Facial Expression Recognition Using Convolutional Neural Network

Ved Agrawal
Computer Science and Engineering
Shri Ramdeobaba College of Engineering And Management, Nagpur, India
agrawalvs@rknec.edu

Harshal Dhunde
Computer Science and Engineering
Shri Ramdeobaba College of Engineering And Management, Nagpur, India
dhundehs@rknec.edu

Harsh Mata
Computer Science and Engineering
Shri Ramdeobaba College of Engineering And Management, Nagpur, India
matahd@rknec.edu

Chirag Bamb
Computer Science and Engineering
Shri Ramdeobaba College of Engineering And Management, Nagpur, India
bambcd@rknec.edu

Dr. Ramchand Hablani
Computer Science And Engineering
Shri Ramdeobaba College of Engineering And Management
Nagpur, India
hablanir@rknec.edu

***Abstract*— Facial expression recognition is important for many domains such as schools [4], hotels and surveys. Facial expression recognition is being used for many applications, such as evaluating student understanding of a subject [4], the health condition of patients in hospitals and customer satisfaction in hotels and restaurants etc. In this paper, we have designed different Convolutional Neural Networks (CNNs) for the recognition of 7 facial expressions. We have achieved 96.35 testing accuracy with CNN having three pairs of convolution and max pooling on the Ryerson Audio-visible Database of Emotional Speech and Music, consisting of seven emotion datasets.**

Keywords*— *Facial expression recognition, Convolutional Neural Network, Real-world Affective Faces Dataset, Max Pooling

I. Introduction

Facial Expressions are non-verbal forms of communication that are used to convey various types of meaning in various contexts. An essential component of human communication is the ability to generate and recognize facial expressions. The basic facial expressions are classified into 7 major categories, i.e., Anger, Happy, Sad, Neutral, Fear, Disgust, and Surprise.

Facial expression recognition improves human-computer interaction, communication, emotional comprehension, healthcare diagnostics, security, and marketing. Facial expression recognition plays a significant role in various fields such as education, healthcare, and marketing[4][5].

Facial expression recognition has substantially benefited from the introduction of deep learning techniques, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)[2][5]. Such methods have proved to extract facial features efficiently which makes our systems more efficient and reliable.

In this research, we designed various convolutional neural network architectures. We also analyse the changes in results by changing various hyperparameters like batch sizes, epochs, and different combinations of layers. This analysis helps us to understand the effect of various hyperparameters while training facial expressions recognition models.

We have used two well-known datasets, i.e., FER-2013, and RAF-DB. Since the images of these datasets have different image sizes, and colours, we get a broad picture of how these affect our results.

II. Literature Review

W. Mellouka, W. Handouzi, 2020 [2] The different infrastructures can achieve better results in this system. The reason for this article is to study facial expressions using deep literacy and create automatic facial expression recognition models. This paper is useful for getting insights and reviews. This paper lacks the deep learning aspects such as batch size, epoch, etc. We are trying to add these aspects to our insights. In a paper by S. Abdullah, A. Abdulazeez,2021 [3], In this paper the author focuses on accuracy and speed of computation. by researching different architecture they create architecture. Facial expressions are a form of verbal communication. This article gives a brief study on various systems of FER. It helped us to create the architecture of our model.But this paper also lacks the deep learning aspects such as batch size, epoch, etc [2].

In a paper by E. Komagal and B. Yogameena, 2020 [4], During the pandemic many teaching institutes shifted to online teaching mode, so there is a need for effective analysis. But Facial expression recognition is useful for both online and

offline analysis of student engagement. Although offline it is difficult and there are many parameters which hinders results. Since it can be fluently acclimated to the online terrain. This paper works on real-time facial expression recognition. It lacks dataset size although they created their own dataset. We want to work and improve real-time facial expression recognition. It is a future plan. In [5], This paperwork on FER methods based on deep mastering and synthetic intelligence (AI) strategies to ensure performance and Realtime processing.They studied datasets in-depth and gave various issues and problems faced by them and troubles are mentioned. Unique measures to evaluate the overall performance of FER methods are comprehensively mentioned. In this paper, they have greatly explained the limitations of the dataset and methods. They have created their own solutions for the problems mentioned which helped us to resolve many problems. But it lacks deep learning aspects the same as [2]. We are working on lacking deep learning aspects. In [6], Conventional strategies in particular examine prototypical facial expressions of 8 different expressions are classified.. In this paper, by using CNNs various methods and models are proposed. The proposed method not only classifies the expression but also can give the intensity of feeling as we feel. In Particular, we first mapped facial expressions into dimensional measures to convert facial expression evaluation from a category problem to a regression one. This paper helped us resolve architectural issues and issues with similar expressions.

