Speech Emotion Detection

Table of Contents

1. Introduction

This is a classification task involving classification of various emotions. For this purpose, the RAVDESS Emotion speech audio dataset is being used. We aim to classify the images using two techniques. The first method involves using only the audio spectrograms. The second method consists of using an ensemble technique by using both audio spectrograms and the face of the person that spoke the words.

The RAVDESS dataset consists of 1440 audios at 16bit and 48kHz in .wav format. The audios are made by 24 different professional actors (12 male, 12 female). There are 7 different emotions in the dataset that include disgust, sad, happy, angry, surprise, fearful, and calm.

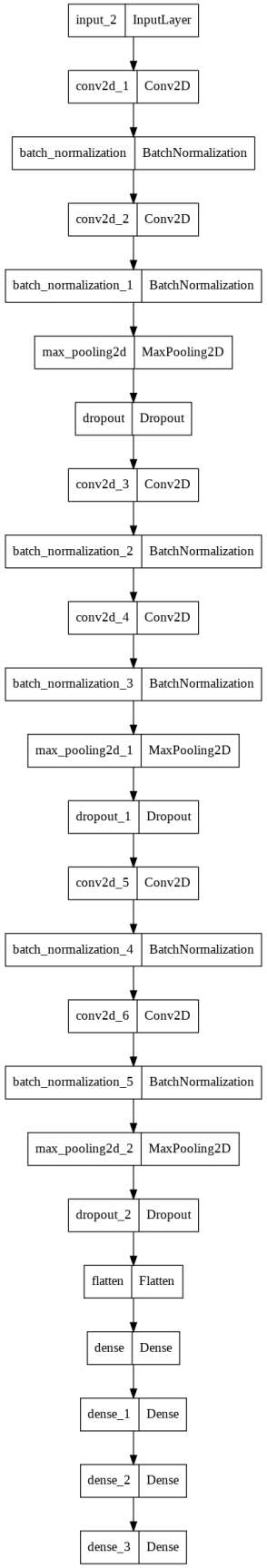
2. Methodology

2.1. Retrieving and preparing the data

The RAVDESS dataset is provided in the form of audio files (.wav file). I have used Librosa package to convert these audios to audio spectrograms and saved them. The generated audio spectrograms have a resolution of 288 x 432. The labels for the audios are read using the file name format that is provided

