Amrita School of Computing, Chennai

Department of Computer Science and Engineering (Artificial Intelligence)

*21AIE205 – Python for Machine Learning*

**Capstone Project Title: Facial Expression Recognition using CNN**

## ABSTRACT

Emotion recognition based on facial expressions is an interesting research area that has been presented and applied in several fields such as security, health, and human-machine interfaces. Researchers in this field are interested in developing techniques to interpret facial expressions, to encode them, and to extract these features in order to get a better prediction from the computer. With the remarkable success of deep learning, different types of architectures of this technique are exploited to achieve better performance. It may also be used in behavioural science and in clinical practice. An automatic Facial Expression Recognition system needs to perform detection and location of faces in a cluttered scene, facial feature extraction, and facial expression classification.

Our facial expression recognition system is implemented using Convolution Neural Network (CNN). CNN model of the project consists of 7 layers. Facial expression dataset with seven facial expression labels as happy, sad, surprise, fear, anger, disgust, and neutral are used in this project. The purpose of this project is to provide an insight on how automated facial emotion recognition FER works through deep learning.

***Keywords: Convolutional Neural Network, Emotion.***

## INTRODUCTION

One of the more crucial components of human communication is facial expressions. The face is in charge of conveying emotions as well as thoughts and ideas. The communication of emotions is fascinating because it seems as though some of these emotional expressions such as contempt, disgust, fear, happiness, sad, surprise, and to a lesser extent, disgust, interest, pain, and shame may be biologically hardwired and are expressed in the same manner by all peoples of all cultures[1]. This stands in contrast to other viewpoints that claim that social learning and culture are the primary causes of all facial expressions.

Human smiles and facial analysis are nearly inextricably linked since every expression, especially for computer algorithms, transforms our faces beyond recognition. Two front-facing facial portraits can be compared by software to see if they show the same person[3]. These solutions evaluate so-called nodal points on human features in a manner similar to portrait artists. These points are utilized to identify our unique faces; various techniques locate

somewhere between 80 and 150 nodal points on a single face.

Facial expression detection uses biometric markers to detect emotions in human faces. More precisely, this technology is a sentiment analysis tool and can automatically detect the six basic or universal expressions: happiness, sadness, anger, surprise, fear, neutral, and disgust[2].

First of all, emotion detection is a very important task for many companies to understand how are their consumers reacting to the products launched by them. Also, it can be used to know whether their employees are satisfied with the facilities given to them. Also, it has many other use cases like checking a person's mood without getting near to him as we are using the camera to detect. Also, the same algorithm just needs a little modification and can be used in other fields like face detection, attendance system, mask detection, and many more. To detect facial expressions using Machine learning algorithms such as CNN

## OBJECTIVE:

The Problem statement is to implement a code and train it in order to recognize facial expressions. To establish this, we'll make an avatar that can watch videos0 and recognize people's facial expressions using webcams. The facial expression recognition project will use convolutional neural networks, a deep learning approach. While working with video, the convolution technique enables us to reduce computation without sacrificing the system's accuracy. We will construct each layer of our convolutional neural network model from scratch using the TensorFlow framework rather than pre-existing models like VGG-16, Reset, ImageNet, etc

## CNN ARCHITECTURE:

Machine learning includes convolutional neural networks, also known as convnets or CNNs. It is a subset of the several artificial neural network models that are employed for diverse purposes and data sets. A CNN is a particular type of network design for deep learning algorithms that are utilized for tasks like image recognition and pixel data processing[4].

Although there are different kinds of neural networks in deep learning, CNNs are the preferred network architecture for identifying and recognizing objects. They are therefore ideally suited for computer vision (CV) activities and for applications where accurate object recognition is crucial, such as facial and self-driving automobile systems.

CNN is another type of neural network that can uncover key information in both time series and image data [5]. This makes it very beneficial for applications involving images, such as pattern recognition, object classification, and picture identification. A CNN makes use of linear algebraic concepts, including matrix multiplication, to find patterns in an image. CNN may categorize audio and signal data as well.