III. DATASETS

FER-2013: FER-2013, also called the "facial expression recognition - 2013" dataset, is a broadly used benchmark dataset on the subject of facial features popularity research. The dataset was created by using the international conference on automated Face and Gesture (FG) and consists of 35,887 grayscale snapshots of faces. Each image in FER-2013 is classified with one of the seven emotion training: anger, disgust, fear, happiness, sadness, surprise, and neutral. The images in FER-2013 have dimensions of 48 pixels by using 48 pixels. Those grayscale pictures capture facial expressions with varying stages of depth and subtlety. The dataset became usually accumulated from the internet, incorporating photos from various sources, ensuing in a sensible representation of facial expressions encountered in real-global situations.Researchers often make use of FER-2013 as a trendy dataset for schooling and comparing deep learning fashions in facial feature recognition. Its vast adoption allows significant comparisons and enables advancements in the discipline. Though, due to the dataset's internet-derived nature, there can be versions in image high-quality and noise, which need to be taken into consideration whilst interpreting results and designing algorithms. FER-2013 Dataset’s Sample images are given in Fig.1

Fig .1 FER-2013 Dataset

RAF-DB: RAF-DB, additionally referred to as the "Ryerson Audio-visible Database of Emotional Speech and Music," is a broadly used dataset within the field of affective computing and emotion popularity. It was advanced with the aid of Ryerson University and consists of audiovisual recordings taking pictures of emotional expressions in speech and music performances. The RAF-DB dataset comprises 12,271 movies, where each clip represents a single instance of an emotional expression. The dataset consists of 8 emotion classes: anger, disgust, fear, happiness, neutral, disappointment, surprise, and different. Similarly to the emotional labels, RAF-DB affords annotations for facial landmarks and motion devices, taking into consideration extra exact analysis of facial expressions. The motion pictures in RAF-DB have dimensions of 720 pixels with the aid of 576 pixels. While RAF-DB offers a complete series of emotional expressions, it is critical to consider a few barriers. The dataset more often than not specialises in posed expressions rather than spontaneous emotions, which may also impact the generalizability of the findings to real-global eventualities. Moreover, because of its especially smaller length as compared to other datasets, the sample diversity and statistical distribution of feelings might be extra confined. In the end, RAF-DB offers a treasured resource for investigating emotional expressions in speech and track performances. With its dimensions of 720 pixels by 576 pixels and annotations for facial landmarks and action gadgets, the dataset offers a wealthy and specified illustration of emotional expressions. Researchers keep leveraging RAF-DB to enhance the sector of affective computing and multimodal analysis, striving to improve emotion reputation algorithms and deepen our know-how of human emotions.RAF-DB Dataset’s Sample images are given in Fig.2

Fig.2 Seven emotion dataset RAF-DB

IV. Proposed Method

In This Project firstly we studied the various processes of building a model like collecting the dataset, pre-processing the dataset, training the model, testing the model and then finally deploying the model. while developing this model we also faced some major challenges like we all know that facial expressions highly depend on the individual [1][2][5], face cut, and culture so it was important for us to choose a dataset that is diverse enough to detect all types of diverse facial expressions also image quality, face blocked by glasses, hair was the major issue every cnn architecture consist of 4 layers -

Con2D layer: A Conv2D layer is a type of convolutional layer used in convolutional neural networks (CNNs). It takes a dot product between two matrices, one of which is the restricted area of the receptive field and the other is the set of learnable parameters also known as a kernel or filter. The filter traverses over the input data throughout the convolution operation, and the dot product between the filter and the input data is calculated at each place. This produces a feature map that summarises the presence of detected features in the input.

MaxPooling2D layer: A MaxPooling2D layer is a type of pooling layer used in convolutional neural networks (CNNs). It performs a downsampling operation along the spatial dimensions of the input data by taking the maximum value over a sliding window of a specified size. This operation reduces the dimensionality of the input data while retaining the most important information.

Flatten Layer: A Flatten layer is used in convolutional neural networks (CNNs) to convert the multi-dimensional output of the previous layer into a one-dimensional vector. This is necessary because the output of convolutional and pooling layers is typically multi-dimensional, while the input to fully connected layers (also known as dense layers) is one-dimensional.

