Multimodal Sentiment Classification using Audio and Visual feed

Adithya Kiran
PES1201700231
Computer Science and Engineering
PES University
Bengaluru , India
adithyakiran1999@gmail.com

Chirag P Tubakad
PES1201700896
Computer Science and Engineering
PES University
Bengaluru , India
chirag.tubakad@gmail.com

***Abstract*— Sentiment analysis aims to automatically uncover the underlying attitude that we hold towards an entity. The aggregation of these sentiments over a population represents opinion polling and has numerous applications. Current text-based sentiment analysis relies on the construction of dictionaries and machine learning models that learn sentiment from large text corpora. Sentiment analysis from text is currently widely used for customer satisfaction assessment and brand perception analysis, among others. With the proliferation of social media, multimodal sentiment analysis is set to bring new opportunities with the arrival of complementary data streams for improving and going beyond text-based sentiment analysis. Since sentiment can be detected through affective traces it leaves, such as facial and vocal displays, multimodal sentiment analysis offers promising avenues for analyzing facial and vocal expressions in addition to the transcript or textual content. These approaches leverage emotion recognition and context inference to determine the underlying polarity and scope of an individual's sentiment.**

Keywords—Deep Learning , Tensorflow , Keras , multimodal Sentiment classification , audio features , facial landmarks

I. INTRODUCTION

multimodal sentiment analysis is a new dimension of the traditional text-based sentiment analysis, which goes beyond the analysis of texts, and includes other modalities such as audio and visual data. It can be bimodal, which includes different combinations of two modalities, or trimodel, which incorporates three modalities. With the extensive amount of social media data available online in different forms such as videos and images, the conventional text-based sentiment analysis has evolved into more complex models of multimodal sentiment analysis, which can be applied in the development of virtual assistants, analysis of YouTube movie reviews, analysis of news videos, and emotion recognition (sometimes known as emotion detection) such as depression monitoring, among others.

Similar to the traditional sentiment analysis, one of the most basic tasks in multimodal sentiment analysis is sentiment classification, which classifies different sentiments into categories such as happy , sad , fearful , surprised , angry neutral etc. Feature engineering, which involves the selection of features that are fed into the deep neural networks, plays a key role in the sentiment classification performance. In multimodal sentiment analysis, a combination of different textual, audio, and visual features are employed.

By combining vocal modulations and facial expressions, it is possible to enrich the feature learning process to better understand affective states of opinion holders. In other words, there could be other behavioral cues in vocal and visual modalities that could be leveraged.

The proposed framework considers both facial landmarks mapping as well as the audio cues that are taken into consideration as features in building the model.

II. DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

The set of recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest inter rater reliability were reported.

The dataset we need is a subset of the original dataset and our dataset contained a total of 1440 files , 60 trials per actor X 24 actors. All the video files are recorded in a white background with proper lighting so as to perceive every minute little textural details of the actors.

From these Audio-Video files , we first separated the audio from the video creating a .mp4 and .wav file for each AV file. For extracting features from these files , pythons' dlib library for mapping the facial landmarks and MFCC for the auditory cues was used.

1. Facial Landmarks

   The pre-trained facial landmark detector inside the dlib library is used to estimate the location of 68 (x, y)-coordinates that map to facial structures on the face. These annotations are part of the 68 point iBUG 300-W dataset which the dlib facial landmark predictor was trained on.

2. Mel Frequency Cepstral Coefficients

   In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip. The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound.

Using the above techniques , we were able to create a CSV file for each AV file. Each of these CSV files consisted of about 110-120 rows depending on the number of frames in each video and a total of 157 columns. The 157 columns are broken down as 68 X coordinates and 68 Y coordinates for the facial landmarks totalling 136 features. The remaining 20 features were from the MFCC feature extraction from each of the .wav audio files. The last column was the truth value or the emotion being portrayed by the actor in the video. There were a total of 8 emotions - Neutral , happy , sad , calm , angry , fearful , disgust and surprise.

We created two separate CSV files, one for binary classification i.e classifying just two emotions - happy and sad. The second CSV is a CSV formed by merging all the individual CSV files created and for all the emotions. These two CSV files formed the dataset for the models that we proceeded to build.

