International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 173-177
Published Online March 2021 in IJEAST (http://www.ijeast.com)

A REVIEW ON FACIAL EMOTION RECOGNITION THAT USES MACHINE LEARNING ALGORITHMS

Maulin Patel
Department of Applied Instrumentation
L. D. College of Engineering, Ahm., Guj., India

Manisha Patel
Department of Instrumentation and control
L. D. College of Engineering, Ahm., Guj., India

Abstract— **For a computer, identification of human emotion from a still image of the human face is a complex, challenging, and heavily calculative task. Classification of human emotion is done by using a different combination of convolutional neural networks (CNN) that task is known as Facial Emotion Recognition (FER). CNN model is achieved by training and testing on lots of same categorical images from the dataset using different hyperparameter tuning. The main contribution of this work is to look for various CNN architectures, hyperparameter tuning and compare the performance of those CNN models based on accuracy and loss while training and testing on Facial Emotion Recognition. This study shall help to provide a guide for the selection of an appropriate CNN model and tuning parameter according to the needs of the applicant.**

Keywords— **Application of CNN, Convolutional Neural Network, Deep-learning, Facial Emotion Recognition, FER2013**

I. INTRODUCTION

As a human we communicate not only through verbal, for conveying our emotions and feeling, we also communicate by gestures and posture of our different body movements. Facial expression is just one of them. As technology progresses, we start to communicate through machines and nowadays we communicate with a machine that is also part of our daily routine.

Now a day we have devices that respond to our vocal commends, there will be a day where we can control our surroundings by just expressing our emotion through facial expression. Facial expression recognition can play a bigger role in that.

For machines, recognition of a facial expression is a well-known task in the branch that studies computer vision. There are two kinds of approaches for facial expression recognition (FER). The first is convention algorithms and the second is a deep learning-based approach. Here we are going to discuss about the latest method that is deep learning-based approach.

The researcher took inspiration from our brain cells to build an intelligent machine by mimicking the architecture of our brain's neurons we created artificial neural networks (ANNs). After that using knowledge of human visual patterns recognition system based on that researchers developed different mathematical models. That models developed over time with the contribution of lots of research and presented today as we know as Convolution neural networks (CNNs) [1].

We give one still image as input to trained CNN and it classifies the input image as one of the given categories it belongs to most.

In 2013 Kaggle introduce “Challenges in Representation Learning: Facial Expression Recognition Challenge”.in these challenges all the top three performed teams used CNN, from there this deep learning-based approach start gaining popularity [2].

In section 2 we are going to see basic blocks of CNN architecture and some terminologies that we used in deep learning. In section 3 we are going to see what different architecture was used in previous research with hyper tuning parameters. In section 4 we concluded this review.

II. TYPICAL CNN’S ARCHITECTURE

CNN, as we knew it today, was first proposed by LeCun et al. [3] that introduce LeNet-5 architecture (Fig.1). In that network, they introduce new building blocks Kwon as convolutional Layers and pooling Layers. also, retain some old concepts of fully connected layers and activation functions.

A. Convolutional Layer

It basically means that the relationship between two layers can be defined by convolutional operation. that have learnable neuron weights.

Neurons in the first convolutional layer are only connected to pixels in their receptive fields of the input image.

