WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Emotion Based Music Recommendation System Using CNN

Ashwini Jadhav[1], Nikita Bhaise[2], Mihir Narwade[3], Ruchita Phalke[4], Yash Talele[5]

[1,2,3,4,5]Dept. of Computer Engineering, DYPIT.

ABSTRACT:

The method of identifying a person's emotions based on different facial clues and visual data is known as emotion detection. Since deep learning became so popular, this discipline has flourished. Additionally, numerous hitherto unimagined uses have been made possible through emotion detection. Music is one of the things that has a strong emotional connection. Music can evoke specific emotions in the listener, and someone experiencing a certain emotion might search for a song that evokes that experience. We link these feelings with a music player that plays music that enhances user experiences using our emotion recognition model. Two convolutional neural network (CNN) models, a five-layer model and a global average pooling (GAP) model, are included in the model we created. These CNN models were integrated with transfer-learning models. We employed three pre- trained models for our transfer-learning models: ResNet50, SeNet50, and VGG16. Our results are equivalent to those of cutting-edge models, but our models perform better overall.

Key Words: Class Weighting, Convolutional Neural Network, Emotion Detection, Ensemble, Global Average Pooling, Transfer Learning

I. INTRODUCTION:

A rapidly growing area in deep learning is facial emotional recognition (FER) [1]. Based on the visual face, FER seeks to anticipate the emotion of the face. Since facial visual cues are not always clear, collecting them from the face is difficult during the sophisticated process of emotion identification. They may even be absent in some circumstances or very subtly present. There are models that can accurately identify emotions, but they can only do it in a controlled setting. The challenge increases significantly in a real-world setting when we must take into account lighting, various facial characteristics, occlusions, head attitude, etc. But over the past ten years, this field has substantially improved, performing far better than average people. This has led to the rise of various applications such as medical treatments, social robotics, driver fatigue surveillance, etc. Our main contribution is on improving a FER model and applying it in the real world. A model that we created combines two standard CNN models with a few transfer learning algorithms. We gave the two vanilla CNN a name. Approaches like the five-layer model and the later-discussed global average pooling (GAP) model. ResNet50, SeNet50, and VGG16 are the transfer learning models that we used; further information on them will be provided later. The GAP model stands out from the competition since it cuts the parameters almost 80% while keeping a respectable level of accuracy. This little model's ease of mounting on portable devices assists us in the real world. We also looked into several different ideas for how to improve the model. Our investigation covered data augmentation, class weighting, the addition of auxiliary data, and ensembling. In order to make music recommendations that are in tune with the user's mood, we want to learn how effectively we can interpret their emotions. Although the areas of facial recognition and music suggestion have been well studied, the two together have not. This is what we wish to investigate. The order of the remaining text is as follows. In Section II, some current models that deal with the issue of face expression recognition are reviewed. Section III provides a thorough explanation of our strategy. The outcomes of our studies are described and discussed in Section IV, which is followed by Section V's conclusion.

II. RELATED WORK

The FER2013 dataset was developed by Goodfellow et al. [2] as a Kaggle challenge to advance emotion recognition research. Some type of convolutional neural network with image manipulation was employed by the top three teams. Yichuan Tang [3], the champion, with a 71.2% accuracy rate. Both an L2-SVM loss function and an SVM loss function were employed. The use of these loss functions was unique at the time and producedperformance that was exceptional. Additionally, the model performs better on benchmark datasets including ICML-2013 [2], MNIST [4], and CIFAR-10 [5.