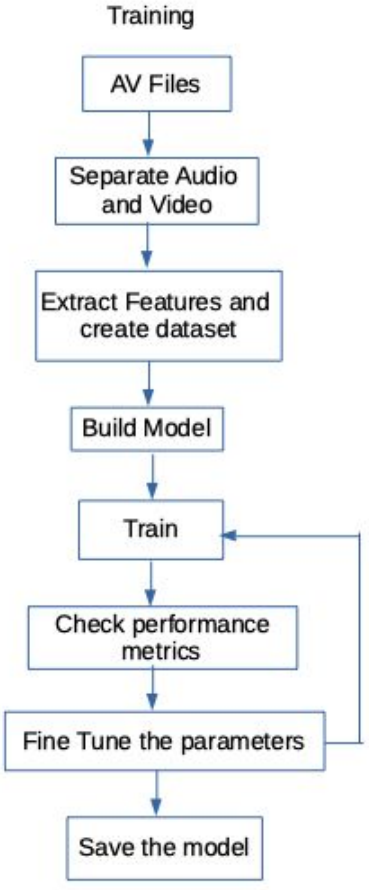

Fig 1

III. MODELS

To showcase versatility we built 4 models , two models for binary classification of 2 emotions amongst which one was in keras and one in tensorflow. The remaining two models were for multiclass classification of 8 emotions amongst which, again, was one in keras and one in tensorflow. The general workflow for all the models remains the same as depicted in Fig 1. Only the deep neural network configurations for each model changes.

A. Keras Binary Emotion Classification Model

We first started off by building the keras model for binary classification of emotions. This model consisted of an input dimension of 156 which is mostly consistent across all the models. It has 3 hidden layers with 4 neurons in each layer and all have ReLu as the activation function. The output layer has 1 neuron with a sigmoid activation unit. The loss or the cost function that is being used is the binary cross entropy coupled with the Adam optimizer. The model was run for 30 epochs with a batch size of 16 and the model gave a training accuracy of 98.81% with a loss of 0.0479. Over the testing set of the data , the model predicted with an accuracy of 98.73%. The plot for the loss and accuracy during training can be found in Fig 2.

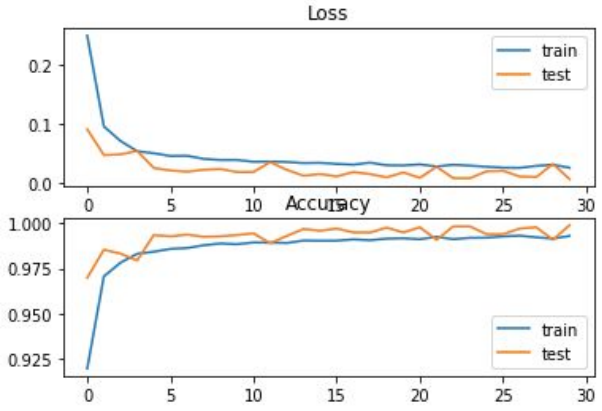

Fig 2

B. Tensorflow Binary Emotion Classification Model

Our second model for the same problem of binary classification of two emotions i.e happy and sad was built using tensorflow. This tensorflow model has the input layer with 156 neurons for each of the 156 input features, followed by 3 hidden layers with 8,8 and 4 neurons each and an output layer with 2 neurons. Each of the 3 hidden layers use ReLu as the activation function. The output layer uses a softmax activation function for classification. The loss function used was the binary cross entropy function. This was coupled with the Adam Optimizer for gradient descent with backpropagation with a learning rate of 0.01. The plot for accuracy during training can be found in Image 4. The train test split was 70-30 and we used a drop out of 20%. The accuracy for this model on the validation (test) set was around 99%. The accuracy and loss during training can be seen in the Fig 4 present below.

