly learn to recognize patterns across time. We are also using some loss functions along with data augmentation techniques in order to train model in all features of input images.

We are using FERC-2013 datasets these datasets are coming from the competition held on Kaggle in 2013 where the winner of this competition is RBM team [9]. The accuracy of public test set is 69.7% and the private test is 71.1%.

The remaining paper is categorised as follows. In section 2 we concisely summarize some related work on facial emotions. In section 3 we illustrate our proposed CNN model. The Results, Experiments are in section 4 and Conclusion is in section 5 respectively.

2. RELATED WORK

One or two different methods are used for facial emotion recognition, both of which include two different methods. Dividing the face into separate action units or further processing it as a whole seems to be the first and main difference between the main methods. In both methods, two different methods can be used, namely

"geometry based" and "appearance based" parameterization.

Using the entire frontal image and processing it, the final classification of six generic facial emotion prototypes: disgust, fear, joy, surprise, sadness, and anger; outlines the first approach. Here, it is assumed that each of the above emotions has a characteristic emotion on the face, which is why it is necessary and sufficient to identify them. Instead of using the facial images as a whole, they are divided into sub-parts for further processing, forming the main idea of the second facial emotion analysis method. Because emotions are related to subtle changes in discrete features such as eyes, eyebrows, and lipids are used to analyze automatic recognition.

These are the two main methods used in the above two methods. Geometry-based parameterization is an ancient way of tracking and processing the motion of certain spots on an image sequence, first used to identify facial emotions [14]. Cohn and Kanade later tried geometric modeling and tracking facial features, claiming that each action unit has a specific facial muscle. The disadvantage of this approach is that the contours of these features and components must be manually adjusted in this framework, with robust and difficult problems in the case of changes in posture and illumination, while tracking is applied to the image because the movements and emotions tend to Changing the shape and dynamic meaning makes it difficult to estimate the general parameters of motion and displacement. Therefore, it is too difficult to make a strong decision to make facial behavior under these changing conditions.

There are a variety of other papers that suggest alternatives to CNN, but not well. Most of these papers use Support Vector Machine (SVM) or Maximum Margin Nearest Neighbor (LMNN) for classification. The main difference between these is the function descriptor. [12] uses a system that extracts pyramids of gradient histogram (PHOG) and local phase quantization (LPQ) features for encoding shape and appearance information. [10] Features identified using AU (Action Unit), which are generated after finding pairs of patches that are important for distinguishing emotion categories. The main point of Yaos is that previous research groups ignored the importance of exploring the potential relationship between changes in facial muscle movement. This approach provides better results than the winning team in 2014, but compared to the 2015 winners, this approach is not as good as [11] focusing on facial expression recognition that is unrelated to humans, and using local Binary mode (LBP) descriptor. Many algorithms that use feature-based models are designed to mimic Ekman's recommendations for human emotions [13].

3. PROPOSED CNN MODEL

In this part, we proposed our CNN Model data flow structure for facial emotion detection problem. In this model, we take the input image size of 48 x 48 pixels. This model architecture is composed of 5 layers. These layers contains 5 convolutional layers and 5 max pooling layers along with 2 fully connected layer and at last layer we classify image through softmax activation function. The output layer consists of 7 neurons corresponding to 7 emotional labels: angry, disgust, fear, happy, sad, surprise and neutral.

A. 5-Layered CNN Architecture

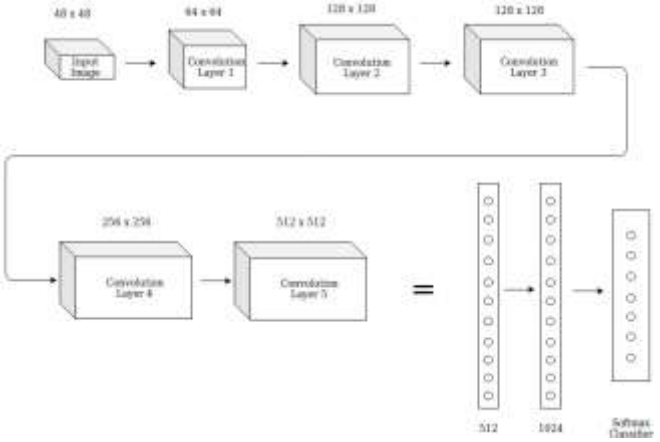

Fig. (1): 5-Layered CNN Architecture

The 5-Layered CNN architecture is represented in the below table. This architecture is composed of 5 Convolution layers and 5 Max Pooling Layers along with 2 Fully Connected Layers and the output layer. Our model uses Rectified Linear Unit (relu) as most precisely used activation function which is applied on all the Convolution Layer and Dense Layer except the last layer (output layer) which is actually Softmax Function. Dropout Layer is also applied after each Convolution, Max Pooling and Dense Layer with the rate of 0.25.