The paper's flaw is that it doesn't investigate different multi-class SVM formulations. A recent assessment by W. Deng and S. Li [6] delves further into the current state of deep learning's application to FER. The study by D. V. Sang et al. [7] is another one that we would like to mention because it had a significant impact on the work completed by us. The main goal of this study was to automate the extraction of semantic information from faces using convolutional neural networks rather than manually developing feature descriptors. It significantly outperforms the competition winner when used with the FER2013 dataset. The study explains this by stating that they experimented with different pairings of training strategies and loss functions. Additionally, the method's lower parameter count and increased efficiency are both mentioned in the paper. This is significant because it makes the method acceptable for real-time systems. An ensemble model of six state-of-the-art Convolutional Neural Networks (CNN)-based predictors was built

by Pramerdorfer et al. [8] and reached a performance of 75.6%. The study takes significant pains to pinpoint bottlenecks. The paper's problem, though, is that it doesn't offer any techniques for data augmentation. Nearly one million photos from the ImageNet dataset were categorised into its many classes using a deep CNN model developed by Krizhevsky et al. [9]. It only used supervised learning and yet had amazing performance. The problem with this model is that if we remove any convolutional layer, the model's performance suffers significantly. A facial emotion detector model is created by Z. Yu et al. [10] employing three states of the art predictors. This model is capable of achieving state-of-the-art performance, but randomly initialising one model results in slightly subpar performance. The concept of taking into account the aligned and non-aligned states of the face was employed by B-K Kim et al. [11] to improve prediction accuracy. However, emotion detection cannot benefit from this research. It is better suited for facial recognition instead.

III. METHODOLOGY

a) Models

Our primary objectives when developing the model architectures were to maintain the real-time component while maximising the accuracy of the test set for the FER2013 dataset. Thus, when we built our models, we tried to strike a compromise between accuracy and the overall amount of parameters. Three transfer learning models and two CNN models that we created are detailed below. The class imbalance in the FER dataset was the main obstacle. We investigated class weighting to solve this issue, and the outcomes were encouraging. Finally, by combining all of these distinct models, we reached our maximum accuracy of 76.12%.

1) Model1-Five-Layer Model:

This model has five layers, as the name would imply. Convolutional and max-pooling layers make up each of the first three stages, which are followed by a fully connected layer with 1024 neurons and an output layer with 7 neurons and a soft-max activation function. The initial convolutional layers used 32, 32, and 64 5*5, 4*4, and 5*5 kernels respectively. Following these convolutional layers are max-pooling layers, each of which employed ReLu as the activation function and used kernels with dimensions 3*3 and stride 2. At each layer, batch normalisation was added. 30% dropout following the final layer with full connectivity. To further enhance performance, this was done. Fig. 1 displays a visual representation of the model architecture. We ran 350 epochs of training on the model. The optimizer was Stochastic Gradient Descent (SGD) [12] and the loss function was cross-entropy. The learning rate was set to 0.01 and the batch size was 256.

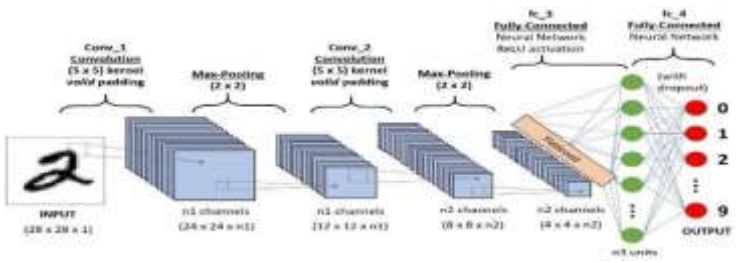

Fig. 1: 5-Layer Model Architecture

Model 2 - Global Average Pooling (GAP) Model:

The enormous number of parameters that are typically present in a standard convolutional network are to be reduced in our second proposed CNN design while yet keeping a respectable level of accuracy. This is required to create a quick real-time CNN and close the performance gap with real-time architectures. By removing the final fully connected layers, it lowers the total number of parameters. Most CNNs typically have completely connected layers at the end that contain more than 80% of the total parameters. Fig. 2's description of the model architecture is accurate. The model comprises of a conventional fully convolutional network with nine convolutional layers, four batch normalisation layers, and a final GAP layer. This last layer is where we get the name of the model. It has 642,935 parameters, of which 641,463 are trainable, which is significantly fewer For a task like this, conventional deep learning models with more than 2 million parameters are used. After each convolutional stage, dropout is employed to regularise the network. The fully connected layers are eliminated by the model using GAP. This is accomplished by using the soft max activation function on each of the feature maps in the final convolution layer, which has the same number of feature maps as the classes in the dataset we wish to forecast. An ADAM optimizer is utilised to train the model.