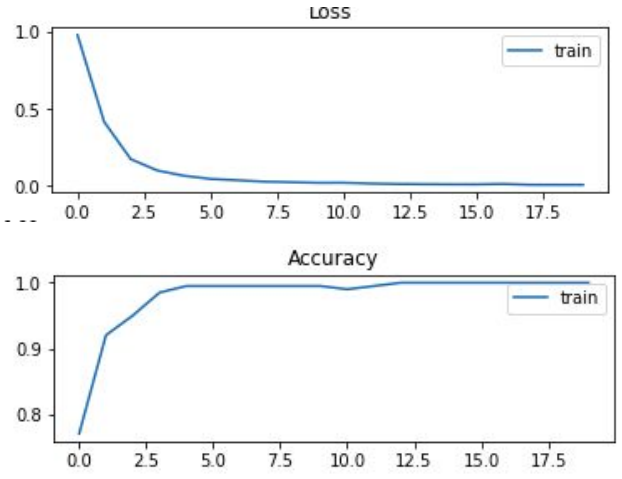

Fig 3

C. Keras 8 Emotion Classification Model

Our third model was a keras model built for the classification of all the 8 emotions. This keras model as usual consisted of an input dimension of 156. The first hidden layer consisted of 32 neurons all activated with a ReLu activation unit. The second hidden layer consisted of 32 neurons with ReLu as the activation unit. The third hidden layer consisted of 16 neurons with ReLu as the activation function. The output layer consists of 8 neurons with softmax as the activation unit as it is the best for multiclass classification and converts all the values into a probability distribution, so we can select the one with the highest probability as our output. The dataset for this model was split 90:10 for the training set and the testing set. The model was trained for a total of 100 epochs with a batch size of 200. For this model, we used the categorical cross entropy loss function as the target columns were one-hot encoded coupled with the Adam optimizer to obtain the best result. This model gave an accuracy of 92.6% over the training set an accuracy of 92.2% over the testing set. The model's loss and accuracy during training can be seen in Fig 4.

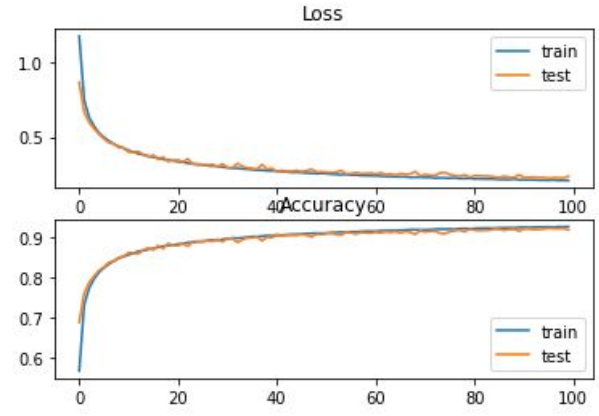

Fig 4

D. Tensorflow 8 Emotion Classification Model

Our fourth model was a tensorflow deep neural network for classifying all 8 emotions available within our dataset. This model is an exact replica of (C.) which was done using keras. The input has 156 neurons, followed by the first hidden layer having 64 neurons, followed by the second hidden layer with 32 neurons and activated with ReLu as their activation functions. The third hidden layer had 16 neurons, with ReLu as the activation function. Finally, our output layer has 8 neurons. The depiction of our neural network can be seen in Fig 5.

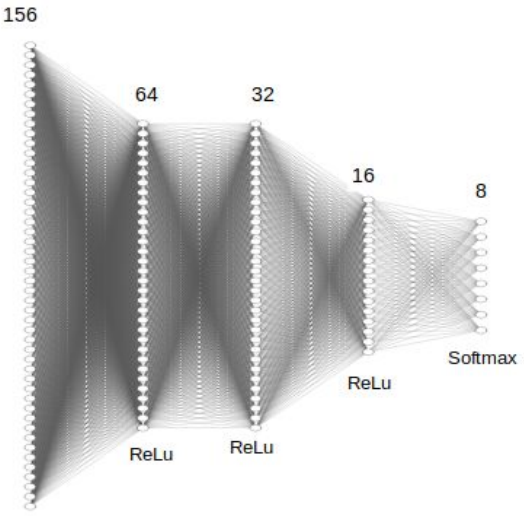

Fig 5

The loss function is softmax cross entropy with logits, which applies softmax activation to our output layer before calculating categorical cross entropy for the loss. This loss is optimized by the Adam optimizer with a learning rate of 0.001. The train-test split was 70-30 and we utilised a keep probability of 80% to prevent overfitting our model. The accuracy over the validation (test) set was 95%.

This matches our keras models accuracy. The loss and accuracy during training over 150 epochs can be seen below in Fig 6.

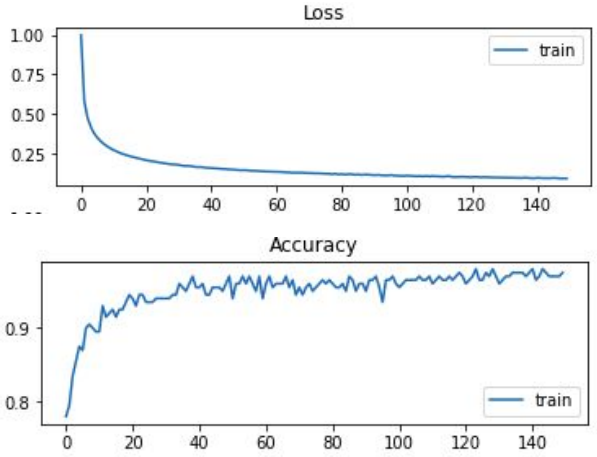

Fig 6

IV. Model Parameters

A. Activation Functions

We have used Rectified Linear Unit (ReLu) activation functions in all our models. The ReLu function is less susceptible to problems such as vanishing gradients, when compared to other activation functions such as sigmoid and tanh. The ReLu function takes the form:

$f(x) = max(0, x)$ $f'(x) = \{ 0 \text{ if } x <= 0 ; 1 \text{ if } x>0$

This function also has the advantage of getting trained quickly when compared to functions such as sigmoid as it does not require much computation. It provides a sparse representation during computations, allowing it to be trained faster. ReLu is however capable of causing some neurons to never get activated, and this is called dying ReLu, which can be overcome using Leaky ReLu. However, we tested our models with different activation functions, and ReLu performed the best, with marginally different accuracies when compared to activations with Leaky ReLu and tanh. Sigmoid as well gave us 90% accuracy, but ReLu proved best with 92%.

For the output classification layer, we apply the softmax function.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

This gives us a probability distribution for all the 8 emotions, and we can easily select the emotion with the highest probability as our prediction.

B. Loss Function

For binary classification, we utilised the binary cross entropy loss function which provided us favourable results. The binary cross entropy function takes the form :

$$-(y \log(p) + (1-y) \log(1-p))$$

Since binary cross entropy worked well, we utilised categorical cross entropy for our eight emotion classification model. This function takes the form :

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

C. Hyperparameters

We utilise a learning rate of 0.001 with our model, which is the recommended/default rate used with adam optimizers.

For 8 emotion classification, using 64, 32 and 16 neurons in the hidden layers proved most effective. Adding more neurons did not provide us with any considerable improvements.

The number of training epochs, was again obtained by trial by testing different numbers, and increasing it beyond 150 epochs did not provide any significant improvement to the models.

V. Conclusion And Future Work

While all the models we built give some exceptionally good results and are very much viable to be used at a production level , the problem arises when the input scales exponentially and handling the workload at that level would pose a great threat. All the current models, given the processing time, one of the bottlenecks would be the underlying hardware used to run the algorithms. We are processing every video per frame and extracting values from each and every one of them and performing computations on the same is a very computationally expensive task. These computations take a monumental amount of time if they have to be performed just on a CPU. Moreover , some of the tensors we are performing operations on are so big to even fit in the memory of an edge system. It is quintessential that all these computations be done with GPU support and on a system with sufficient memory. Hardware constraints is always a factor that needs to be taken into consideration when building and scaling Deep Learning Models.

One of the aspects that could possibly be taken into considerations and worked on in the future would be incorporating LSTM in our current solution framework. All the current models process the features extracted from the video files and return the best estimate of the emotion being portrayed in the video. Looking ahead , we could include another dimensionality by giving a relative dependence between successive frames. At the starting of every video , the actor or the subject of the video when looking into the camera , the first few seconds the features extracted are mostly the same for any video irrespective of the true emotion class because they are just looking into the camera and getting started . We wouldn't be able to predict with current models if we were using still images and all the models would fail. That would be something exciting to work on and make our models as robust as possible.

Finally, our current model has been trained only on two statements, "Kids are talking by the door" and "Dogs are sitting by the door". It would be better to train the model on a larger variety of videos with a much better vocabulary. Our current model serves as a proof of concept, that the method works, and that working with a more diverse dataset, would surely allow us to make huge strides in the field of sentiment analysis via audio and visual feeds.

VI. Acknowledgement

We would like to specifically thank Professor Srinivas K S for guiding us, believing in us and bestowing us with this wonderful opportunity to showcase our skills and in the process giving us hands-on experience in one of the most trending domains in the field of computer science. It has been a great learning curve and absolute fun to dabble in the field of Deep Learning.

References

[1] *Ameya Rajendra Bhamare, Srinivas Kathargуppe, Silviya Nancy J, "Deep Neural Networks for Lie Detection with Attention on Bio-signals"*

[2] *Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391*

[3] *Mohammad Soleymania , David Garciab , Brendan Jou , Björn Schuller , Shih-Fu Chang , Maja Pantic , "A survey of multimodal sentiment analysis"*

[4] https://medium.com/@himanshuxd/activation-functions-sigmoid-relu-leaky-relu-and-softmax-basics-for-neural-networks-and-deep-8d9c70eed91e

[5] https://en.wikipedia.org/wiki/Multimodal_sentiment_analysis

[6] https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html#cross-entropy