**Live Class Monitoring System**

**(Face Emotion Recognition)**

**Akash Salmuthe**

**Data Science Enthusiast**

**AlmaBetter, Bangalore**

***Abstract*— the rapid growth of artificial intelligence has contributed a lot to the technology world. As the traditional algorithms failed to meet the human needs in real time, Machine learning and deep learning algorithms have gained great success in different applications such as classification systems, recommendation systems, pattern recognition etc. Emotion plays a vital role in determining the thoughts, behavior and feeling of a human. An emotion recognition system can be built by utilizing the benefits of deep learning and different applications such as feedback analysis, face unlocking etc. can be implemented with good accuracy. The main focus of this work is to create a Deep**

**Convolutional Neural Network (DCNN) model that classifies 7 different human facial emotions. The model is trained, tested and validated using the manually collected image dataset.**

***Keywords— Facial Emotion Recognition, Deep Convolutional Neural Network, Classification, Adam.***

I. INTRODUCTION

The face is the most expressive and communicative part of a human being [1]. It’s able to transmit many emotions without saying a word. Facial expression recognition identifies emotion from face image, it is a manifestation of the activity and personality of a human. In the 20th century, the American psychologists Ekman and Friesen [2] defined six basics’ emotions (anger, fear, disgust, sadness, surprise and happiness), which are the same across cultures.

Facial expression recognition has brought much attention in the past years due to its impact in clinical practice, sociable robotics and education. According to diverse research, emotion plays an important role in education. Currently, a teacher use exams, questionnaires and observations as sources of feedback but these classical methods often come with low efficiency. Using facial expression of students the teacher can adjust their strategy and their instructional materials to help foster learning of students.

The purpose of this article is to implement emotion recognition in education by realizing an automatic system that analyze students’ facial expressions based on Convolutional Neural Network (CNN), which is a deep learning algorithm that are widely used in images classification. It consist of a multistage image processing to extract feature representations. Our system includes three phases: face detection, normalization and emotion recognition that should be one of these seven emotions: neutral, anger, fear, sadness, happiness, surprise and disgust.

II. PROBLEM STATEMENT

The Indian education landscape has been undergoing rapid changes for the past 10 years owing to the advancement of web-based learning services, specifically, eLearning platforms. Global E-learning is estimated to witness an 8X over the next 5 years to reach USD 2B in 2021. India is expected to grow with a CAGR of 44% crossing the 10M users mark in 2021. Although the market is growing on a rapid scale, there are major challenges associated with digital learning when compared with brick and mortar classrooms. One of many challenges is how to ensure quality learning for students. Digital platforms might overpower physical classrooms in terms of content quality but when it comes to understanding whether students are able to grasp the content in a live class scenario is yet an open-end challenge. In a physical classroom during a lecturing teacher can see the faces and assess the emotion of the class and tune their lecture accordingly, whether he is going fast or slow. He can identify students who need special attention. Digital classrooms are conducted via video telephony software program (ex: Zoom) where it’s not possible for medium scale class (25-50) to see all students and access the mood. Because of this drawback, students are not focusing on content due to lack of surveillance. While digital platforms have limitations in terms of physical surveillance but it comes with the power of data and machines which can work for you. It provides data in the form of video, audio, and texts which can be analyzed using deep learning algorithms. Deep learning backed system not only solves the surveillance issue, but it also removes the human bias from the system, and all information is no longer in the teacher’s brain rather translated in numbers that can be analyzed and tracked.

We will solve the above-mentioned challenge by applying deep learning algorithms to live video data. The solution to this problem is by recognizing facial emotions.

**Face Emotion Recognition**: This is a few shot learning live face emotion detection system. The model should be able to real-time identify the emotions of students in a live class.

III. DATASET INFORMATION

The data comes from the past Kaggle competition “Challenges in Representation Learning: Facial Expression Recognition Challenge [3]”: defined the image size to 48 so each image will be reduced to a size of 48x48. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image.