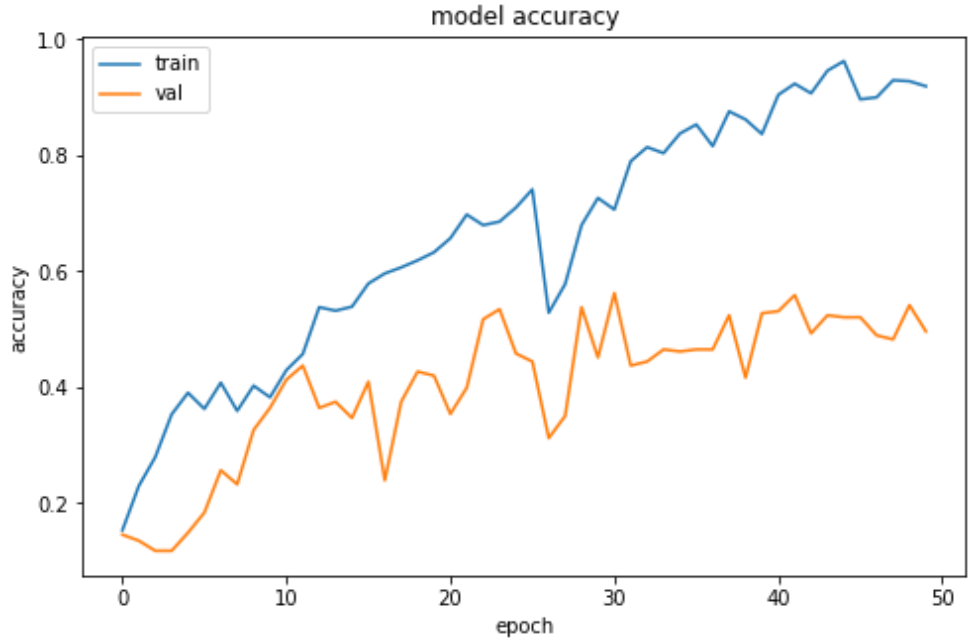
2.2. Defining first model

For creating models, I have used TensorFlow Keras. This model only makes use of the audio spectrograms. This model uses convolution layers, batch normalization, max pooling and dropout. Accuracy is being as the evaluation metric for both training and evaluation. It uses the categorical cross-entropy as the loss function and Adam optimizer. For the output softmax activation function is being used which return the probability of each emotion. The full architecture of the models is as follows:

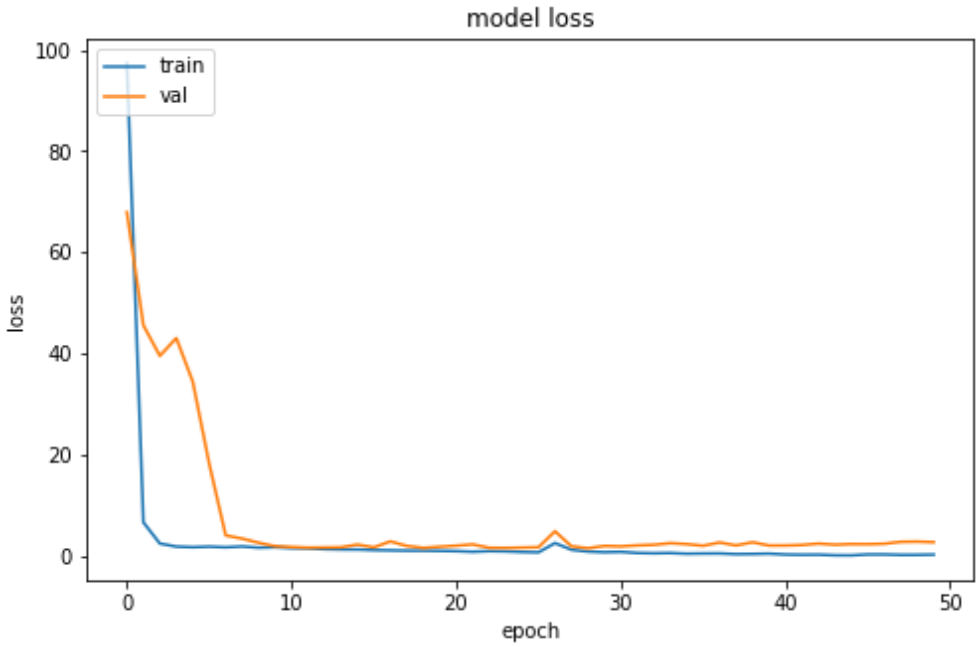

input_2 InputLayer
conv2d_1 Conv2D
batch_normalization BatchNormalization
conv2d_2 Conv2D
batch_normalization_1 BatchNormalization
max_pooling2d MaxPooling2D
dropout Dropout
conv2d_3 Conv2D
batch_normalization_2 BatchNormalization
conv2d_4 Conv2D
batch_normalization_3 BatchNormalization
max_pooling2d_1 MaxPooling2D
dropout_1 Dropout
conv2d_5 Conv2D
batch_normalization_4 BatchNormalization
conv2d_6 Conv2D
batch_normalization_5 BatchNormalization
max_pooling2d_2 MaxPooling2D
dropout_2 Dropout
flatten Flatten
dense Dense
dense_1 Dense
dense_2 Dense
dense_3 Dense

2.3. Results

This model did not give satisfactory results. The model was not able to generalize using only the audio spectrograms. The model was trained with 80% training data and 20% testing data on 50 epochs. However, the model only gave a training accuracy of 91.93% and a validation accuracy of 49.65%.



