2022 2nd International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 24-26, 2022

Music Recommendation System based on Emotion Detection using Image Processing and Deep Networks

Oindrella Ghosh
Sardar Patel Institute of Technology
oindrella.ghosh@spit.ac.in

Dr. Reena Sonkusare
Sardar Patel Institute of Technology
reena_kumbhare@spit.ac.in

Sudhanshu Kulkarni
Sardar Patel Institute of Technology
sudhanshu.kulkarni@spit.ac.in

Sanskar Laddha
Electronics and Telecommunication
sanskar.laddha@spit.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847888

***Abstract*—There are several techniques in which a user can decide the music they want to listen to from a variety of collections. Playing music based on the mood of the user is an application of deep learning which is introduced to the listeners. This can be done by figuring out the user's facial expression which depicts their mood. Expressions talk a lot more about humans than words do. Due to increased numbers of songs being produced every day, it is becoming difficult for the users to figure out the song they would want to listen to. Our Recommendation System aims to solve this problem by providing users a list of songs directly into their personalised playlist based on the mood. The aim is to build a music recommendation system using image processing which is based on neural networks. The novelty lies in the fact that we plan to leverage a set of model architecture which has been developed by the scientist at Meta. The Multitask Cascaded Convolution Neural Network(MTCNN) and FaceNet Architecture have been used for face detection and recognition through the embedding generated. We then run a Convolution Neural Network Model to predict the emotion. Finally a classifier is used to recommend music based on their emotion from a spotify dataset which consists of almost 6 hundred thousand songs. The user can set the number of songs he/she wants to be recommended. We will also add the recommended songs back to user playlist.**

***Index terms*—Face Detection, Face Embedding, Face Extraction, Mood Predictions, Music Recommendation**

I. Introduction

Human beings have a tendency of judging a person's mood based on the facial expressions. Applications having great utility can be developed if this ability of humans is mastered by computers or other electronic devices using deep learning. People have always enjoyed music. It can make us expressive and aids in analysing our feelings and emotions in a better way. Music does wonders for our mood. For example, if we want to feel ecstatic, we can listen to songs that are happy and vice versa. In case of an unfortunate event, listening to sad songs may help. According to science, listening to sad songs can actually elevate the mood. This paper proposes a similar application which is a music recommendation system using emotion detection.

Using traditional music players is quite time-consuming as users themselves have to have to go through their song list and find music that would suit their mood. This task is laborious and one can often be found facing a dilemma of selecting the perfect song based on the mood. Other recommenders which detect the user's mood have high complexity which affects the real time performance of the application. Our product will not only find emotion of the user with the help of their facial expressions but also provide the user with any number of songs the user desired in a personalized spotify playlist.

II. Literature Review

In the paper [1] the authors proposed a system which uses the algorithm of point detection for the extraction of features from the input images and the classification algorithm OpenCV for the purpose of training the input images for facial emotion detection. Image is taken from webcam and the extracted image is subjected to pre-processing. Edge detection technique is applied using canny edge detection. Segmentation is applied to the edge detected image. After this face detection takes place followed by feature extraction. This is then deployed on web services.

A VGG-16 CNN based facial expression recognition[2] is also developed. Once the emotion was detected, the song matching the user's emotions would be played from the user's personalised playlist.

The researchers in their work [3] developed a new model for facial image recognition which makes use of three Convolution neural networks cascaded together to recognize face and detect facial landmarks. Each of the CNN layers in the model detects different facial feature which makes processing of facial image easier.

The authors in the paper [4] design a system which can detect the facial expressions of the user. They used clustering techniques after facial recognition and detection. It uses deep convolution networks along with triplet loss to achieve state of the art accuracy.

The paper [5] discusses a system which detects the emo-

tions. On identifying negative emotion of the subject, suitable playlist for improving the mood is recommended. If positive emotion is detected, an appropriate playlist for boosting the positive emotions is recommended.

III. Methodology

Our proposed methodology primarily has three stages. Face detection and recognition, Mood Prediction and thirdly music recommendation. Each of the following steps have been detailed in the given section.

A. Face Detection and Recognition

Face detection is the process in which a single face is extracted from an image. This can be extracted by drawing a box around the face or just windowing the face from the image.

Multi-Task Cascaded Convolution Neural Network(MTCNN) which detects faces and contours like the facial landmarks from images has been used to create a face detector model. MTCNN is a three stage neural network where each stage has a network to perform some sub-task. The final stage or the third stage gives the face detected image. The three stages consist of multiple CNN's with different complexities.