Dense Layer: A fully connected layer, usually referred to as a dense layer, is a type of layer used in neural networks where each neuron is connected to every neuron in the previous layer. In a convolutional neural network (CNN), Dense layers are typically used at the final stage of the network to perform classification.the params and outputs are calculated Using the below formulas

A. Equations

$$N_{out} = \lfloor \frac{N_{in}+2p-k}{S} \rfloor + 1$$

$$j_{out} = j_{in} \times S$$

$$r_{out} = r_{in} + (k + 1) \times j_{in}$$

$$cen_{out} = cen_{in} + \left(\left(\frac{k-1}{2}\right) - p\right) \times j_{in}$$

N_{in} : Number of input features

N_{out}: Number of output features

j_{in}: Distance between two consecutive input features

j_{out}: Distance between two consecutive output features

S: Convolutional stride size

k: Convolutional kernel size

cen_{in}: Input centre coordinate of first feature

cen_{out}: Output centre coordinate of first feature

p: Convolutional padding size

r_{in}: Input receptive field size

r_{out}: Output receptive field size

The params and outputs are calculated Using the above formulas.

So by using the above-mentioned four layers which are Con2D, MaxPooling, Flatten and Dense we built our own Convolutional Neural Network (CNN) architecture as shown in figure 3.

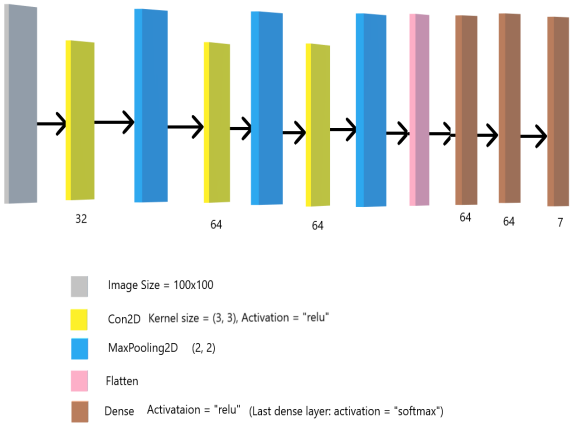

Fig .3 Layers of CNN Architecture

V. Experiments And Results

Working with FER dataset -

Initially we have started our work with the FER dataset, and built a Convolutional Neural Network (CNN) architecture consisting of four pairs of convolution and max pooling followed by one flatten and three dense layers having relu as the activation function for all the layers and softmax as the activation function for the last dense layer, and train the model using number of epochs as 50, batch size as 32 which results in training accuracy of 92, validation accuracy of 84 and testing accuracy of 84.85 as shown in the Table 1.

Epochs	Training Accuracy(%)	Validation Accuracy(%)	Testing Accuracy(%)
50	92	84	84.85

Table. 1 Accuracies of Model on FER dataset

Confusion matrix for the model trained using FER dataset is drawn in Table 2

True/Predicted	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	321	2	16	8	19	28	4
Disgust	1	32	0	1	1	1	0
Fear	27	2	303	12	24	32	25
Happy	6	0	10	656	25	20	10
Neutral	15	0	13	6	460	26	1
Sad	12	1	12	7	44	389	3
Surprise	6	0	10	7	0	4	310

Table. 2 Confusion Matrix for FER dataset

Working with RAF dataset -

Firstly we have built a Convolutional Neural Network (CNN) architecture consisting of four pairs of convolution and max pooling followed by one flatten and three dense layers having relu as the activation function for all the layers and softmax as the activation function for the last dense layer. On this architecture we have varied the Batch Size and have achieved the following accuracy.

Batch-Size	Training Accuracy(%)	Validation Accuracy(%)	Testing Accuracy(%)
32	97.99	92.64	93.50
64	98.52	95.77	94.90
128	97.80	93.29	95.00
256	97.00	92.84	93.80

Table. 3 Accuracies of Model on Varied Batch-Size

We have taken different batch sizes such as 32, 64, 128, 256 and trained the model having four pairs of convolution and max pooling, and got the accuracy as mentioned in the Table. 3.

From the above analysis we were able to conclude that with batch size as 128 our model gives the best testing accuracy among the others that is 95.

Taking the result into consideration from Table 1 we have taken the batch size as 128, and now we have varied the pairs of convolution and max pooling of the CNN architecture.

Firstly we have added a pair of layer consisting of convolution and max pooling to the CNN architecture which results in five pair of convolution and max pooling and then train the model on this new architecture which results in a training accuracy of 98, validation accuracy of 92 and testing accuracy of 93 as shown in the Table. 4.