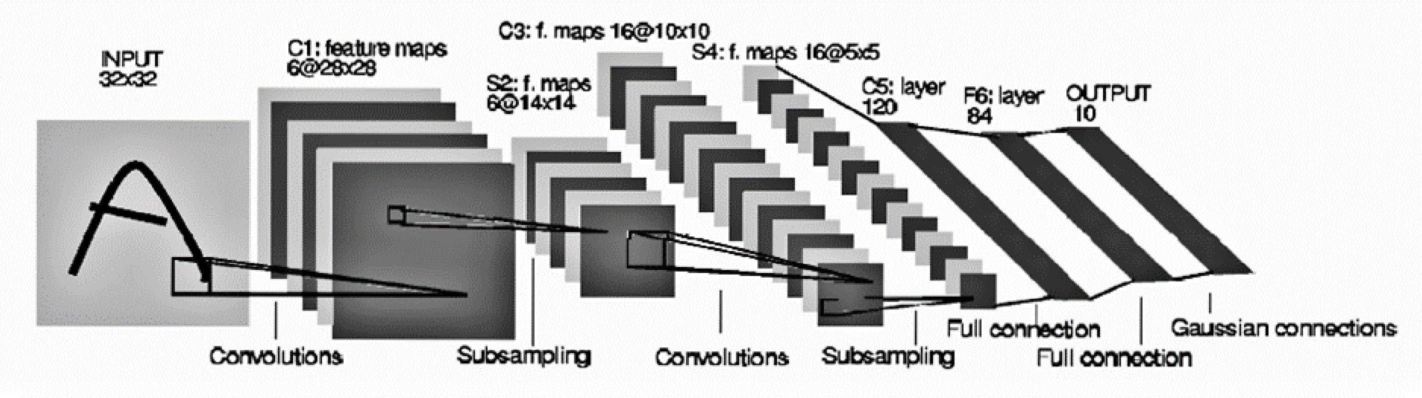

Fig. 1. Typical CNN(LeNet) [3]

We slide a receptive field over the first layer that has learnable weights neurons to get the second layers' pixel values. these windows have different sizes like 1x1, 3x3,5x5 These are rectangular windows but there are other kinds of vector windows too like 1x5,7x1. Where, first value represents numbers of raw in the receptive field and the second value represents numbers of the column in the receptive field, and also you can do depth-wise convolution to generate the next layer.

B. Strides

While sliding an operational window over the image, how much gap of distance to live between two windows known as a stride (how much gap required between two windows is known as stride). It has numbers of row pixels to jump and numbers of column pixels to jump. For example, stride size can be 1x1, 2x2, 2x1, etc.

C. Padding

In order for a layer to have the same height and width as the previous layer, it is common to add zeros around the inputs. This is called zero padding.
There are two kinds of padding that is mostly use:
(i) valid padding in which input image doesn't get any padding before operation and
(ii) same padding where the image gets padded to get the same dimension of the output image as an input image.

D. Pooling Layer

For subsampling input image pooling layers are used and they reduce the number of parameters. There are many types of pooling layers like max-pooling, min pooling, average pooling, etc.
Each neuron in a pooling layer is connected to the outputs of a limited number of neurons in the previous layer, located within a small rectangular receptive field. This receptive field has the size, stride, and padding type.

E. Activation Function

These are mathematical functions that determine whether a particular neuron should generate output or not from given values of the input, weight, and bias. There are two basic categories of activation functions first one is linear activation function and non-linear activation function. Rectified Linear Unit Functions (ReLu) is a wildly used non-linear activation function.

F. Fully Connected Layers

Fully connected layers are commonly referred to as a neural network. Where each and every neuron of the presentation layer is connected to each and every neuron of the next layer.

G. Data Set

For training and testing of deep neural networks we need lots of data. That data has to have an input image and correspondingly have a category it belongs to that has to be mention in given data.
Creating our own data set is a very time-consuming and expensive task so generally, we try to get open-source data set that we can use to training and validate our network for a particular application. FER2013 is one of the data sets provided by Kaggle for the facial emotion recognition challenge in 2013 [2].

III. DIFFERENT CNN ARCHITECTURE

In, this section we have discussed some of the important convolutional neural network's architecture that have been used for research purpose and discussed their training parameter's.

A. Shallow CNN vs Deep CNN [4]:

Shallow CNN has only one convolution layer with 5x5x64(height, width, channel depth) filters, astride of size 1, along with max-pooling of size 5x5 and strides of 2x2. After that 3 FC layers flatten 25600 neurons followed by 1024 and 7 neurons. The last 7 neurons have softmax as the loss function along with dropout. This network has around 26M+ trainable parameters. In all the layers they used ReLu as an activation function. This model gave 48% accuracy on test data.

Deep CNN consists of 8 convolutional layers,4 max-pooling layers, and 3FC layers. Filter window size is 3x3 for all convolutional layers and the numbers of channels are 32 for the first two layers, 64 for the third and fourth layer, 96 for the fifth and sixth layer, and 128 for a seventh and eighth layer. This network has around 470K+ trainable parameters. They used the swish activation function for FC layers. Summary of Deep CNN as follows:

INPUT→ [(CONV → ReLu) * 2 → Max Pool] * 4 → [FC → SWISH] * 2 → FC

Train, validation, and testing performed on a FER2013 Data set.