The three stages of MTCNN can be simplified as follows:

- The first stage neural network is called Proposal Network abbreviated as P-Nets. It creates several segments or frames as the image is scanned through this network. Creating such divisions makes it easy for the further networks to process the facial image. P-Net is a shallow convolution neural network with a feed-forward network attached to it.

- The output of proposal networks are fed into the next neural network called Refinement networks or R-Nets. This networks rejects most of the frames which does not contain facial data.The network refines the image by discarding non facial frames.

- The third stage network is the most sophisticated of the three neural networks, It is a powerful network which gathers the pixel information of the facial landmarks like eyes, nose, cheeks, lips and ears, etc, thus detecting a complete face. This is called the Output Network (O-Net).

It assumes that the photograph contains only one face and returns the detected face after defining a lower left corner and width/height. A function accumulates all the faces detected in the photographs present in the specified path. Further, the dataset of the Indian celebrities is loaded. An array of detected faces and their respective labels is obtained.
FaceNet is a face recognition system that was developed at Google in 2015 [5]. This was also a CNN based architecture which made use of triplet loss function in training. The embeddings or vectors on the face map of the same ID or the same person were similar as compared to the different IDs (having more distance between the vectors). This is a system that takes facial images, extracts high quality features, and gives the predicted facial embedding. Face embedding is a vector of 128 elements that is a representative of unique facial features like facial landmarks. A linear support vector machine (SVM) [6] classifier predicts facial identities. The classifier model takes the embedding as input and predicts the identity of the face. The model is trained using this embedding. The inputs are processed via normalization and the target variable for each celebrity name must be converted to an integer. Linear SVM is applied to the training data by using a linear kernel. Fig 1. shows the input image and how the face is detected using MTCNNs and FACENETs. The input image is 400x375x3 while the detected face is 160x160x3 which is also the input to the following convolution neural network.

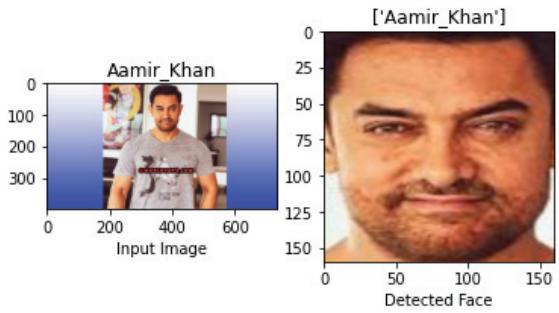

Fig. 1: Face is detected and extracted from a raw input image

B. Mood Prediction

In order to identify the facial expression it is important that we first detect the face in an image. Once we have face being identified we can proceed with mood prediction.This task involves a brief data preprocessing and augmentation which ensures that the images are properly rescaled, shifted and rotated.The next step involved is passing the training images into a Convolution Neural Network[7] which will extract pixel information and finally predict the mood shown by the person in the image.This the heart of this step.The stages of this network are explained below.Fig 2. shows the various steps involved in the process.A point to be noted here is that there are multiple convolution,Max Pooling and Dropout layers followed by the remaining layers in the end.

1) Convolution Layer: The convolutional layer is a major part of the CNN and performs most of the calculations. This layer performs a mathematical operation called dot product between input image and filters.Which are in the form of matrices. The first matrix is an image matrix that contains the pixel values of the image (typically 0-255), and the second matrix is called the filter or kernel. Function detection. Most images consist of 3 channels (RGB), so the kernel and image matrix also cover 3 channels. An illustration of the convolution operation is shown below.

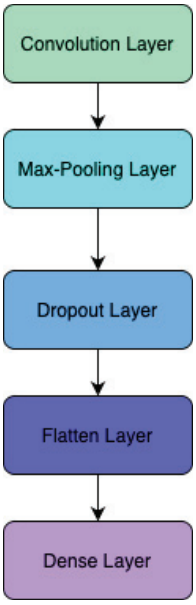

Fig. 2: Flow of Mood Prediction Model

2) Pooling Layer: Dimension reduction of the previous output layer is the primary use of pooling layer. Reducing the size of the feature matrix reduces computational time and makes features robust to noise and outliers. There are several ways to do pooling: B. Maximum and average pooling. However, the most common is Max Pooling. It calculates the maximum output from a small matrix and moves further through the smaller windows of the matrix.