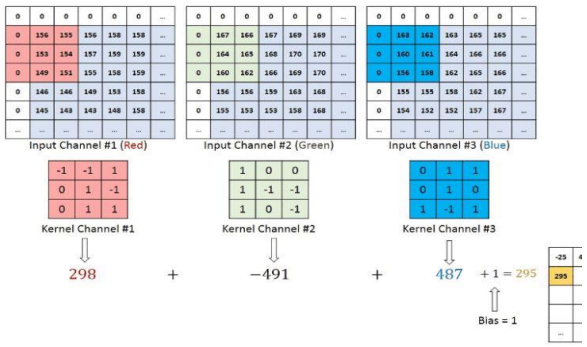
### CNN layers

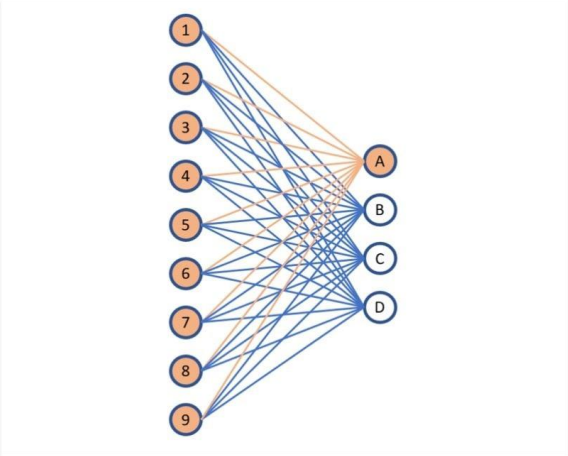
A deep learning CNN consists of three layers: a convolutional layer, a pooling layer and a fully connected (FC) layer.

*Convolutional Layer:*

A convolutional layer is the main building block of a CNN. It contains a set of filters (or kernels), parameters must be learned over the course of training [6]. Typically, the filters' size is smaller than the original

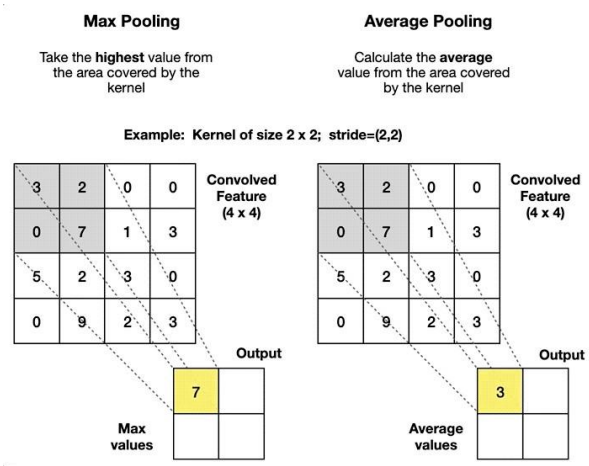
image. Each filter produces an activation map after it convolves with the image.




*Pooling Layer:*

Pooling layers are used to reduce the dimensions of the feature maps. As a result, it minimizes the quantity of network computation and the number of parameters that must be trained. The feature map created by a convolution layer's feature pooling layer summarises the features that are present in a certain area.




*Fully connected Layer:*

A neural network with fully connected layers is one in which each neuron uses a weights matrix to apply a linear transformation to the input vector. As a result, every input of the input vector influences every output of the output vector, signifying that all layer-to-layer linkages are present.

**TYPES OF CNN ARCHITECTURE:**

**LeNet**: LeNet is the first CNN architecture. The LeNet architecture consists of multiple convolutional and pooling layers, followed by a fully-connected layer.

**AlexNet**: AlexNet is the deep learning architecture that popularized CNN. The AlexNet architecture was designed to be used with large-scale image datasets and it achieved state-of-the-art results at the time of its publication. AlexNet is composed of 5 convolutional layers with a combination of max-pooling layers, 3 fully connected layers, and 2 dropout layers [7].