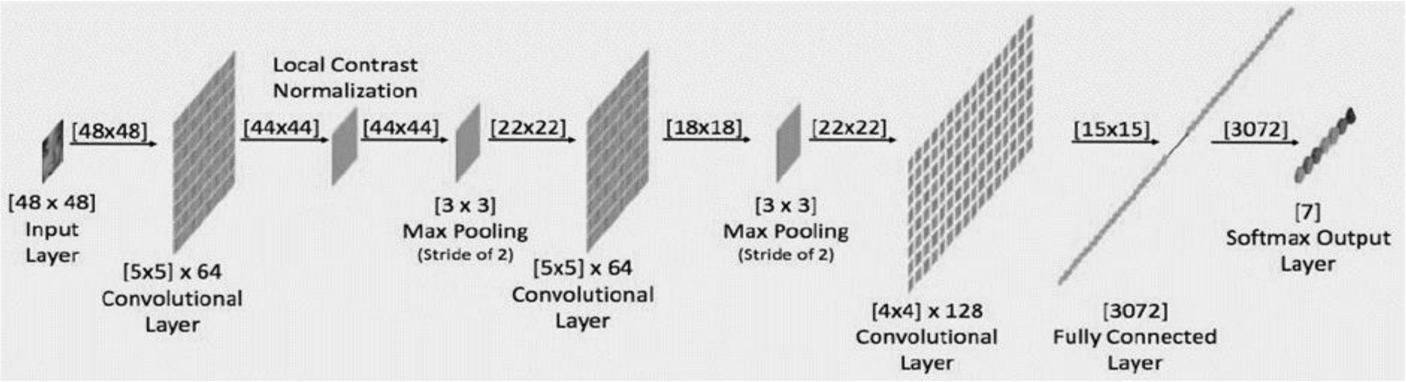
Local Contrast Normalization
[48x48]
[44x44]
[44x44]
[22x22]
[18x18]
[22x22]
[15x15]
[3072]
[48 x 48] Input Layer
[5x5] x 64 Convolutional Layer
[3 x 3] Max Pooling (Stride of 2)
[5x5] x 64 Convolutional Layer
[3 x 3] Max Pooling (Stride of 2)
[4x4] x 128 Convolutional Layer
[3072] Fully Connected Layer
[7] Softmax Output Layer

Fig. 2. Proposed Systems Network Model [10]

B. MTCNN and miniShuffleNet V2 [5]:

CNN is used for two problems first is that too find a human face from a still image and second for recognizing expressions of that human face. For the first problem, they use MTCNN Multi-Task Convolutional Neural Network (MTCNN) and for the second problem, they use miniShuffleNet V2. By Combing these two CNNs they achieved a test accuracy of 71.19%.

They used the FER2013 Data set. they achieved remark-able accuracy by normalizing the data set by augmentation of data.

ACNN for Occlusion-aware FER [6]:

They used two different versions of ACNN: patch-based ACNN (pACNN) and global–local-based ACNN (gACNN). Further, pACNN and gACNN divided into two schemes

Two schemes of pCNN has

i. Region decomposition in which 24 facial landmark points selected.
ii. Occlusion perception in which Patch-Gated Unit (PG-Unit) used to learn, weigh the patch's local representation by its un-obstructedness.

Two schemes of gACNN are

i. integration with full face region and
ii. global-gated Unit (GG-Unit).

Input image decomposes into feature to obtain local patches mappedmaps by ACNN. Then feature maps are sent to gg-unit for identification of occlusion. After that pg-unit and gg-unit are concatenated and softmax loss is used for output.

Indian Spontaneous Expression Database (ISED) is used. Accuracy is only high for happy emotion other than that accuracy rate is under 40%.

Li et al. [7] show that gACNN gives better accuracy than pACNN, pCNN, and gCNN. Because gACNN combines local representations at patch-level with the global representation of an image.