Epochs	Batch Size	Training Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
50	128	98	92	93

Table. 4 Accuracies on 5 pair convolution and max pooling architecture

Then we have also removed a pair of layer consisting of convolution and max pooling from the CNN architecture as shown in figure 3 which results in three pair of convolution and max pooling and then train the model on this new architecture which results in a training accuracy of 100 ,

validation accuracy of 95.05 and testing accuracy of 96.35 as shown in the Table 5.

Epochs	Training Accuracy(%)	Validation Accuracy(%)	Testing Accuracy(%)
50	100	95.05	96.3

Table. 5 Accuracies on 3 pair convolution and max pooling architecture

We have a batch size as 128 and now taking the result into consideration from Table 4 and Table 5 based on testing accuracy, we have taken the CNN architecture having three pairs of convolution and max pooling as it gives a better testing accuracy as compared to the other. And now have varied the number of epochs of three pairs of convolution and max pooling CNN architecture.

Epochs	Training Accuracy (%)	Validation Accuracy (%)	Testing Accuracy (%)
15	0.93	0.90	0.91
25	0.96	0.92	0.92
35	0.98	0.94	0.95
50	1.00	0.95	0.96

Table. 6 Accuracies of Model on Varied Epochs

We have taken different epochs such as 15, 25, 35, 50 and trained the model having three pairs of convolution and max pooling, and got the accuracy as mentioned in Table. 6.
From the above analysis from Table. 6 we were able to conclude that with epoch as 50 our model gives the best testing accuracy among the others that is 96.

Confusion matrix -

True/Predicted	Fear	Surprise	Disgust	Happy	Sad	Anger	Neutral
Fear	109	5	3	9	10	2	18
Surprise	6	15	1	8	6	4	0
Disgust	2	0	37	10	8	9	13
Happy	7	2	4	515	22	2	25
Sad	6	4	8	20	177	8	40
Anger	4	6	7	8	6	49	12
Neutral	15	0	19	16	33	3	243

Table. 7 Confusion Matrix for RAF dataset

So finally after varying different parameters such as batch size, epoch and layers we were able to conclude that the architecture consisting of three pairs of convolution and max pooling followed by one flatten and three dense layers having relu as the activation function for all the layers and softmax as the activation function for the last dense layer with batch size as 128 and the number of epoch as 50 gives the best result taking into account the testing accuracy of the model which gives training accuracy of 100, validation accuracy of 95.05 and testing accuracy of 96.35.
A confusion matrix for the same architecture is drawn in the Table. 7.

VI. CONCLUSION

We have successfully achieved the objective of building a model and implementing an accurate model for the recognition of facial expressions, in which we have used machine learning techniques including data collection, preprocessing, feature extraction, model training, and evaluation.
The CNN-based model gives a better performance for the RAF dataset, achieving high accuracy in facial expression recognition.
We were able to detect the facial expressions with the maximum testing accuracy of 96.35 on three pairs of convolution and max pooling architecture with 128 batch size with training accuracy of 100, and validation accuracy of 95 after encountering various challenges as mentioned above.
We will extend our work to improve performance on new datasets and real-time implementation.

REFERENCES

[1] Shan Li and Weihong Deng* , Member, IEEE, “Deep Facial Expression Recognition”, 2018

[2] Wafa Mellouka*, Wahida Handouzia, “Facial emotion recognition using deep learning: review and insights”, 2020 Journals -This is an open access article under the CC BY-NC-ND licence

[3] Sharmeen M Saleem Abdullah, Adnan Mohsin Abdulazeez. “Facial Expression Recognition Based on Deep Learning Convolution Neural Network: A Review”, 2021 Journals - Journal of Soft Computing and Data Mining

[4] E. Komagal and B. Yogameena, “PTZ-Camera-Based Facial Expression Analysis using Faster R-CNN for Student Engagement Recognition”, 2023

[5] Muhammad Sajjad, Fath U Min Ullah, Mohib Ullah, Georgia Christodoulou, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad Joel J.P.C. Rodrigues, “A Comprehensive Survey on Deep Facial Expression Recognition: Challenges, Applications, and Future Guidelines”, 2023

[6] Feng Zhou, Shu Kong, Charless C Fowlkes, Tao Chen, Baiying Lei “Fine-Grained Facial Expression Analysis Using Dimensional Emotion Model”, 2020