Even though both training and validation had low loss values the model was overfitting the training even with high dropout being used.



Possible solutions to this include changing the hyperparameters of the model or increasing the dataset size.

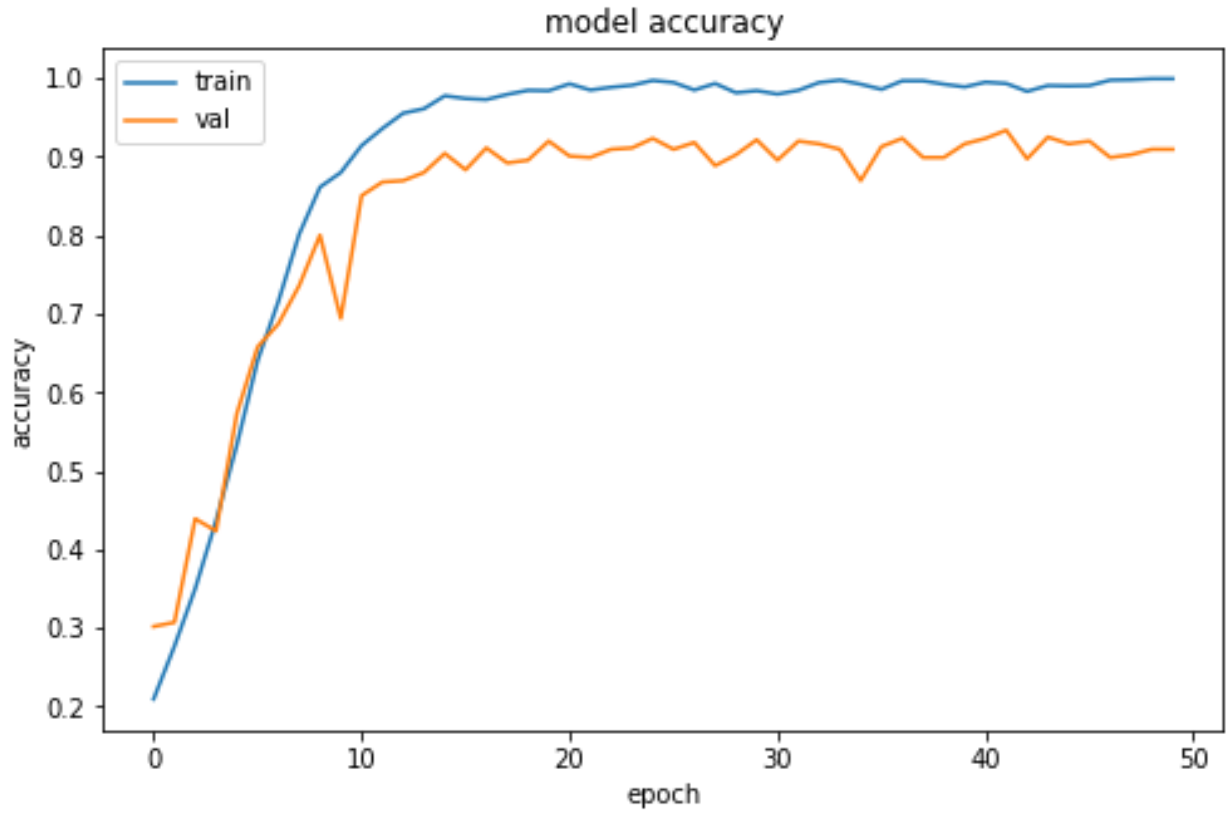
2.4. Defining the second model

This model makes use of both audio spectrograms and actors face pictures. This model uses an ensemble learning technique. The audio spectrograms and facial pictures are processed by two different models and then their outputs are combined and fed into fully connected layers.

