Loop Q PRIZE 2022 - EMOTION DETECTION FROM SPEECH

Davide Baldelli, davide.baldelli4@studio.unibo.it

Abstract

Speech emotion recognition continues to be a difficult task. There are still several open problems: which are the best input features, and which is the most effective neural architecture. I have adopted a combination of input features, that include Mel spectrogram, Mel-frequency cepstral coefficients (MFCCs), chromagram, spectral contrast and Tonnetz representation. I propose an architecture based on bidirectional long-short term memory (LSTM) layers, that fully exploit the temporal information of audio recordings. I have trained the network on audio files from four different origins: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Crowd-sources Emotional Multimodal Actors Dataset (CREMA), Surrey Audio-Visual Expressed Emotion (SAVEE), Toronto emotional speech set (TESS).

1. Introduction

In this work I propose an approach to speech-based emotion recognition system. The analyses were carried out on audio recordings from four different datasets: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Crowd-sources Emotional Multimodal Actors Dataset (CREMA), Surrey Audio-Visual Expressed Emotion (SAVEE), Toronto emotional speech set (TESS).

I have extracted five different features from the audio files: Mel Spectrogram, MFCCs (Mel-frequency cepstral coefficients), chromagram, spectral contrast, and Tonnetz representation, as suggested in [1].

My model architecture consists of five parallel networks, one for each feature. Those networks are a combination of 1-dimensional convolutional layers and bidirectional LSTM layers. The LSTM layers are particularly effective for the analysis of temporal data and are largely adopted in audio classification tasks [2]. The five networks are then merged to produce the class label.

The project has been developed using Tensorflow and Keras. To extract the features from the audio files I have used the package librosa. I have also exploited scikit-learn for the data preprocessing.

2. Data

The merged dataset is composed of 10109 recordings: 6305 from CREMA, 1041 from RAVDESS, 397 from SAVEE, 2366 from TESS. The emotion labels are not equally distributed among the different origin, especially because the dataset CREMA does not contain any recording labelled as surprise (Figure 1).

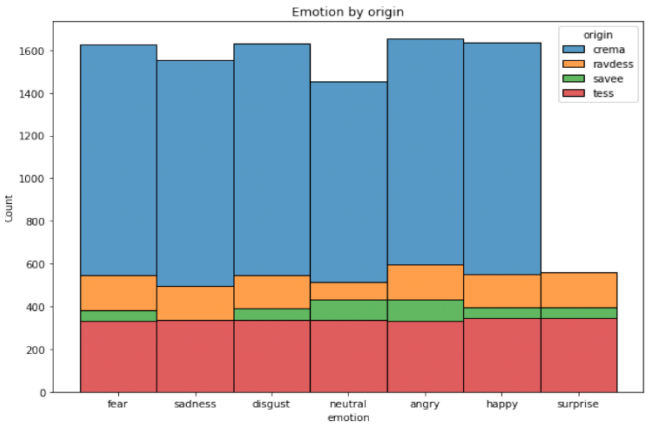

Figure 1. Histogram of the distribution of emotion by origin.

The original recordings have different sampling rates: RAVDESS 48KHz, CREMA 16KHz, SAVEE 44.1KHz, TESS 24.414KHz. Therefore, I have resampled the audio files to 16KHz. Regarding the length of the recordings, the maximum duration is 7.14 s, but 96% of the recordings are shorter then 4s. Consequently, I have decided to pad or cut every recording to fit that duration. Instead

of padding the audios shorter then 4s with silence I have repeated the recording until they reached the desired length.
I have augmented the dataset by adding random noise and random time shifting. For each recording I have created two augmented copies so to have a final dataset three times bigger than the original.

I have split the dataset into train and test with 20% of test size.

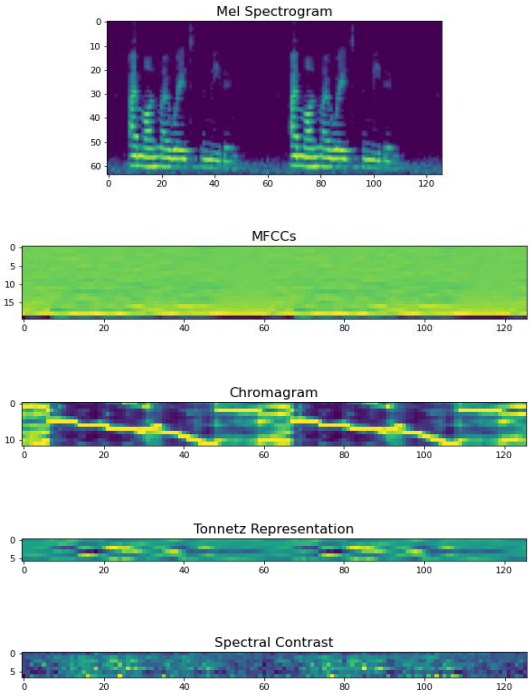

Figure 2. Model inputs.

Once the audio files have been normalized to the same sampling rate and duration, I have extracted the features. I have used a length of the fast Fourier transform (fft) window of 1024, that, with a sampling rate of 16KHz, corresponds to 0.064s, and a number of samples between successive frames of 512, that means the half of the window size. To compute the Mel Spectrogram, I have set the number of Mel bands to 64.

Finally, the processed features are 2-dimensional arrays that share one dimension, the timestep dimension, that is 126.