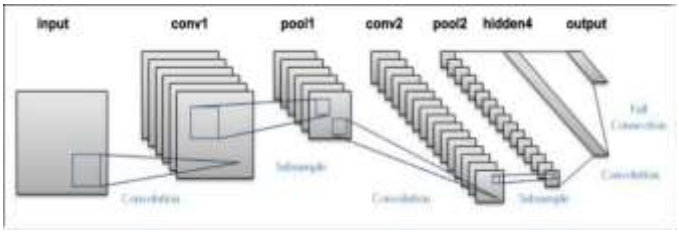

Fig. 2: GAP Model Architecture

b) Transfer Learning

The significant challenges of the FER2013 dataset were the small size of the dataset and imbalance in the classes. The accuracy of such a dataset was improved with the use of transfer learning. We used SeNet50 [13], ResNet50 [14], and VGG16 [15] as the pre-trained models. These sophisticated models include a lot of parameters, and they are known to perform well when captioning images. Therefore, we used it for our purpose of recognising face expressions. For the same, we made use of the Keras library. The 48*48 grayscale photos from FER2013 have to be resized and recolored to fit the input specifications for each transfer learning model during training.

1. Model3-Fine-Tuning ResNet50:

The first transfer learning model investigated was ResNet50. Residual Network is referred to as ResNet. There are fifty layers in this model. The initial step was to resize the photos from the FER2013 dataset during training to satisfy the input requirement of this transfer learning model. After removing the original output layer, we added two completely connected layers with a total of 1024 and 4096 bytes each. Finally, a 7-layer output layer with a softmax activation function was added (one for each class). The initial few layers of ResNet were frozen in order to further optimise accuracy, but the remaining layers were remained trainable. The optimizer used Stochastic Gradient Descent (SGD) [12]. Batch size and learning rate were fixed at 64 and 0.001, respectively. This model was trained for 100 epochs and achieved 73.22% accuracy on the test set.

2. Model4-Fine-Tuning SeNet50:

The second transfer learning model investigated was SeNet50. Squeeze and Excitation Network is known as SeNet. There are fifty layers in this model. It has a structure similar to ResNet50. So, using the same batch size and learning rate as ResNet50, we trained the model using these parameters. After 150 training iterations, this model's accuracy on the test set was 72.19%.

3. Model5-Fine-Tuning VGG16:

VGG16 was the third and final transfer learning model explored. This model is much shallower than ResNet50 and SeNet50 as it consists of only 16 layers. However, VGG16 is more complicated in its architecture and has much more parameters. We froze all the pre-trained layers and added two fully connected layers of 1024 and 4096 with fifty percent dropout. Adam optimizer was used while training. Batch size and learning rate were fixed at 128 and 0.01. This model was trained for 120 epochs and achieved 69.30% accuracy on the test set

c) Implementing The Models

1. Data Preparation:

The FER2013 dataset exists in a number of variations. These vary in terms of directory organization, image size, and labelling. To address these discrepancies, we divided all the input datasets into seven directories, each of which represented a different class in the FER2013 dataset. During training, Keras data generators were used to resize the photos once they had been loaded in batches from disc.

2. Data Augmentation: To expand the dataset and boost accuracy, we used data augmentation.

Horizontal mirroring, image zooms, degree rotations, a nd horizontal/vertical shifting were a of the data augm entation methods used.