**ZF Net**: ZFnet is the CNN architecture that uses a combination of fully-connected layers and CNNs. ZF Net CNN architecture consists of a total of seven layers: Convolutional layer, max-pooling layer (downscaling), concatenation layer, convolutional layer with linear activation function, and stride one, dropout for regularization purposes applied before the fully connected output.

**GoogLeNet**: GoogLeNet is the CNN architecture used by Google to win ILSVRC 2014 classification task. GoogleNet CNN architecture is computationally expensive. To reduce the parameters that must be learned, it uses

heavy unpooling layers on top of CNNs to remove spatial redundancy during training and also features shortcut connections between the first two convolutional layers before adding new filters in later CNN layers.

**VGGNet:** The VGG CNN model is computationally efficient and serves as a strong baseline for many applications in computer vision due to its applicability for numerous tasks including object detection.

**MobileNets**: MobileNets are CNNs that can be fit on a mobile device to classify images or detect objects with low latency. The architecture is also flexible so it has been tested on CNNs with 100-300 layers and it still works better than other architectures like VGGNet.

**GoogLeNet_DeepDream**: The architecture is often used with the ImageNet dataset to generate psychedelic images or create abstract artworks using human imagination

## WORKING PRINCIPLE OF CNN ALGORITHM

Multiple layers of a CNN are possible, and each layer trains the CNN to recognize the many aspects of an input image. Each image is given a filter or kernel to create an output that gets better and more detailed with each layer. The filters may begin as basic characteristics in the lower layers.

At each successive layer, the filters increase in complexity to check and identify features that uniquely represent the input object. As a result, the partially recognized image from each layer's output, or convolved image, serves as the input for the subsequent layer [8]. The CNN recognizes the image or objects it represents in the final layer, which is an FC layer.

The input image is processed through a number of different filters during convolution. Each filter performs its function by turning on specific aspects of the image, after which it sends its output to the filter in the subsequent layer. The operations are repeated for dozens, hundreds, or even thousands of layers as each layer learns to recognize various features. Finally, the CNN is able to recognize the full object after processing all the picture data through its many layers [9].

## DATASET USED:

The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589.

## IMPLEMENTATION

We are using TensorFlow for data pre-processing

*Tensorflow:*

TensorFlow, a rival to PyTorch and Apache MXNet frameworks, can train and run deep neural networks for image recognition, word embeddings, recurrent neural networks, sequence-to-sequence models for machine translation, natural language processing, and PDE (partial differential equation)-based simulations. The best part is that TensorFlow uses the same models that were used for training to provide production prediction at scale [13].

*Working Principal of TensorFlow:*

TensorFlow allows developers to create dataflow *graphs*—structures that describe how data moves through a graph, or a series of processing nodes. Each node in the graph represents a mathematical operation, and each connection or edge between nodes is a multidimensional data array or *tensor*.

TensorFlow applications can be run on almost any target that's convenient: a local machine, a cluster in the cloud, iOS and Android devices, CPUs, or GPUs. If you use Google's own cloud, you can run TensorFlow on Google's custom TensorFlow Processing Unit (TPU) silicon for further acceleration. The resulting models created by TensorFlow, though, can be deployed on almost any device where they will be used to serve predictions [14].

Keras is used for image data generation and recognition

Seven different layers are used in the CNN model

- Input Layer
- Convo Layer
- Maxpooling Layer
- Dropout Layer
- Dense Layer
- Flatten Layer
- Output Layer

Training the model in all these later, it will be tested and find the accuracy of the model.

## RELATED WORKS

Yu and Zhang used a five layer ensemble CNN to achieve a 0.612 accuracy. They pre-trained their models on the FER-2013 dataset and then finetuned the model on the Static Facial Expressions in the Wild 2.0 (SFEW) dataset. They used an ensemble of three face detectors to detect and extract faces from the labelled movie frames of SFEW[11]. They then proposed a data perturbation and voting method to increase the recognition performance of the CNN. They also chose to use stochastic pooling layers over max pooling layers citing its better performance on their limited data.

