ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 5 May 2024

A Survey on Facial Expression Recognition Using Convolutional Neural Network

[1] Vaibhav Govindwar, [2] Aman Akbani, [3] Aishwarya Wanjari, [4] Aachal Nadeshwar, [5] Prachi Aghase, [6] C R Pote

[1] [2] [3 [4] [6] Department of Computer Technology, Priyadarshini College of Engineering, Nagpur
[5] Computer Technology, Priyadarshini College of Engineering, Nagpur
Email: [1] govindwarvaibhav@gmail.com, [2] amanakbani786@gmail.com,
[3] aishwaryawanjari264@gmail.com, [4] sweetynandeshwar58@gmail.com, [5] agasheprachi33@gmail.com,
[6] scrpote@gmail.com

Abstract*— Understanding others' intentions through nonverbal cues like facial emotions is crucial in human communication. To design and train Deep Learning Models, this paper describes in detail how Convolutional Neural Network Models are developed using tf. Keras. The aim is to Sort facial photos into one of the seven face detection classifiers, our model is developed in such a manner that it learns hidden nonlinearity from the input facial images, which is critical for discriminating the type of emotion a person is expressing. The model proposed on the Lenet-5 architecture by Yann LeCun uses the subsampling, feature map, and activation function (ReLu) in between the convolutional layer and fully connected layer for the output soft-max activation function will be used. The FER-2013 dataset, which consists of 35,887 structured 48x48 pixel grayscale images, was used to train the CNN models. The training dataset has 28,709 elements, while the testing dataset has 7,178 elements. Train and test are the two folder names used to organize the FER dataset. separated even further into distinct files, each holding a different kind of FER dataset class. To mitigate the overfitting of the dropout, batch normalization and the model are employed. Since this is a multiclass classification problem, we are utilizing the Soft-max activation function and the Rectified linear unit for non-linear operation (ReLu). We are training a categorical cross-entropy and matrix for accuracy based on the parameters to assess the constructed CNN model's performance by examining the training epoch history. [13].

***Index Terms*— Deep Learning, CNN, LeNet-5, FER-2013.**

I. INTRODUCTION

Lung disease is common Understanding others' intentions through nonverbal cues like facial emotions is crucial in human communication. People often read other people's facial expressions to interfere their emotional states, such as happiness, sadness, disgust, surprise, neutral or anger by reading their facial expressions. Additionally, it makes use of machine learning (ML) and computer vision algorithms to identify and categorize distinct face patterns and attributes connected to a range of expressions. Recently, convolutional neural networks (CNNs) have demonstrated promise in the field of object categorization, and this ability extends to the difficulty of recognizing facial expressions [4].

Artificial neural networks with one or more convolution layers are called convolutional neural networks (CNNs), and they are primarily utilized for image processing, classification, segmentation, and other tasks involving autocorrelated data. CNN's capacity to create an internal representation of a two-dimensional image is a benefit. This enables the model to pick up on the scale and location of faces in an image. CNN can identify a face in an image when it has been trained.

The study of facial expression recognition using convolution neural networks (CNN) is being proposed and aims to classify the expression of different faces. The model based on LeNet-5 and the algorithm of CNNs is used to train with the FER2013 dataset. In the LeNet-5 model, we are using the 3 convolutional layers with the ReLu, filters, and the 2 max-pooling layers and the fully connected layer with activation functions i.e., soft max for a probability distribution.[5].

II. LITERATURE REVIEW

[Hai-Duong, et-al., 2018] have proposed to 18-layer of CNN inspired by VGG net and the activation function ReLu. The MLCNN architecture that automatically selects high-level feature for classification and takes mid-level features. This feature map originates from the second, third, and fourth blocks of the network, whereas the initial block specifically incorporates filters designed for recognizing facial expressions. Trained the data from the FER2013 dataset in 48x48 pixels in gray-scale images. The accuracy of MLCNN is 73.03%.

[Joshua. G., et al., 2019] have proposed a model using CNN and HOG classifier to improve facial expression recognition. CNN is a deep learning system that can recognize between different aspects of a picture and learn its features. FER research studies are carried out to address problems related to orientation and varying light conditions. CNN is used to identify and categorize the expressions and the HOG classifier is used to extract features. With an optimization result of 77.2% and an overall high accuracy, this approach provided greater accuracy than using SVM algorithm and HOG classifier with accuracy of 55%.