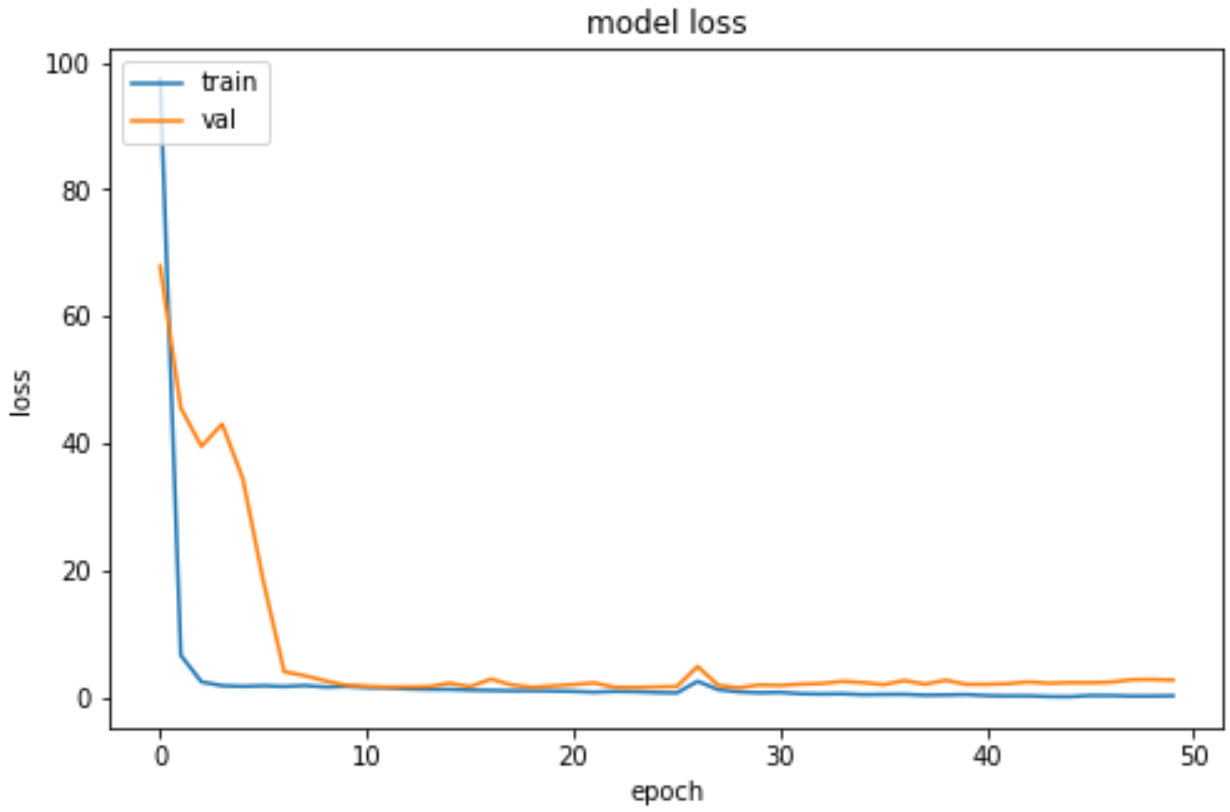
Both models make use of convolution layers and max pooling. Accuracy is being as the evaluation metric for both training and evaluation. It uses the categorical cross-entropy as the loss function and Adam optimizer. For the output SoftMax activation function is being used which return the probability of each emotion.

Results

This model gave good results and was able to generalize on the data. The model was trained with 80% training data and 20% testing data on 50 epochs. The model gave a training accuracy of 99.96% and a validation accuracy of 90.97%.



Both the training accuracy and validation accuracy were almost the same meaning that the model did not overfit the training data. The model loss on training data is also constant near the end meaning the model has reached global minima.


model loss
train
val
loss
epoch
100
80
60
40
20
0
0
10
20
30
40
50