Table I Proposed Model Architecture

Name of Layer	No. of Filters / Filter Size / Rate / Neurons	Activation Function
Conv2D	**64x3x3**	**Relu**
Dropout	0.25	-
MaxPooling2D	2x2	-
Dropout	0.25	-
Conv2D	**128x5x5**	**Relu**
Dropout	0.25	-

MaxPooling2D	2x2	-
Dropout	0.25	-
Conv2D	**128x3x3**	**Relu**
Dropout	0.25	-
MaxPooling2D	2x2	-
Dropout	0.25	-
Conv2D	**256x5x5**	**Relu**
Dropout	0.25	-
MaxPooling2D	2x2	-
Dropout	0.25	-
Conv2D	**512x3x3**	**Relu**
Dropout	0.25	-
MaxPooling2D	2x2	-
Dropout	0.25	-
Dense	**512**	**Relu**
Dropout	0.25	-
Dense	**1024**	**Relu**
Dropout	0.25	-
Dense	**7**	**Softmax**

The first convolution layer composed of 64 filters with size of 3 x 3 with activation function. After each successful convolution layer max pooling layer is applied which is composed of 2 x 2 as our standard size of the pooling layer. On second layer we increased the no. of filters by 2 times than the previous convolution layer so in this layer the no. of filters are 128 with 5 x 5 filter size

Similarly, the next layers are (no of filters, filter size) (64, 3 x 3), (128, 5 x 5), (128, 3 x 3), (256, 5 x 5), (512, 3 x 3) respectively. In pooling layer, on each successful convolution layer we applied max pooling layer however, for each convolution layer we applied max pooling layer with standard size of 2 x 2 to all the 5 pooling layers in order to detect or capture every feature of the image. Before and After each pooling layer we add dropout layer to randomly select the neurons and ignore them in the next layer in order to avoid the overfitting problem. The next layer is fully connected layer or hidden layer in traditional neural network as we are using 2 FC layers including output layer. The fully connected layer consists of 512, 1024, 7 neurons respectively. The 7 neurons represents the 7 labeled classes of human emotions as mentioned earlier.

At last, we get the 7 neurons with probability distribution as our output of the model where the one with highest probability will be our final answer.

B. Convolution Network Layers and their terminologies:

1. Input Image:

Our model needs 48 x 48 x 1 input image where 48x48 is the height and width of that image and 1 is the channel (channel = 1 (grayscale) / 3 (RGB)). In our case, our model is trained on grayscale image.

2. Pooling Layer:

Pooling layer basically calculates or in fact chooses the one pixel from the specified dimensions. For example if we have 2 x 2 matrix of pixel values then it will select the value from the matrix according to type of pooling layer applied to that model. So in order to compute pooling there are two popular methods used such as max pooling and average pooling where max pooling calculates maximum value and average pooling calculates mean value respectively from the matrix.

3. Convolution Layer:

Convolution Layer basically, extract the feature from the image. The main purpose of the Convolution is to analyze the visual imagery.. In our case, we take 48 x 48 x 1 as a input image for first layer and with filter size and stride and as 64 and 3 respectively. To calculate the output layer the above formula mentioned in pooling layer is also utilize here.

$$O = \frac{(n - f + 2p)}{s} + 1$$

where O is the output of convolution layer,

n is input height/ length

f represents filter size

p represents padding parameter

s represents stride

The above are the hyper parameters of Convolution Neural Network model.

4. Fully Connected Layer:

The Fully Connected Layer is the layer perceptron which uses rectified linear unit activation function for the FC Layers and for the final layer we use softmax function. It is Basically a hidden layer from the typical neural network model.

5. Stride:

Stride is basically a window of specified dimensions which travels through the matrix (image) and calculates for the specified layers. For example in max pooling if we specify stride as 2 x 2 then it will travels through the whole matrix and calculate the maximum value from that 2 x 2 window.

6. Dropout Layer:

Dropout Layer is a method of preventing the neural network model from overfitting problems. Dropout Layer randomly selects neurons and ignore them in while training the model. In our case our model is trained on rate of 0.25.

C. Data Preparation

1. Data Pre-processing

The datasets are composed of 35887 training and testing samples. The training samples are then divided into two sets namely; Training Set and Validation Set. Training set samples composed of 80% of the original dataset samples and 20% of the samples are specified for validation. Hence Training Set and Validation Set will be 22966 and 5741 respectively. Preprocessing is performed to prepare images for the feature extraction stage. A set of facial feature points is extracted from the images then facial features derived from these points. Different sets of facial features are used for both training and validation classifiers.