Uddin et al. [8] proposed volume symmetric local graph structure (VSLGS) for spatiotemporal future from video data and CNN(VGG-11) is used for deep spatial information. These both features were fused and given to the Spark MLlib Multilayer Perceptron (MLP) classifier. They achieved 98.2% accuracy with CK+ dataset and 80.35% accuracy with Oulu-CASIA dataset.

Cornejo et al. [9] show that rather than giving the original image as input to CNN(VGG) Census-Transformed (CT) image as input to CNN gives a more accurate result in the task of emotion recognition.

Jadhav et al. [10] proposed a network as in Fig.2 They achieve a testing accuracy of 63%.

Mollahosseini et al. [11] use the Inception module [12] in between convolution layers and fully connected layers. The network architecture of the proposed model is shown in Fig.3. They achieve testing accuracy on FER2013 is 66.4%. during training, they use a Polynomial learning rate because that test loss converges fast.

Singh et al. [13] use fixed-size 3x3x128 convolution in all 6 layers of convolution layers and achieved 61.7% accuracy with 1.2M trainable parameters.

Yu et al. [14] show that the use of Log-Likelihood Loss and Testing accuracy from 71.08% to 72%. Also, they used Stochastic Pooling instead of popular max-pooling because of that they reported a loss in accuracy.

Raghuvanshi et al. [15] noted that Fractional max-pooling does not outperform standard max pooling and also increases training time. though a higher dropout rate reduces overfitting it will increase training time to achieve the same performance as the model without dropout. An increase in the numbers of first layer filters costs you more taring time and computational cost but does not guarantee an increase in accuracy.

IV. CONCLUSION

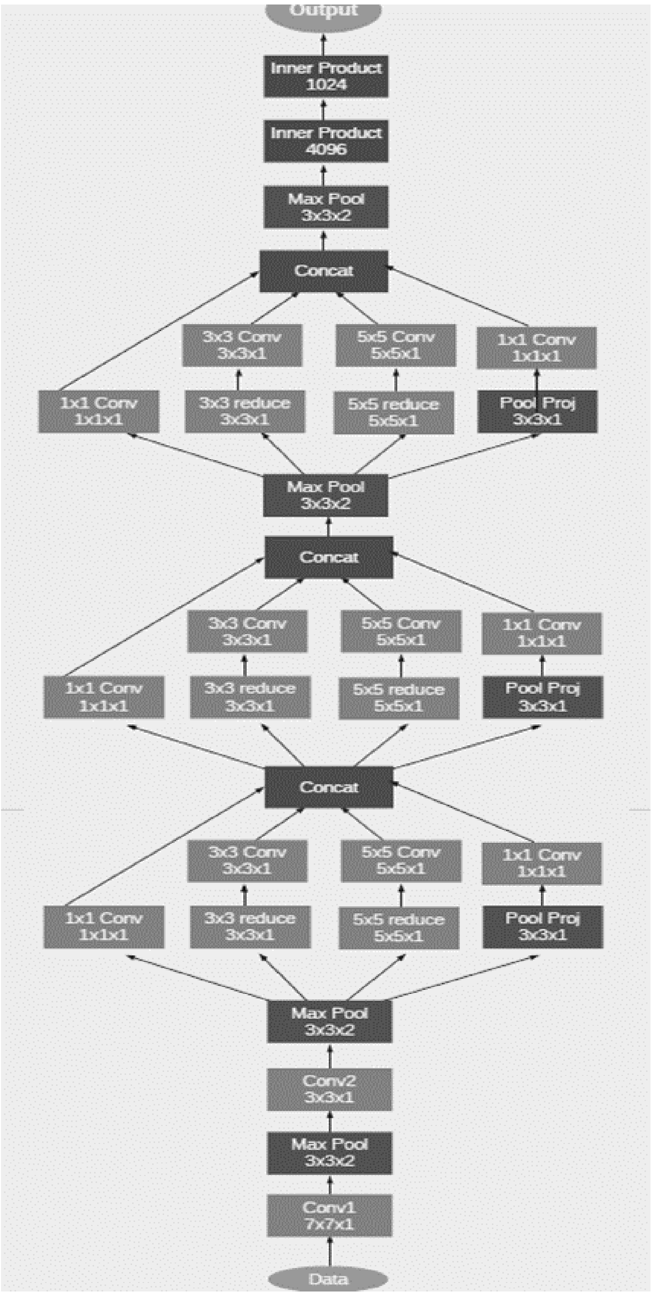

Fig. 3 Network Architecture [11]