The dataset contains approximately 36K images. Each image corresponds to a facial expression in one of seven categories (0=Angry,

1=Disgust,

2=Fear,

3=Happy,

4=Sad,

5=Surprise,

6=Neutral)

IV. APPROACH

In this section, we describe our proposed system to analyze students’ facial expressions using a Convolutional Neural Network (CNN) architecture. First, the system detects the face from input image and these detected faces are cropped and normalized to a size of 48×48. Then, these face images are used as input to CNN. Finally, the output is the facial expression recognition results (anger, happiness, sadness, disgust, surprise or neutral). Figure1 presents the structure of our proposed approach.

A Convolutional Neural Network (CNN) is a deep artificial neural networks that can identify visual patterns from input image with minimal pre-processing compared to other image classification algorithms.



Fig 1: The Structure of facial emotion recognition system

This means that the network learns the filters that in traditional algorithms were hand-engineered. The important unit inside a CNN layers is a neuron. They are connected together, in order that the output of neurons at a layer becomes the input of neurons at the next layer.

In order to compute the partial derivatives of the cost function the backpropagation algorithm is used. The term convolution refers to the use of a filter or kernel on the input image to produce a feature map. In fact, CNN model contains 3 types of layers as shown in Figure 2:

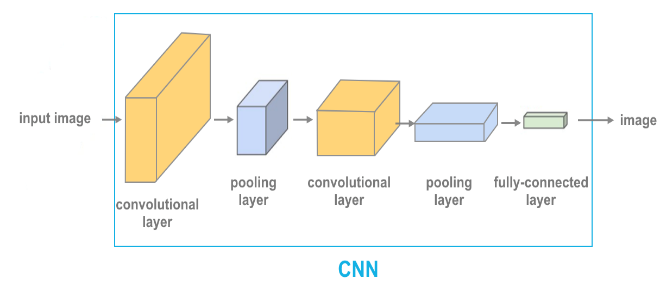


Fig 2: Generic CNN Architecture

**Convolution Layer:** is the first layer to extract features from an input image. The primary purpose of Convolution in case of a ConvNet is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data [21]. It performs a dot product between two matrices, where one is the image and the other is a kernal. The convolution formula is represented in Equation 1

C:\Users\AKASH\Desktop\Emotion Images\ex 1.png

Where net (*t, f*) is the output in the next layer, *x* is the input image, *w* is the filter matrix and \* is the convolution operation.

**Pooling Layer:** reduces the dimensionality of each feature map but retains the most important information. Pooling can be of different types: Max Pooling, Average pooling and Sum Pooling.



Fig 3: Pooling Layer

The function of Pooling is to progressively reduce the spatial size of the input representation and to make the network invariant to small transformations, distortions and translations in the input image. In our work, we took the maximum of the block as the single output to pooling layer as shown in Figure 3.

**Fully connected layer:** is a traditional Multilayer Perceptron that uses an activation function in the output layer. The term “Fully Connected” implies that every neuron in the previous layer is connected to every neuron on the next layer. The purpose of the Fully Connected layer is to use the output of the convolutional and pooling layers for classifying the input image into various classes based on the training dataset. So the Convolution and Pooling layers act as Feature Extractors from the input image while Fully Connected layer acts as a classifier.