[S. Marry Hima Preethi, et al., 2020] have put forward a model that can identify the expression on an individual's face and categorize the emotion on it. It requires multiple phases and is constructed using a convolutional neural network. This CNN model's performance is assessed using a dataset FER2013 consisting of 35,887 structured, 48x48 pixels of Gray-scale images. By introducing non-linearity and utilizing hierarchical anti-overfitting techniques like batch normalization and dropout, deep networking has helped to lessen the overfitting tendency of the learning model. The model's accuracy has the range between 60% to 70%.

[J. Bodapati, et al., 2021] have proposed a novel deep learning based strategy to address the challenges of facial expression recognition from the image. The deep convolutional neural network model consists of multiple convolutional layer and sub-sampling layer. The model is trained on FER2013 dataset. The task is to achieve the accuracy of around 69.57%.

[Akash Kumar, et al., 2021] have proposed the architecture of The CNN model has eight layers total, the first five of which are max pooling layers through the ReLu function and convolutional layers for model learning. The soft-max classifier, which is the last eighth layer, is used for either image recognition or classification. An initial attempt The VGG model implements CNN with a fixed depth of five convolutional layers. The model was trained utilizing the architecture. Have achieved an accuracy of 53% .

III. SYSTEM OVERVIEW

The proposed model is implemented by providing input image through the Python library. This image comes under the image preprocessing. this goes under preprocessing, the various steps like Resize image, image augmentation, and image normalization. Neural networks, like neurons in the human brain, operate through multiple layers to process data. They typically consist of three layers: input layer, hidden layers, and output layers. In this model, we use the FER2013 dataset, which is a CSV file consisting of pixel values and labels for each image. In the feature extraction, we are using the LeNet-5 model in the CNN algorithm for expression detection and the fully connected layer comes under the classification where 84 neurons are there in that layer, and for the output, the activation function Soft-Max is used to convert a vector of numbers into vectors of probability. The output layer is the same as the Soft max layer, which has seven neutrons and is mostly used to anticipate the seven different ways that human face emotions can be expressed [10]

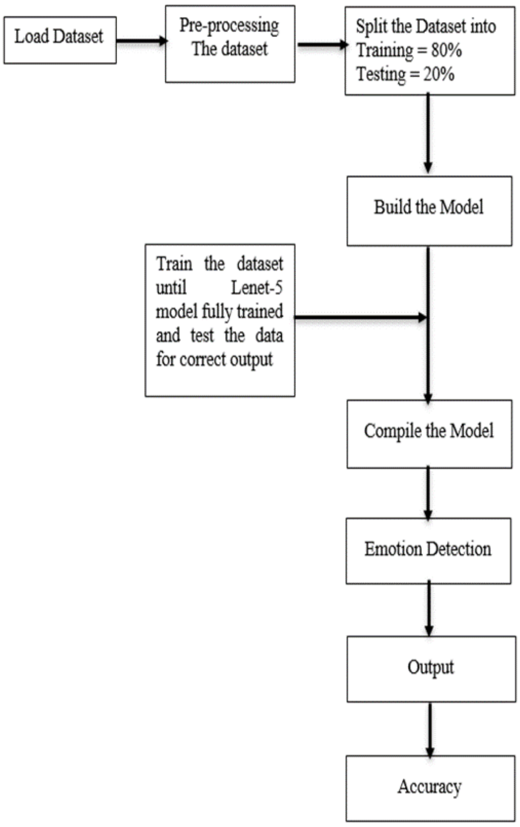

Load Dataset
Pre-processing The dataset
Split the Dataset into Training = 80% Testing = 20%
Build the Model
Train the dataset until Lenet-5 model fully trained and test the data for correct output
Compile the Model
Emotion Detection
Output
Accuracy

Fig. 1. System overview Architecture

IV. DATASET

The FER2013 dataset, which is approximately 35,887 in size and comprises a training dataset of 28,709 and a testing dataset of 7,178 well-structured 48x48 pixel grayscale photos of faces with various expressions, is used in this proposed model. Each image needs to be categorized into one of the seven classes, each representing a specific facial emotion. These expressions fall within the following categories: 1. Surprise, 2. Fear, 3. Angry, 4. Neutral, 5. Sad, 6. Disgust, and 7. Happy. Our primary method for classifying the expressions is to use the characteristics that convolution layers provide from the raw pixel data. [16]