This paper review facial emotion recognition (FER) by using deep learning-based approaches. The deep learning-based approach provides much higher accuracy than the algorithmic-based approach in the FER application [16]. The deep learning-based approach involves CNN architectures, the dataset for training and testing purposes, and training parameters. In this review, we have discussed some conventional hyperparameters and CNN architecture. By normalizing the FER2013 dataset we can achieve higher average accuracy [5]. Pre-processing input image with some transformed gives little more accurate results [9]. The architecture of CNN has no versatility for FER applications mostly they are derived from VGG. We can use the state-of-art CNN architecture of the ImageNet challenge for our application. We have to start training from scratch, fine-tuning those networks do not give more accurate performance than the newly trained model [14, 15]. There is no particular way in deep learning, we can use different architectures, different tuning parameters that give more accurate classification and meet our needs in limited computation power.

V. REFERENCE

[1] Géron A. . (September 2019) . Deep Computer Vision Using Convolutional Neural Networks. In: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* s.l.:O'Reilly, (pp. 445-495).

[2] Goodfellow I.J., Erhan D., Carrier P.L., Courville A., Mirza M., Hamner B., Cukierski W., Tang Y., Thaler D., Lee D.H. and Zhou Y. . (November 2013) . Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117-124). Springer, Berlin, Heidelberg.

[3] LeCun Y., Bottou L., Bengio Y. and Haffner P. . (1998) . Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), pp.2278-2324.

[4] Pathar R., Adivarekar A., Mishra A. and Deshmukh A. . (April 2019) . Human Emotion Recognition using Convolutional Neural Network in Real Time. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (pp. 1-7). IEEE.

[5] Ghofrani A., Toroghi R.M. and Ghanbari S. . (2019) . Realtime face-detection and emotion recognition using mtcnn and minishufflenet v2. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)* (pp. 817-821). IEEE.

[6] Engoor S., SendhilKumar S., Sharon C.H. and Mahalakshmi G.S. . (March 2020) . Occlusion-aware Dynamic Human Emotion Recognition Using Landmark Detection. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 795-799). IEEE.

[7] Li Y., Zeng J., Shan S. and Chen X. . (2018) . Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, *28*(5), pp.2439-2450.

[8] Uddin M.A. and Lee Y.K. . (February 2020) . Dynamic Facial Emotion Recognition Using Deep Spatial Feature and Handcrafted Spatiotemporal Feature on Spark. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)* (pp. 21-27). IEEE.

[9] Cornejo J.Y.R. and Pedrini H. . (October 2019) . Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (pp. 3396-3402). IEEE.

[10] Jadhav R.S. and Ghadekar P. . (December 2018) . Content based facial emotion recognition model using machine learning algorithm. In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)* (pp. 1-5). IEEE.

[11] Mollahosseini A., Chan D. and Mahoor M.H. . (March 2016) . Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

[12] Li H., Su P., Chi Z. and Wang J. . (August 2016) . Image retrieval and classification on deep convolutional SparkNet. In *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (pp. 1-6). IEEE.

[13] Singh S. and Nasoz F. . (January 2020) . Facial expression recognition with convolutional neural networks. In *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0324-0328). IEEE.

[14] Yu Z. and Zhang C. (November 2015) . Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 435-442).

[15] Raghuvanshi A. and Choksi V. . (2016) . Facial expression recognition with convolutional neural networks. *CS231n Course Projects*, *362*.

[16] Ko B.C. . (2018) . A brief review of facial emotion recognition based on visual information. *sensors*, *18*(2), p.401.

[17] Alizadeh S. and Fazel A. . (2017) . Convolutional neural networks for facial expression recognition. *arXiv preprint arXiv:1704.06756*.

[18] Taha B. and Hatzinakos D. . (May 2019) . Emotion recognition from 2d facial expressions. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)* (pp. 1-4). IEEE.