2. Data Augmentation

In order to avoid overfitting and improve recognition accuracy we applied data augmentation techniques on each training and validation samples. For each image we performed following transforms:

a. Rescale (1. / 255)
b. Shear (0.2)
c. Zoom (0.2)
d. Flip (horizontal)

3. Datasets

There are various datasets available on internet which has its own experimental purposes as per their research, some of them are Cohn-Kanade (CK+), MMI Facial Expression Database, Belfast Datasets, FERC-2013 from kaggle. An appropriate and suitable datasets to choose is very important part of the given problem. So in order to get the best results for a given problem we are using FERC-2013 datasets which are coming from kaggle platform for data science. FERC-2013 contains fer2013.csv which includes 3 columns (emotion, pixels, usage). The below table represents no of samples used for each emotional class.

Table II Training Sample of Emotion Classes

Class	No. of Samples
Angry	3995
Disgust	436
Fear	4097
Happy	7215
Neutral	4830
Sad	3171
Surprise	4965

4. EXPERIMENTS, RESULTS AND EVALUATION

We train and test our model on FERC-2013 datasets which is publically available under some terms and conditions. The datasets includes 7 different emotion classes (angry, disgust, fear, happy, sad, surprise and neutral) which contains training set, private test set and public test set. The Datasets contains pixel values which can be converted to gray scale images of 48 x 48 dimensions. The images are shown in the Figure(2).

Fig. (2): Snapshots of FERC-2013 datasets

In order to increase the accuracy of the model we applied data preparation techniques on datasets as covered in the section 3.C. These preparation and all the other implementation were conducted on Kaggle Kernels which allows us to use cuda and other deep learning activities on cloud. The following are the system configuration of Kaggle Kernel Cloud Platform.

- Intel(R) Xeon(R) CPU @ 2.20GHz
- 13 GB Ram
- 16 GB Nvidia Tesla P100 Graphics
- 5 GB Disk

The following are the results of our 5 Layered Convolution Neural Network Model. First Graph represents Model accuracy and the second represents Model loss. X-axis denotes no. of epochs whereas y-axis denotes Accuracy and Loss respectively.

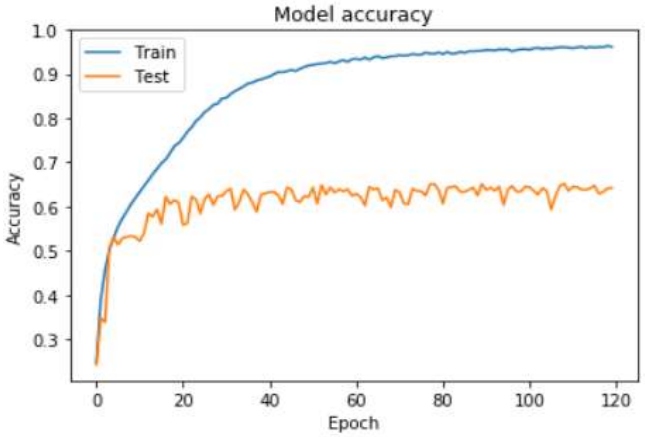

Fig. (3): Model Accuracy

In the above figure represents model accuracy. The x-axis shows the no. of epochs and y-axis shows the accuracy. The blue line represents the training curve and the orange line represents the testing curve. Since our model is trained on 120 epochs the training curve goes beyond 0.95 whereas the test curve fluctuates between 55 to 64

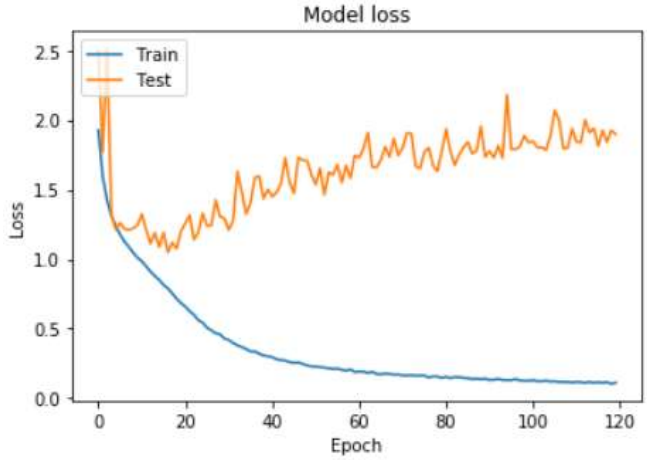

Fig. (4): Model Loss

This figure shows the Model Loss of our model. Similarly, the x-axis represents the no of epochs and y-axis represents the loss. The training curve is continuously decreasing whereas test curve is still fluctuating and as the epochs are increased.