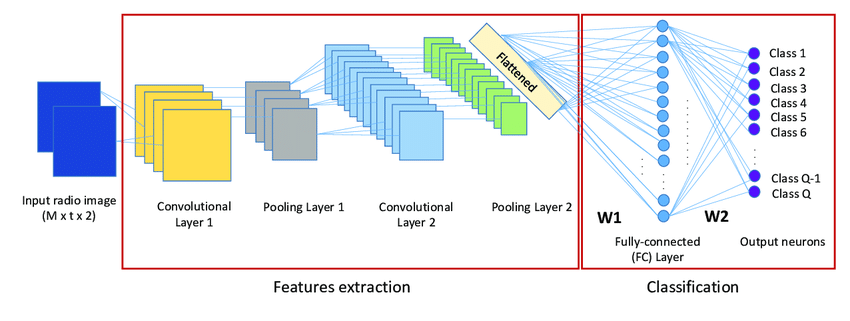
Fig 4: CNN Model

Figure 4 represents our CNN model. It contains 4 convolutional layers with 4 pooling layers to extract features, and 2 fully connected layers then the softmax layer with 7 emotion classes. Input image is grayscale face image with a size of 48×48. For each convolutional layer we used 3×3 filters with stride 2. For the pooling layers, we used max pooling layer and 2×2 kernels with stride 2. Thus, to introduce the non-linearity in our model we used the Rectified Linear Unit (ReLU), defined in Equation 2, which is the most used activation function recently.

R(*z*) = max(0,z)

As shown in Figure 5, R(*z*) is zero when *z* is less than zero and R(*z*) is equal to *z* when *z* is above or equal to zero. Table I presents the network configuration of our model.

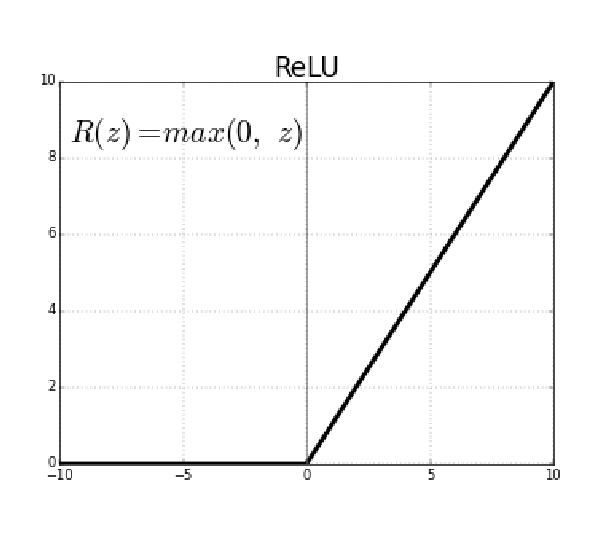


Fig 5: ReLu Function

V. IMPLEMENTATION

Here we are using mainly two approach, ResNet50 and Convolution Neural Network respectively

* **ResNet50:**

ResNet50 is a variant of the ResNet model which has 48 Convolution layers along with 1 MaxPool and 1 Average Pool layer. It has 3.8 x 10^9 Floating points operations.

The ResNets were initially applied to the image recognition task but as mentioned in the paper that the framework can also be used for non-computer vision tasks to achieve better accuracy. Started with ResNet50 and added 2 FC layers and trained the model freezing all Conv layers except the last 4 and on the second run, we fine tune the model by unfreezing all the layers. And got an accuracy of 40% and 50% respectively

* **CNN:**

1. 4 convolutional layers
2. 2 fully connected layers

Also, we use some common techniques for each layer Batch normalization: improves the performance and stability of NNs by providing inputs with zero mean and unit variance. Dropout: reduces overfitting by randomly not updating the weights of some nodes. This helps prevent the NN from relying on one node in the layer too much.

VI. RESULTS

Model can be evaluated by various metrics such as Accuracy,categorical cross-entropy (Loss) and Confusion matrix

* **Accuracy:**

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost the same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.66 which means our model is approx. 66% accurate.

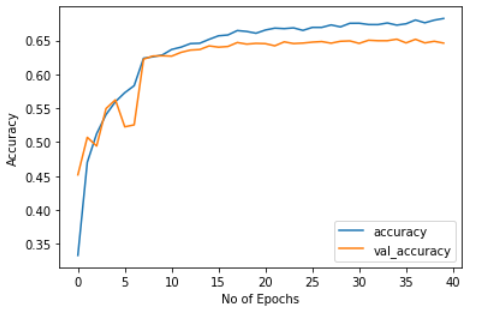
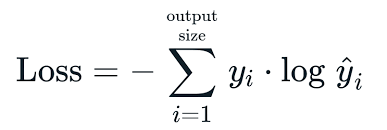


Fig 6: Accuracy Plot

* **Loss**

The categorical cross-entropy loss function calculates the loss of an example by computing the following sum:



where y^i is the i-th scalar value in the model output, yi is the corresponding target value, and the output size is the number of scalar values in the model output.