I have one hot encoded the target labels.

3. Model architecture

I have designed the same architecture for each feature. It consists of two 1D convolutional block, two Bidirectional LSTM layers and two dense layers. The convolutional blocks extract low level features and reduce the size of the time dimension.

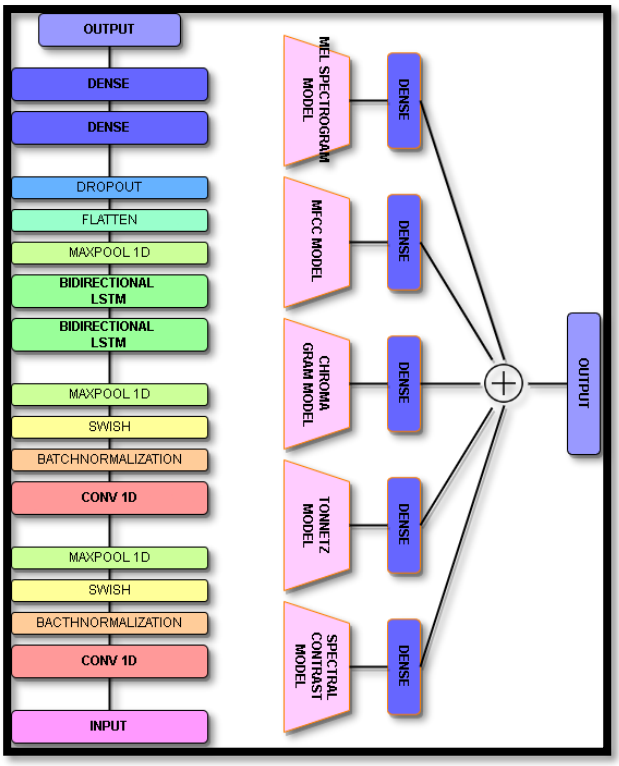

Figure 3. On the left, architecture of the neural network for each feature, on the right, ensemble architecture.

Each 1D convolutional layer has 32 channels and kernel size of 5 in the time dimension, and it is followed by a Batch Normalization layer, a swish activation function and a 1D Max Pooling layer, whose pool size is 2. Each LSTM layer has 32 units per direction, dropout, and recurrent dropout of 0.3 and both return the full sequence. The last LSTM layer is followed by a 1D Max Pooling layer, after whom the output is flattened. The top of the model consists of a dropout layer and two dense layers with 256 neurons each and swish activation function. The classification layer has dimension 7, the number of different emotions.

Instead of computing an average of the different predictors' outputs, I have merged the models as follows: I have removed the last layer from each network that would result in a bottleneck; I have stacked a 64-dimensional Dense layer to each model and summed up the outputs; I have added the classification layer on top of the network. During the training, the pretrained models' weights are kept frozen. In this way the top of the model learns how to extract information from the 256-dimensional internal representations of each model.

4. Results

I have reached 0.726 of accuracy and 0.745 of F1-Score in the test set, as shown in figure 4.

The quality of the performance depends both on the origin of the file audio and on the emotion. Indeed, we can see that the recordings that express surprise are more easily detected, while the ones that express fear are the most difficult to classify.

Regarding the origins of the files, I have obtained the results shown in the Figure 5.

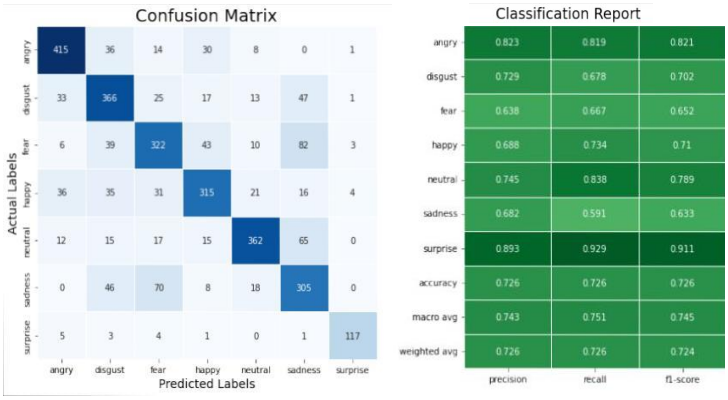

Figure 4. Evaluation of the proposed model in the merged dataset.

Figure 5. Evaluation of the proposed model on the different datasets.

5. Conclusion

In this work I propose an ensemble neural network architecture based on bidirectional LSTM layers.

The fact that the performance of the model depends strongly on the origin of the data could be an indicator that none of the proposed datasets is enough large and heterogeneous to train a model effectively for real-world applications.

Therefore, I suspect that the main limitation of my work is the quality and the size of data. Such a hard task could require a huge amount of more complex and varied recordings.

Without further testing and the introduction of additional data, I cannot conclude that the results will extrapolate to all other speech data.

Nonetheless, given the complexity of the task, the results are promising: I believe that extracting several input features and the employment of LSTM architecture can be the base ideas for more performing models.

6. References

[1] Dias Issa, M Fatih Demirci, and Adnan Yazici. “Speech emotion recognition with deep convolutional neural networks.” Biomedical Signal Processing and Control, 59:101894, 2020.

[2] Scarpiniti, Michele, et al. “Deep recurrent neural networks for audio classification in construction sites.” *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.