Face Emotion Recognition

Shubham Srivastava

Data science trainee

AlmaBetter, Bengaluru

Abstract:

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. It provides data in the form of video, audio, and texts which can be analyzed using deep learning algorithms

We will solve the above-mentioned challenge by applying deep learning algorithms to live video data.

Our experiment can help understand the overall student's sentimental behavior in the class. whether they found the class interesting or not which help to improve the web learnings.

Keywords: Deep learning, CNN Keras, batch size, epochs, transfer learning

1.1 Problem Statement and Introduction:

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 rapidly, there are major challenges associated with digital learning 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 can 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 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 analysed 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 analysed and tracked.

1.2 Project Mission:

This project is part of the research performed under Almabetter. The Almabetter is interested in the design algorithm and the develop a app that are able to learn to interact socially with humans. The idea is that society can benefit from the use of webcam in areas such as education, e-learning, health care, and assistive technology. Technically, the project's goal consists on training a deep neural network with labeled images of static facial emotions. Later, this network could be used as part of a software to detect emotions in real time. Using this piece of software will allow webcam to capture their interlocutor's inner state This capability can be used by machines to improve their interaction with humans by providing more adequate responses. Thus, this project fits the purpose.

Finally, this is a multidisciplinary project involving affective computing, machine learning and computer vision. Learning how these different fields are related, and to understand how they can provide solutions to complex problems is another project's goal.

2.1 Model Building

This phase consisted on the use of a facial emotion labeled data set to train a deep learning network. Additionally, evaluations were performed on several network topologies to test their prediction accuracy. The use of convolutional neural networks on the topologies was preferred given its great achievements on computer vision tasks

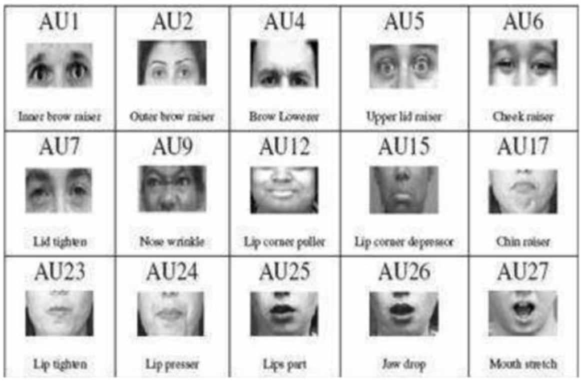
The idea is to make a comparison on both data sets, and evaluate the generalization property of the network. Also, a focus on some parameters and its effect on the model's accuracy prediction was performed. These parameters were chosen because their influence over the network's behaviour:

- Network loss
- Learning rate
- Dropout
- Epochs
- Optimers

2.2 Transfer learning:

In transfer learning, the knowledge of an already trained machine learning model is applied to a different but related problem. For example, if you trained a simple classifier to predict whether an image contains a backpack, you could use the knowledge that the model gained during its training to recognize other objects like sunglasses.

2.3 Affective Computing:



Face actions units.

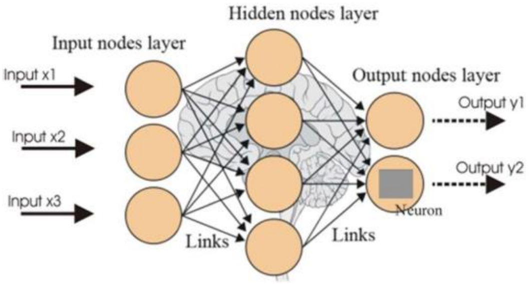
Empathy is a human capacity that makes us aware and provides us with understanding about what other beings might be experiencing from their current's position it is very important for affective computing to develop ways to properly measure these particular modulations since they can lead to a better understanding of a subject's emotional state. The two main ways to do so is by detecting facial and vocal emotions. However, in this project, only facial emotions were used.

2.4 Facial Emotion Background

The work by psychologist Paul Ekman has become fundamental to the development of this area. Nowadays, most of the face emotion recognition studies are based on Ekman's Facial Action Coding System With respect to computers, many possibilities arise to provide them with capabilities to express and recognize emotions. Nowadays, it is possible to mimic Ekman's facial units. This will provide computer with graphical faces that provide a more natural interaction [30]. When it comes to recognition, computers have been able to recognize some facial categories: happiness, surprise, anger, and disgust

2.5 Artificial Neural Networks

An ANN can be explained trough the following three steps:
1. Input some data into the network.
2. Transformation over the input data is accomplished by means of a weighted sum.
3. An intermediate state is calculated by applying a non-linear function to the previous transformation. From the previous steps, it can be said that all of them constitute a layer. A layer represents the block with the highest-level on a network. The transformation is usually referred as a unit or neuron, although the latter is more related with neurobiology. Finally, the intermediate state acts as the input into another layer or into the output layer



3 Steps involved:

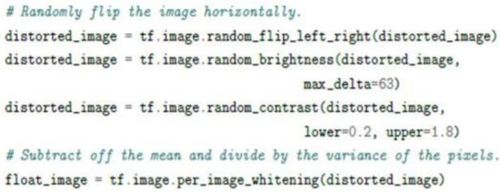
- **Exploratory Data Analysis**

  After loading the dataset, we performed this method by comparing with the different emotions in the data, and their label with display pictures of random choice

  Data augmentation

  One of the main drawbacks behind supervised learning is the need of labeled data.