3. Class Weighting:

The imbalance in the amount of samples for various cla sses was one of the main issues with the FER2013 datas et.By using class weighting [16] that is inversely proport ional to it, we may correct this.Performance was signific antly improved, especially in the disgust class where th e misclassification rate fell from 62% to 35%.

$$W_{n,c} = \frac{1}{Number\ of\ Samples\ in\ Class\ c}$$

4. Ensembling: To increase the accuracy of our top test, we did ensembling using soft voting of the five models.

d) Web-App

The Web-App was ran using the Chrome browser programme. This application's flow is as follows:

1) Start the capture.py programme. It will then launch the file, which will display the music player built using python (webpage)
2) Click on the play button next to any song to start listening, or use the + sign to add it to the queue.
3) On the right upper side, a different option based on emotion will be displayed in frame; choose it. The python function will be triggered.
4) The camera will begin recording the back image and attempt ten successful shots to capture any face.
5) Predict the emotions based on those images; then, compile all ten results; select the suitable emotion.
6) Player a randomly selected song from that genre.
7) To prevent the user from being aware of the end of a song, the identical back-end process is repeated.

IV. EXPERIMENT AND ANALYSIS

A. data set

We used the FER2013 dataset [17] for this research.

A well-liked and intricate benchmark dataset called FER2013 is used in numerous competitions and studies. Human accuracy is 64.5 percent. It is made up of over 36,000 grayscale photos with a 48x48 normalisation. Each of the seven classes that make up the photos represents a different face emotion. Happy (8,988), Neutral (6,197), Sad (6,076), Angry (4,954), Surprise (4,001), Fear (5,120), and Disgust are the various classes that are offered. (548). The great degree of class disparity makes this a very difficult dataset.

C. Accuracy

Table I :

displays both the accuracy of the ensemble model and the accuracy of the individual models. Because of the complexity of their design, transfer learning models have higher accuracy. As a result, there are a lot more parameters to consider, which is not a favourable trade- off for practical applications. We have further demonstrated the accuracy attained by class weighting, when the individual model accuracy does not always increase. However, it illustrates how handling the class imbalance issue in our dataset improved the ensemble model's overall accuracy. After using class weighting, our ensemble model had the maximum test accuracy of 76.12%.

Model	Accuracy	Class Weighted
ResNet50	73.22%	72.28%
SeNet50	72.19%	70.99%
VGG16	69.30%	69.15%
5-Layer Model	65.67%	-
GAP Model	66.54%	-
Ensemble	**74.84%**	**76.12%**

Table II

The accuracy and total number of parameters for our model for the FER2013 dataset are compared with those of other models in Table II. The findings demonstrate that, among all models, the GAP model has the fewest parameters (642,935) while still obtaining a respectable level of accuracy. In contrast to other models that take up more than 200 MB of space, the weight file for the GAP model takes up only 20 MB. Due to this, the model may be mounted even on small gadgets, which calls for such an application. Additionally, we showed the outcomes of Deep-Emotion [18] and Pramerdorfer et al. [8].

Model	Parameters (in Million)	Accuracy
Human-Level	-	64±5%
Deep-Emotion [18]	43 M	70.02%
Pramerdorfer et al. [8]	5.3 M	75.2%
5-Layer Model (Our Model)	2.5 M	65.67%
GAP Model (Our Model)	**0.64 M**	66.54%
Ensemble (Our Model)	-	**76.12%**

D. Confusion Matrix

Figure 3 displays the Ensemble model's confusion matrix. The columns match our forecasts, while the rows match the actual values. As is evident, Happy is the most successful class while Fear performs the poorest for our network. Another intriguing finding is that our algorithm incorrectly predicts the emotion of 18% of the images with a fear label as sad, which is comparable to human error on the same image.