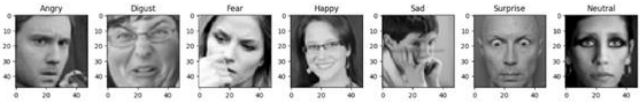


Fig. 2. Sample dataset FER 2013

V. ANALYSIS

a. Image preprocessing:

The original images of the dataset are gray scale and resized to the shape of 48X48. Data augmentation is very beneficial if works with limited data sets because it prevents the model from doing so overfitting. For multiple expressions, the number of images is weak and is therefore increased by throwing and spinning of them. Resized images are flipped and rotated horizontally randomly to generate different versions of the original images. Additionally, the data is normalized to adjust to the same scale. Pixels range in value from 0 to 255 in every image. Normalization requires us to change the range from 0 to 1 before sending it to the model. [13].

b. Convolutional Neural Network:

One of the Deep Learning architectures that is most frequently used for image recognition and classification is CNN (Convolutional Neural Networks). Convolutional neural networks excel in extracting image features and recognizing patterns, rendering them ideal for tasks such as object detection, image segmentation, and classification. Comprising input, pooling, output, fully connected, and convolutional layers, these networks employ filters in convolutional layers to extract features from input images. To minimize computation, the Pooling layer down samples the image. Finally, the fully connected layer generates the final prediction [3].

c. LeNet-5 Architecture:

LeNet-5 is a convolutional neural network (CNN) architecture. This architecture was Developed by Yann LeCun and his colleagues in the 1990s. The convolution-pooling alternation of the LeNet-5, a classical CNN model, is a special structure that may be successfully extracted to translate the input image's invariant properties. The LeNet -5 architecture has 5 learnable layers with Parameters hence named LeNet-5. There are three sets of convolutional layers with maximum pooling combination. Then we have two fully connected layers with the convolution and max-pooling layers. Finally, a SoftMax classifier that classifies the images into corresponding classes. In summary, the performance of these three models is outstanding. As previously mentioned, the benefits of each model that combined to create a fused CNN model [1]

$$\left(\frac{n+2p-f}{s}+1 \quad * \quad \frac{n+2p-f}{s}+1\right)*F$$

Where n is input size, p is padding, f is kernel size, s is stride if stride is 1 then its 1 and if stride is 2 then its 2 and F is kernel type.

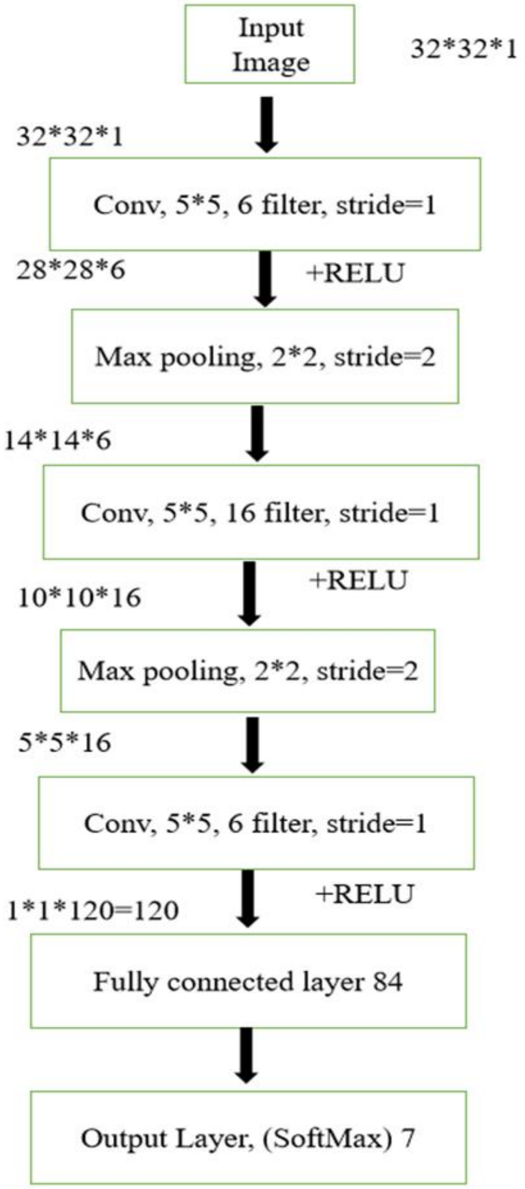

Input Image
32*32*1
32*32*1
Conv, 5*5, 6 filter, stride=1
28*28*6
+RELU
Max pooling, 2*2, stride=2
14*14*6
Conv, 5*5, 16 filter, stride=1
10*10*16
+RELU
Max pooling, 2*2, stride=2
5*5*16
Conv, 5*5, 6 filter, stride=1
1*1*120=120
+RELU
Fully connected layer 84
Output Layer, (SoftMax) 7