3) Dropout Layer: The Dropout layer is a masking layer that makes contributions of few neurons to zero and leaves unmodified all others. We can apply a Dropout layer to the feature matrix, in which case it nullifies some of its features; but we can also apply it to a hidden layer, in which it makes contribution of few neurons to zero.Dropout is a great technique to reduce overfitting and improve accuracy.

4) Flatten Layer: A Flatten layer reshapes the feature matrix to have a shape that is equal to the number of pixels contained in the feature matrix.This is the same thing as making a one dimensional array of elements

5) Dense Layer: Dense layer is a artificial neural network layer connected after CNN. It is the most common and frequently used layer.This layer comprises the artificial neural network which is added at the end of convolution layer to get final prediction.

6) Activation Function and Optimizers: Activation functions is mentioned in convolution layers which adds non-linearity to the output which is required while working with image data.We have trained the model on Exponential Linear Unit (eLu)[8].The mathematical expression for the function is shown in the equation given below. Optimizers are mathematical functions and algorithms which help us in reducing loss functions by changing the values of weights and biases. The weights are initialized with some initial functions. In our case we have used a he normal intializer which initializes the weights from a truncated normal distribution centered around zero.

$$f(k) = \begin{cases} k & : if\ k > 0 \\ \alpha(exp(k) - 1) & : else \end{cases} \quad (1)$$

C. Music Recommendation

To get access to the raw data of different songs we made use of Spotify Web API. Music recommendations use Spotify data, including track names, IDs, song features, artist names, and more. We divide the data into clusters and use them to collect the data points (in this case songs) into a series of songs that will be the playlist.As we are using K-Means[9] we specify the number of clusters we intend to divide our song dataset. Elbow Method is one of the method used to make these clusters more precise which can otherwise be inaccurate. Different cases for varying number of cluster is calculated and then plotted on a line graph.

To boost model performance it's important to pre-process our data. We used MinMaxScaler for pre-processing our data.The main goal of the K Means clustering is to group the songs together which are similar.To make these predictions easier to grasp, we have transformed them into a data frame and concatenated it to our original dataset as a new column.Since,Light Gradient Boosting Machine (LGBM) [10] can handle the large size of data and runs on low memory, we used it for training our song recommender model and finally we then rank songs according to the popularity.The most common emotions are considered which are happy, sad and angry all other emotions associated with music overlap with these three and thus lead to false predictions.Then the emotion is mapped to an integer code i.e. 0,1,2 .If emotion is Sad then assigned code is 0 and likewise for happy mood it is 1 and angry is 2. Number of song recommendation is also taken as an input from the user ultimately displaying the predicted songs according to the mood.

IV. System

To extract data from Spotify account service, intially a test application must be developed on spotify developer's site. All requests must go through this application. To create new playlists on Spotify we used the library Spotipy that helps to automate processes and get a lot of different features of a wide variety of tracks. We just require Client ID, Client Secret, and Username Number to use the Spotify's API and manipulate our library music and account data. Authorization scope of various methods we would like to automate should also be specified. Fig 3. shows the workflow of the system.

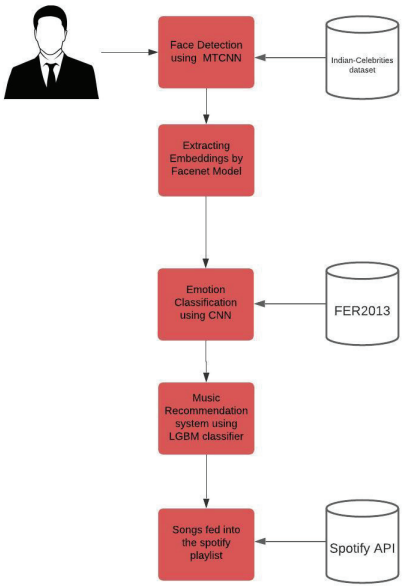

Fig. 3: Flowchart for the proposed system

Algorithm	Training Accuracy	Test Accuracy
VGG 16	97.899%	69.1418%
HAAR Cascade+CNN	-	90.23%
Segmentation+Point Detection	-	90.23%
FaceNets+SVC+CNN	99.8%	99.6%

TABLE I: Accuracy Score of various mood detection models

V. Results

Accuracy Score is a Sklearn metric which gives fraction of number of correctly classified samples. The model trained on mood prediction gave an accuracy score of 99.8%. The table [1] shows various mood prediction algorithms previously developed and their performance on the given matrix. We can observe that our method performs the best on both training and test data. The deep CNN is trained to monitor on validation loss. To prevent over-fitting model is stopped from training as soon as the accuracy remains unchanged for three consecutive epochs.