  The manual work involved on data labeling demands many people following a strict set of rules The bigger the dataset, the more complex to label it. Deep Learning requires big amounts of data for training. Since this is a very expensive task, data augmentation has been proven an efficient way to expand the dataset by following:



```
# Randomly flip the image horizontally.
distorted_image = tf.image.random_flip_left_right(distorted_image)
distorted_image = tf.image.random_brightness(distorted_image,
                                             max_delta=63)
distorted_image = tf.image.random_contrast(distorted_image,
                                           lower=0.2, upper=1.8)
# Subtract off the mean and divide by the variance of the pixels.
float_image = tf.image.per_image_whitening(distorted_image)
```

- **Encoding of categorical columns**

  Since your training data was already correctly labeled into 7 sections of emotions

- **Dropout**:

1. Dropout minimizes the impact of units that have a strong activation. This method
2. shutdowns units during training, so other units can learn features by itself.
3. Providing with more independence to all units reduces the strong unit bias leading
4. to strong regularization and better generalization.

- **Fitting different models**

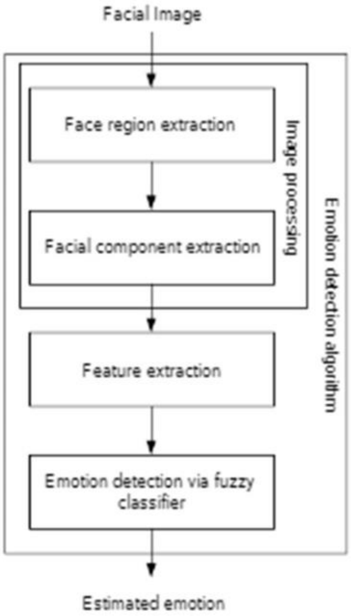
  For modelling we tried various classification algorithms like:

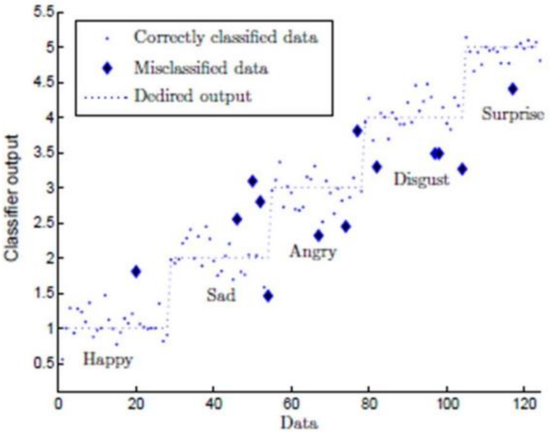
CNN with Keras
DeepFace
Transfer learning

- **Tuning the hyperparameters for better accuracy**

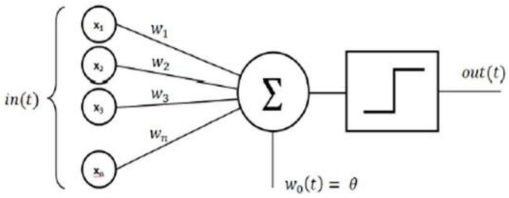
Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of number of layers, Mapping, dropout, epochs, batch size etc.

4.1 Algorithms:


Facial Image
Face region extraction
Facial component extraction
Image processing
Feature extraction
Emotion detection via fuzzy classifier
Emotion detection algorithm
Estimated emotion

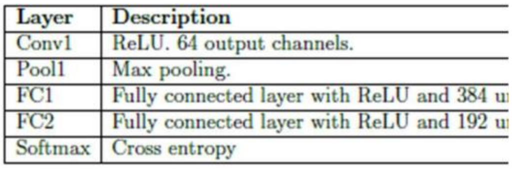


4.2 ANN:



$$y = f(t) = \sum_{i=1}^{n} X_i * W_i | + \Theta$$

The latest reincarnation of ANN is known as Deep Learning (DL). According to Yann LeCun, this term designates “any learning method that can train a system with more than 2 or 3 non-linear hidden layers.



Layer	Description
Conv1	ReLU. 64 output channels.
Pool1	Max pooling.
FC1	Fully connected layer with ReLU and 384 u
FC2	Fully connected layer with ReLU and 192 u
Softmax	Cross entropy

4.3 Rectified linear unit:

The activation function of a unit (neuron) is an essential part of an ANN architecture.