Fig. 3. LeNet-5 architecture

d. Activation Function:

In this model, we are using two activation functions. Generally, the activation function is used in the dense layer or output layer, but to extract the feature we use the activation function after the convolutional and before the pooling layer [14].

e. Rectified Linear Unit (ReLu)

The activation function that deep learning models most frequently use is the Rectified Linear Unit (ReLu). If the function receives a negative input, it returns 0; alternatively, it returns a positive value x. The formula for it is:

$$\mathrm{f(x)} = \max(0, \mathrm{x}).$$

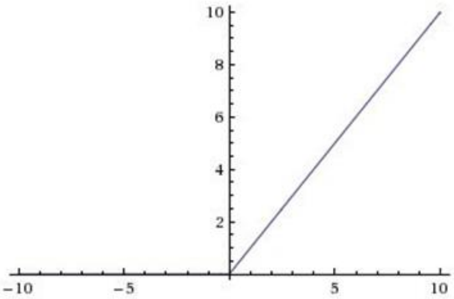


Fig. 4. Graphical Representation of ReLu

f. Soft-Max

The Soft-Max function is designed to convert a vector of numbers into a vector of probabilities. The probabilities connected with each value in the vector are directly proportional to their respective scales.

$$f(x)_i = \frac{e^{xi}}{\sum_{j=1}^{K} e^{xj}}$$

where '*x*' is a vector of the previous layer's outputs, *'K'* is the vector length, *'j'* is the index of a vector element, and *'i'* is the index of the corresponding soft-max output element

g. Stride:

The stride is the rate at which the filters move horizontally and vertically across the input pixel picture during the convolution. A stride refers to the number of pixels shifts in the input matrix. When stride is 1, we shift the filters by one pixel. When stride is 2, we shift the filters two pixels at a time, and so on. The figure below illustrates how convolutional would work with a stride of two. [6]

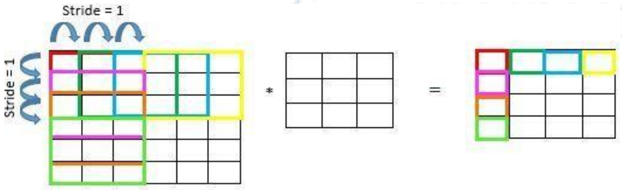
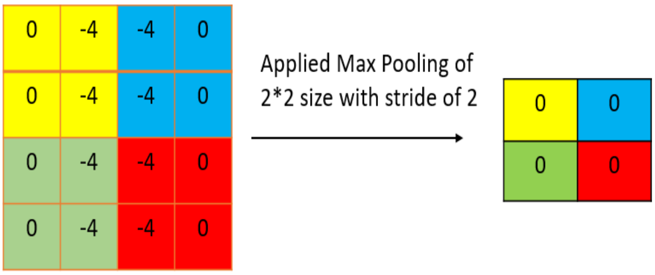


Fig. 5. Stride during convolutional

h. Polling layer:

Before the convolutional layer, the pooling layer is typically implemented, and it is used to reduce the dimensions of feature maps which helps in preserving important information or properties of the input image and shortens the calculation time. Here we use a pooling layer of size 2*2 with a step of 2. The maximum aa value is taken from each highlighted area a new 2*2 version of the input image is obtained so after using pooling it has dimension of feature map reduced. The most common the pooling types are maximum pooling and average pooling. The image below shows how Max Pooling works. Using the feature map, we got from the example above, use pooling. [9]



0	-4	-4	0
0	-4	-4	0
0	-4	-4	0
0	-4	-4	0

0	0
0	0

Fig. 6. Polling layer in convolutional

i. Fully connected layer:

The input image is categorized into a label with the fully connected layer. This layer classifies the input to the appropriate designation by attaching data derived from previous stages (i.e. pooling layers and convolutional layers) output layer. The feature map matrix will be transformed to a vector. We put all these features together to create a model using fully connected layers. Finally, to classify the outputs, we have an activation function with a soft max [11].