The LGBM classifier which is used in classification of music gives 99.8% accuracy train and 99.6% on test data. This classification is executed after applying k-means clustering on music dataset. The classifier segregates the songs into three classes, namely Sad, Happy and Angry which are labelled as 0,1 and 2 respectively. The key parameters for classification of songs include feature like energy, loudness etc. Fig 4. shows the confusion matrix which shows the performance of the classifier in predicting the emotion and recommending the appropriate song.The confusion matrix shows that the classifier recommended songs for 'Sad' emotion with an accuracy of 99.18%, Songs for 'Angry' mood was predicted with an accuracy of 99.54% and 'Happy' Song with an accuracy of 99.51%.

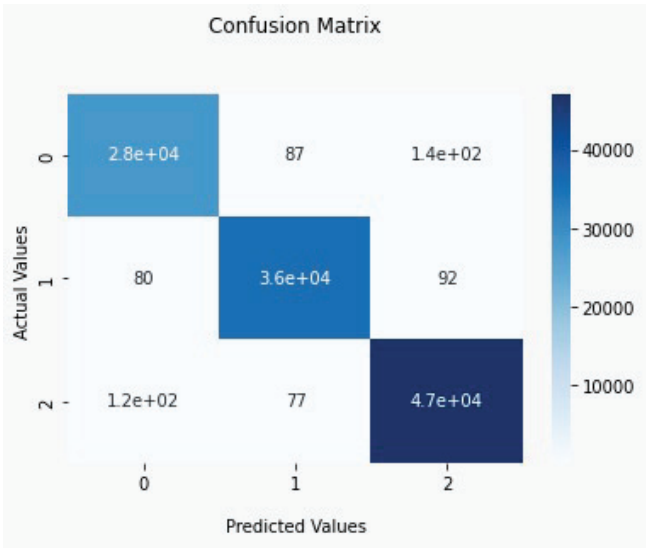

Fig. 4: Flowchart for the proposed system

VI. Conclusion

The aim of our study was to build a deep learning based system for recommending song to the user based on their mood which can be fed into playlist directly.We intended to achieve this by incorporating facial detection systems. Music recommendation has been one of the latest areas of exploration in deep learning. Various tools and technologies have been developed and advancements have been made to improve these systems. Our system is one such addition to it. A huge range of image processing technologies have been developed to support facial expression recognition. In addition to the theoretical background, our contributions provide an approach for sketching and implementing emotion-based music players. The proposing system can process facial images to recognize basic emotions, play music in response to these emotions, and propose music that further improves the user's mood. We have tried to ensure that the detected songs are put directly into the Spotify playlist. Since Spotify is one of the widely used streaming music platforms, it also provides various developer tools that makes integrating with other systems hassle-free. In future, we would like to focus on building music recommender which can detect multiple emotions not only from an image but also from videos or live camera feed. So that this can be incorporated in electronics and automotive systems like mobile phones,computers and cars etc.

References

[1] ShanthaShalini. K, Jaichandran. R, Leelavathy. S, Raviraghul. R, Ranjitha. J and Saravanakumar. N (2021), "Facial Emotion Based Music Recommendation System using computer vision and machine learning techniques", Turkish Journal of Computer and Mathematics Education Vol.12 No.1 (2021), pp. 9012-917, 05 April 2021.

[2] G.Chidambaram, A.Dhanush Ram, G.Kiran,P.Shivesh Karthic and Abdul Kaiyum (2021),"Music Recommendation System Using Emotion Recognition", International Research Journal of Engineering and Technology (IRJET) Vol. 08 Issue: 07, July 2021.

[3] A. Alrihaili, A. Alsaedi, K. Albalawi and L. Syed, "Music Recommender System for Users Based on Emotion Detection through Facial Features," 2019 12th International Conference on Developments in eSystems Engineering (DeSE), 2019, pp. 1014-1019, doi: 10.1109/DeSE.2019.00188.
[4] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, Vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
[5] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.
[6] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.
[7] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.
[8] Clevert, Djork-Arné Unterthiner, Thomas Hochreiter, Sepp. (2015). "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". Under Review of ICLR2016 (1997).
[9] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.
[10] M. Osman, J. He, F. M. M. Mokbal, N. Zhu and S. Qureshi, "ML-LGBM: A Machine Learning Model Based on Light Gradient Boosting Machine for the Detection of Version Number Attacks in RPL-Based Networks," in IEEE Access, vol. 9, pp. 83654-83665, 2021, doi: 10.1109/ACCESS.2021.3087175.