The use of different functions has been used by researchers since ANN early days. the step function was introduced as the activation function. However, the binary nature of the step function does not allow to have a good error approximation.

$$f(x) = max(0, x)$$

4.4 Use of GPU:

The use of GPU for training has become fundamental for training deep networks because of practical reasons. The main reason is the reduction of the training time compared to CPU training. While different speedups are reported depending on the network topology, it is common to have around 10 times speed when using GPU.

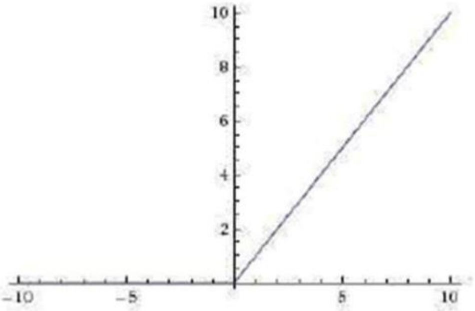


Figure 2.4: Rectified Linear Unit (ReLU) [3]

The difference between CPU and GPU is how they process tasks. CPU is suitable to perform sequential serial processing on few cores. On the other hand, GPU encompasses a massive parallel architecture. This architecture involves thousands of small cores designed to handle multiple tasks simultaneously

4.5 CNN:

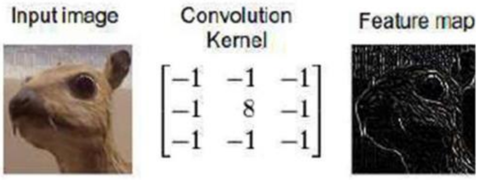


Figure 2.6: Convolution operation [3]

everything that is not important for the feature map, only focusing on some specific information.

In order to execute this operation, two elements are needed:

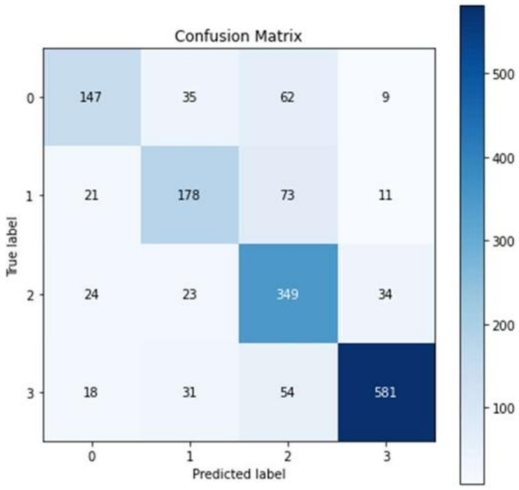
- The input data
- The convolution filter (kernel)

for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits). XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

5. Model performance:

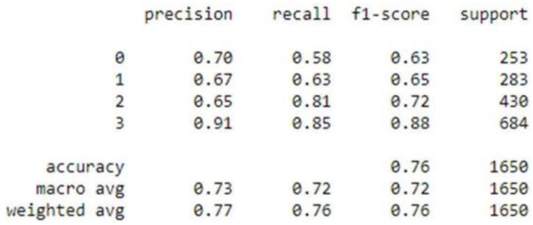
In Transfer Learning Model can be evaluated by various metrics such as:

1. **Confusion Matrix-**



The confusion matrix is a table that summarizes how successful the classification modelis at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

2. **Precision/Recall-**



```
              precision    recall  f1-score   support

           0       0.70      0.58      0.63       253
           1       0.67      0.63      0.65       283
           2       0.65      0.81      0.72       430
           3       0.91      0.85      0.88       684

    accuracy                           0.76      1650
   macro avg       0.73      0.72      0.72      1650
weighted avg       0.77      0.76      0.76      1650
```

Precision is the ratio of correct positive predictions to the overall number of positive predictions: TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

3. Accuracy-

3. **Accuracy-**

Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

4. **Area under ROC Curve(AUC)-** ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

6. Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

1. **Epochs:** In terms of artificial neural networks, an epoch refers to one cycle through the full training dataset
   #Epochs = 150
2. **Batch Size:** It refers to the number of training examples utilized in one iteration. higher batch sizes lead to lower asymptotic test accuracy
   #Batch size = 128
3. **Learning rate:** The learning rate is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function.
   # Learning rate=0.0001

7. Conclusion:

That's it! We reached the end of our exercise. Starting with loading the data so far, we have done EDA, Data Augmentation, model building using various methods and hyper tuning.
In all of these models our accuracy revolves in the range of 82-97%.
And there is improvement in accuracy score even after hyperparameter tuning. So Implementation of CNN layer learning was the accuracy of our best model is 82% which can be said to be good for this large dataset.

References-

1. Daniel Llatas Spiers Research paper
2. GeeksforGeeks
3. Analytics Vidhya