Summary:

A convolutional neural network algorithm is the source Forward Network for image data processing and recognition using grid version. As one of the neural networks, it also consists of an input layer where the image is inserted, followed by a series of layers including convolutional layers, ReLu layers, pooling layers, and fully connected layers, culminating in a Soft-max layer.

Convolution Layer – Converts an image into an array, and the size of the feature map is determined by three variables

Depth: The term "depth" specifies the number of filters that are used in the process of convolution.

Stride - The number of pixels the filter matrix moves through the input matrix.

Padding - It is a good to input matrices with zero around the boundary matrix.

VI. CONCLUSION

We have studied numerous literature survey papers on Facial recognition using CNN. Considered a modified model for Facial Expression Recognition utilizing a Convolutional Neural Network, the CNN-based LeNet-5 architecture was developed. The CNN neural network is composed of three layers: the convolutional layer, the pooling layer, and the fully connected layer. In this, we have 3 convolutional layers and 2 pooling, in between the activation function ReLu has been used. To train the dataset, we have used the FER2013 dataset which is in a Gray-scale image. To get better accuracy than all previous modes, we are developing our model using LeNet-5 architecture with soft-max for the output layer. This Network can work for various fields of Applications like Medical Diagnosis, Business Analysis, surveillance, Security purposes, and many more.

REFERENCES

[1] YANN LECUN, MEMBER, IEEE, L´ EON BOTTOU, YOSHUA BENGIO, AND PATRICK HAFFNER, Gradient-Based Learning Applied to Document Recognition, PROCEEDINGS OF THE IEEE, VOL. 86, NO. 11, NOVEMBER 1998

[2] Shima Alizadeh, Azar Fazel. Convolutional Neural Networks for Facial Expression Recognition, Stanford University, 2016

[3] Raghuvanshi, Arushi, and Vivek Choksi. "Facial Expression Recognition with Convolutional Neural Networks." Stanford University, 2016

[4] Alizadeh, Shima, and Azar Fazel. "Convolutional Neural Networks for Facial Expression Recognition." Stanford University, 2016

[5] Diah Anggraeni, Pitalokaa,, Ajeng Wulandaria, T. Basaruddina, Dewi Yanti Lilianaa. Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition. 2nd International Conference on Computer Science and Computational Intelligence 2017, ICCSCI 2017, 13-14 October 2017.

[6] Chu, William Wei-Jen Tsai, Hui-Chuan, YuhMin Chen and Min-Ju Liao. Facial expression recognition with transition detection for students with high-functioning autism in adaptive e-learning.” Soft Computing: ,2017.

[7] “End-to-end multi-modal Expressions recousing neural networks.” IEEE Journal of Topics in Signal Processing 11, no. 8: 13011309, 2017

[8] Ravichandra ginne, krupa Jariwala. FACIAL EXPRESSION RECOGNITION USING CNN. International Journal of Advances in Electronics and Computer Science, 2018

[9] Hai-Duong Nguyen, Soojan Yeom, Kyoung-Min Kim, Facial Expression Recognition Using Multi-Level Convolutional Neural Network, International Journal of Pattern Recognition and Artificial Intelligence, 2018

[10] Guan Wang, Jun Gong. Facial Expression Recognition Based on Improved LeNet-5 CNN. IEEE, 2019

[11] Joshua. G. Okemwa, Victor Mageto, Facial expression Recognition Using CNN and HOG classifier, IJRASET, 2019.

[12] S. Marry Hima Preethi, P. Sobha, p. Rajalakshmi, k. Gowri Raghavendra Narayan, Facial Expression Recognition Using CNN, IJSRCSEIT 2020.

[13] J. Bodapati, U. Srilakshmi, N. Veer Anjaneyulu, FERNet: A Deep CNN Architecture for Facial Expression Recognition, published in Journal of the Institution of Engineers in 2021

[14] Akash Kumar, Athira B. Nair, S. Jena, Debaraj Rana, Subrat.K. Pradhan, Facial Expression Recognition using python using CNN model, Journal of Applied Science, and Technology, 2021.

[15] Raheena Bagwan1, Sakshi Chintawar1, Komal Dhapudkar1, Alisha Balamwar1, Mr. Sandeep Gore2. FACIAL EMOTION RECOGNITION USING CONVOLUTION NEURAL NETWORK. IJCRT, 2021

[16] https://www.kaggle.com/datasets/deadskull7/fer2013