Kahou et al. used a CNN-RNN architecture to train a model on individual frames of videos as well as static images. They made use of the Acted Facial Expressions in the Wild (AFEW) 5.0 dataset for the video clips and a combination of the FER-2013 and Toronto Face Database for the images. Instead of using long short-term memory (LSTM) units, they used IRNNs which are composed of rectified linear units (ReLUs). These IRNNs provided a simple mechanism for dealing with the vanishing and exploding gradient problem. They achieved an overall accuracy of 0.528 [15].

Mollahosseini et al. proposed a network consisting of two convolutional layers each followed by max pooling and then four Inception layers. They used this network on seven different datasets including the FER-2013 dataset. They also compared the accuracies of their proposed network with an AlexNet [12] network trained on the same datasets. They found that their architecture had better performance on the MMI and FER-2013 datasets with comparable performances on the remaining five datasets. The FER-2013 dataset in particular managed to reach an accuracy of 0.664. Ming Li et al. [10] propose a neural network

## CONCLUSION

In this project, the aim was to classify facial expressions into one of seven emotions by using various models on the FER dataset.

The ability of the model to make predictions in effectively real-time, indicates that real world uses of facial emotion recognition is barred only by the relative inaccuracies of the model itself. In the future, an in-depth analysis of the top- 2 predicted emotions may lead to a much more accurate and reliable system. Further training samples for the more difficult to predict emotion of disgust will definitely be required in order to perfect such a system. The real-time capacity of the model in addition to its quick training time and near-state-of-the-art accuracy allows the model to be adapted and used in nearly any use-case. This also implies that with some work, the model could very well be deployed into real-life applications for effective utilization in domains such as in healthcare, marketing and the video game industry.

## REFERENCE


**[1]** Meng, Z., Liu, P., Cai, J., Han, S. and Tong, Y., 2017, May. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 558-565). IEEE.

**[2]** Li, S. and Deng, W., 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*.

**[3]** Meng, Z., Liu, P., Cai, J., Han, S. and Tong, Y., 2017, May. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 558-565). IEEE.

**[4]** Dobs, K., Schultz, J., Bülthoff, I. and Gardner, J.L., 2018. Task-dependent enhancement of facial expression and identity representations in human cortex. *NeuroImage*, *172*, pp.689-702.

**[5]** Cheng, S., Kotsia, I., Pantic, M. and Zafeiriou, S., 2018. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5117-5126)

**[6]** Zhang, Z., Luo, P., Loy, C.C. and Tang, X., 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, *126*(5), pp.550-569

**[7]** H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by deexpression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2168–2177.

**[8]** K. Zhang, Y. Huang, Y. Du, and L. Wang, ``Facial expression recognitionbased on deep evolutional spatial-temporal networks," *IEEE Trans. ImageProcess.*, vol. 26, no. 9, pp. 4193_4203, Sep. 2017.

**[9]** Li, M., Xu, H., Huang, X., Song, Z., Liu, X. and Li, X., 2018. Facial expression recognition with identity and emotion joint learning. *IEEE Transactions on affective computing*, *12*(2), pp.544-550.

**[10]**

**[11]** Huang, W., Zhang, S., Zhang, P., Zha, Y., Fang, Y. and Zhang, Y., 2021. Identity-aware facial expression recognition via deep metric learning based on synthesized images. *IEEE Transactions on Multimedia*.

**[12]**

**[13]** Zhang, S., Pan, X., Cui, Y., Zhao, X. and Liu, L., 2019. Learning affective video features for facial expression recognition

via hybrid deep learning. *IEEE Access*, *7*, pp.32297-32304.

**[14]** Li, Y., Wang, S., Zhao, Y. and Ji, Q., 2013. Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on image processing*, *22*(7), pp.2559-2573.

**[15]**