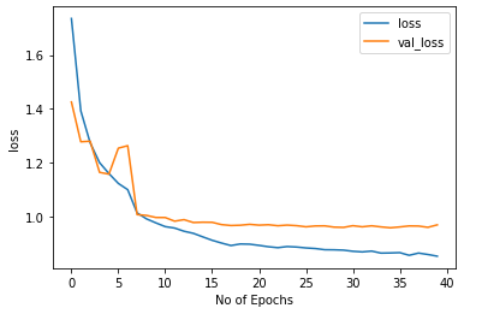


Fig 6: categorical cross-entropy

* **Confusion Matrix**

A confusion matrix, also known as an error matrix.

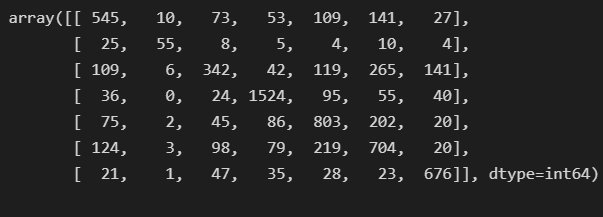
****

Fig 7: Confusion Matrix

Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances an actual class while each column represents the instances in a predicted class, or vice versa.

* **Live Face Emotion Recognition**

Here we are, some live webcam samples with single person and multiple faces

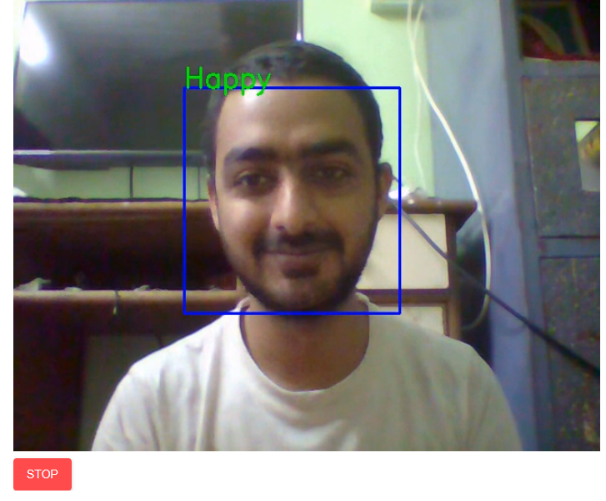


Fig 8: Happy face

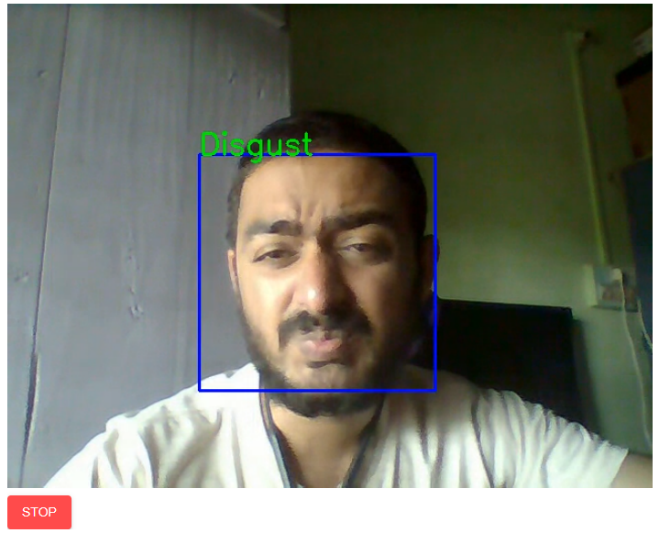


Fig 9: Disgust face

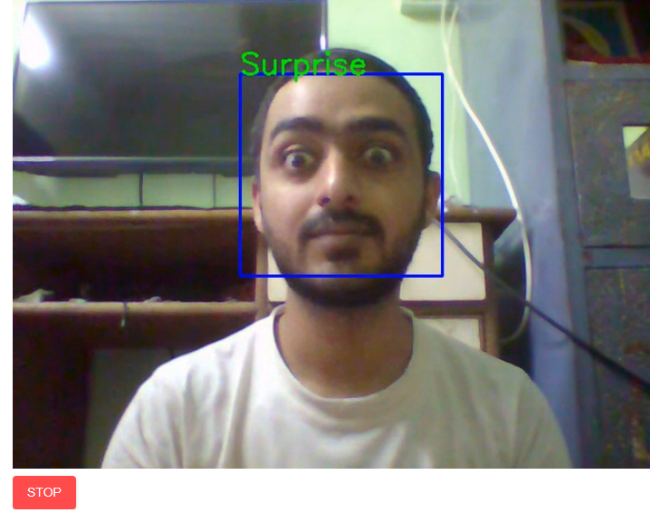


Fig 10: Surprise face

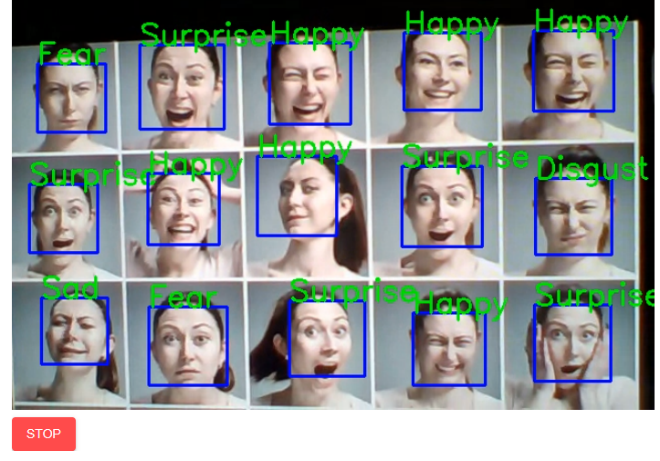


Fig 11: Multiple face emotion recognition

VII. DEPLOYMENT

There is two patterns for detecting and predicting single faces and as well as multiple faces using OpenCV video capture in local. For web app, OpenCV can’t be used. Thus, using Streamlit-Webrtc for front-end application.

Streamlit doesn’t provide the live capture feature itself, need third party API. I have used streamlit-webrtc which helped to deal with real-time video streams. Image captured from the webcam is sent to VideoTransformer function to detect the emotion. Then this model was deployed on heroku and streamlit platform with the help of buildpack-apt which is necessary to deploy opencv model on heroku and streamlit.

VII. CONCLUSION

* Build a FERweb app using streamlit and deployed on Heroku, with live webcam detection.
* The model created with CNN layers gave training accuracy of 68%and validation accuracy of 64%after 50 epochs.
* Difficult to classify disgust images
* Model also work for multiple face detection.

**REFERENCES:**

1. R. G. Harper, A. N. Wiens, and J. D. Matarazzo, Nonverbal communication: the state of the art. New York: Wiley, 1978.
2. P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” Journal of Personality and Social Psychology, vol. 17, no 2, p. 124‑129, 1971.
3. <https://www.kaggle.com/msambare/fer2013>
4. https://towardsdatascience.com/face-detection-recognition-and-emotion-detection-in-8-lin es-of-code-b2ce32d4d5de
5. <https://towardsdatascience.com/video-facial-expression-detection-with-deep-learning-applying-fast-ai-d9dcfd5bcf10>
6. <https://github.com/atulapra/Emotion-detection>
7. https://medium.com/analytics-vidhya/building-a-real-time-emotion-detector-towards-machine-with-e-q-c20b17f89220