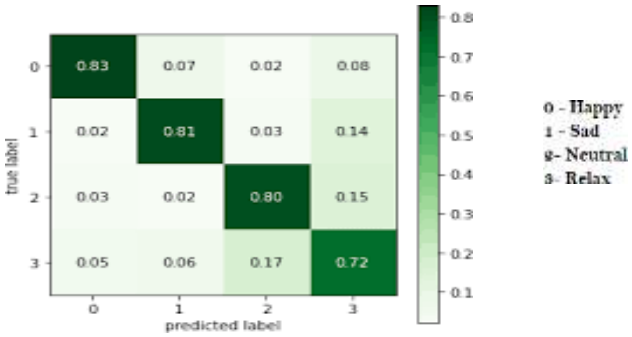

Fig. 3: Confusion Matrix

CONCLUSION AND FUTURE WORK

This study suggests an emotion recognition algorithm for mood-based music recommendations. In order to adapt to real-world scenarios, our work attempts to attain the utmost accuracy while maintaining the real- time component. We investigated a number of models that were constructed differently, such as plain CNNs and pre-trained networks based on ResNet50, SenNet50, and VGG16. The GAP model, which reduced the number of parameters by over 80% while maintaining an accuracy of 66.54%, was one model that stood out. This was a breakthrough since a model this lightweight is simple to mount on compact devices, increasing the applicability to real-world applications. Using class weighting, we were able to further resolve the difficult FER2013 dataset class imbalance problem.

Future study could increase the accuracy of our models by using landmark detection methods that exclude unimportant facial characteristics from the image while training. Through the use of a multi-label classification technique, we could handle photos with various classes of emotion better.We want to adapt our model for use in some other real-world situations, such as a teaching- learning environment where a teacher could enhance their instruction based on the feedback they receive from using the model, or in psychology where it would aid in the analysis and study of a person's behaviour.

REFERENCES

a) S L Happy and Aurobinda Routray, “Automatic Facial Expression Recognition using Features of salient Facial Patches,” in IEEE Trans. On Affective Computing, January- March 2015, pp. 1-12.

b) Hafeez Kabani, Sharik Khan, Omar Khan and Shabana Tadvi, “Emotion based Music Player,” Int. J. of Eng. Research and General Sci., Vol. 3, Issue 1, pp. 750- 756, January- February 2015.

c) Li Siquan, Zhang Xuanxiong. Research on Facial Expression Recognition Based on Convolutional Neural Networks [J]. Journal of Software, 2018, v.17; No.183 (01): 32-35.

d) Hou Yuqingyang, Quan Jicheng, Wang Hongwei. Overview of the development of deep learning [J]. Ship Electronic Engineering, 2017, 4: 5-9.

e) Liu Sijia, Chen Zhikun, Wang Fubin, et al. Multi-angle face recognition based on convolutional neural network [J]. Journal of North China University of Technology (Natural Science Edition), 2019, 41 (4): 103-108.

f) Li Huihui. Research on facial expression recognition based on cognitive machine learning [D]. Guangzhou: South China University of Technology, 2019.

g) Li Yong, Lin Xiaozhu, Jiang Mengying. Facial expression recognition based on cross- connection LeNet-5 network [J]. Journal of Automation, 2018,44 (1): 176-182.

h) Yao L S, Xu G M, Zhap F. Facial Expression Recognition Based on CNN Local Feature Fusion[J]. Laser and Optoelectronics Progress, 2020, 57(03): 032501.

i) Xie S, Hu H. Facial expression recognition with FRR- CNN [J]. Electronics Letters, 2017, 53 (4): 235-237.

j) Zou Jiancheng, Deng Hao. An automatic facial expression recognition method based on convolutional neural network [J]. Journal of North China University of Technology, 2019,31 (5): 51-56

k) A. Anand, G. Pugalenthi, G. Fogel, and P. Suganthan, “An approach for classification of highly imbalanced data using weighting and undersam□pling,” Amino acids, vol. 39, pp. 1385–91, 11 2010.

l) Wolfram Data Repository, “FER 2013.” [Online]. Available:https://datarepository.wolframcloud.com/res ources/FER-2013

m) S. Minaee and A. Abdolrashidi, “Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network,” ArXiv, vol. abs/1902.01019, 2019