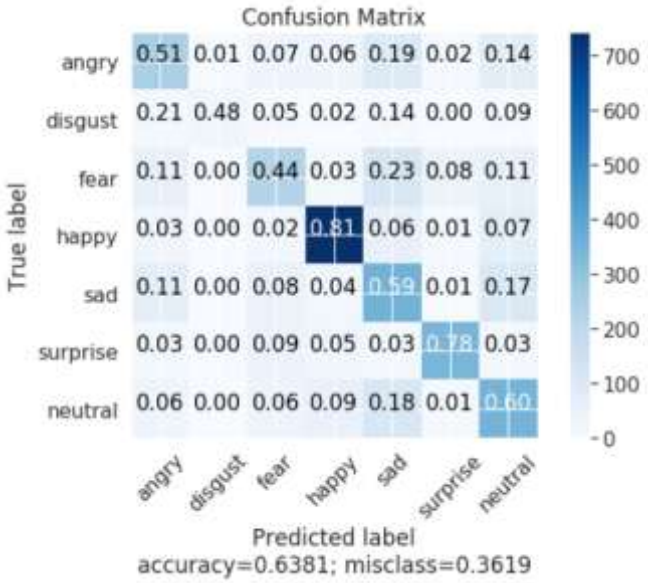

Fig. (5) Confusion Matrix of 7 emotions

In the above figure 5 represents the confusion matrix of 7 emotional classes. The results shows that class happy works most accurately among all the other classes because of its large number training samples. On the other hand, surprise, neutral, sad, angry, disgust and fear comes accordingly. The following table III represents prediction of emotional classes in decreasing order of their accuracies.

Table III Prediction of Classes

Classes	Accuracy
Happy	81%
Surprise	78%
Neutral	60%
Sad	59%
Angry	51%
Disgust	48%
Fear	44%
Average Accuracy	63.81%

5. CONCLUSIONS

In this paper, we used basic LeNet architecture of convolution neural network and is supposed to be an effective architecture among all the other models as discussed in related work. CNN performs more accurately and effectively as we used 5 convolution layers along with 5 max pooling layers followed by the 2 fully connected layers including the output layer. Experiments were conducted on FERC-2013 datasets which is brought from kaggle website. Talking about the 5 layers if we add more layers to this model along with experimental hyperparameters the accuracy will definitely increase.

Furthermore, we also apply data augmentation on the input images in order to prevent the model from overfitting problem. The confusion matrix demonstrates the precision as 63.81% and accuracy as 65%. On the other hand, results shows that our model performs pretty well on the given datasets and it will improve as it gets new set of images to train. Our Model is trained on the kaggle kernels which is quite good platform for deep learning with Nvidia Tesla P100 GPU despite of its 9 hours of session.

REFERENCES

[1] Matthew Turk. Perceptive media: machine perception and human computer interaction. CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION-23(12):1234-1244, 2000.

[2] Google Research. Research blog: A picture is worth a thousand (coherent) words: building a natural description of images.

[3] Michael Cooney. Ibm: In the next 5 years computers will learn, mimic the human senses - network world. http://www.networkworld.com/article/2162228/data-center/ibm--in-the-next-5-years-computers-will-learn--mimic-the-human-senses.

[4] Geoffrey Hinton et al. Deep Neural Networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6):82-97, 2012

[5] G.Mather. Essentials of Sensation and Perception. Foundations of Psychology. Taylor & Francis, 2014.

[6] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In advances in neural information processing systems., pages 1097-1105, 2012

[7] Howard, J. Jeremy howard: The wonderful and terrifying implications of computers that can learn.

[8] Shrikant S. Narayanan: Emotion Recognition System.

[9] Y. Tang. Deep learning using support vector machines. CoRR, abs/1306.0239, 2, 2013.

[10] Anbhag Yao, Junchao Shao, Ningning Ma, and Yurong Chen. Capturing au-aware facial features and their latent relation for emotion recognition in the wild. In Proceeding of the 2015 ACM on International Conference on Multimodal Interaction, pages 451-458. ACM, 2015

[11] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing. 27(6):803- 816, 2009

[12] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In Automatic Face & Gestures Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pages 878-883. IEEE, 2011

[13] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. Emotion in the human face: Guidelines for research and an integration of findings. Elsevier 2013.

[14] Suwa, M.; Sugie N. and Fujimora K. A Preliminary Note on Pattern Recognition of Human Emotional Expression, Proc. International Joint, Pattern Recognition, pages 408-410, 1978.