Performance Comparison of LSTM Models for Speech Emotion Recognition

Tanuhree Swain, Utkarsh Anand, Yashaswi Aryan, Soumya Khanra, Abhishek Raj

Abstract-Speech Emotion Recognition can be essentially categorized into a sequence-related task and the performance of LSTM models is an excellent benchmark for inspecting. We attempt to assess the performance of the combined Convolutional Neural Network and Long Short Term Memory and Long Term Memory only for recognition of emotions from a speech corpus.We propose some modifications to the existing CNN-LSTM based methods that rely on feature extraction to achieve better performance. We perform the analysis on the Ryerson Audio-Visual Database of Emotional Speech and Song. After pre-processing raw audio files, the Mel-Frequency Cepstral Coefficients feature was considered. Data augmentation techniques approaches are also used to boost the performance of the approaches.

Keywords:Speech Emotion Recognition, Convolutional Neural Network ,Long Short Term Memory,Mel-Frequency Cepstral Coefficients

INTRODUCTION

As virtual agents begin to permeate our daily lives, there is a raising need for intelligent human interactive computer systems. And although some of these systems have mastered the art of speech recognition and transcription, they still lack human-like aspects of the conversation. In conversation, apart from the textual content, people receive a lot of specific information, such as diction and emotion, loudness, tones, etc., that attributes to the true meaning of a word. When systems such as voice assistants are armed with such information, they can contribute to a more emotionally sensitive dialogue. Speech Emotion Recognition(SER) is a challenging problem because people articulate emotions in different ways and auditory variation due to specific words, voice rhythm, dialect, and speaking volume, influence the interpretation of the underlying emotions. Extracting low-level descriptors and training the machine correctly by learning certain features are conventional techniques for solving this problem. The selection of the most optimal feature set for the representation of emotional responses is still a matter of ongoing research. As a result, the traditional trend in speech/audio information retrieval is to focus on the use of powerful semantic analysis strategies, often relying on model selection to optimize results.

Unparalleled success of deep learning in various domains has led to the emergence of neural network architectures as successful approaches across many tasks due to their ability to share low-level representations and naturally progress from low-level to high-level structures. Neural network architectures like Convolutional Neural Networks show remarkable performance for learning representations from spatial data like images , Recurrent frameworks such as Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) have demonstrated state of the art results in tasks where there is a need to take note of the of the interdependency of input data. They use their internal memory to compute arbitrary sequences such as text and speech.

In this work, we seek to evaluate the performance of LSTM and CNN-LSTM architectures for speech emotion recognition.We also learn about how data augmentation techniques in audio can help enhance the performance of these architectures.

RELATED WORK

The earliest work on SER focused on Hidden Markov Models, Gaussian Mixture Models, Support Vector Machines, Decision trees, and K Nearest Neighbors. Some other notable systems suggested by researchers were the modified brain emotional learning model (Sara et al., 2017), a multiple kernel Gaussian process classification (Chen and Jin, 2015) and the Voiced Segment Selection (VSS) algorithm (Yu et al . , 2016)

In 2014, Han et al. presented one of the first deep learning frameworks for SER. Their approach was to divide each phrase into frames and calculate the low-level features in the first step and then a densely connected neural network is used to give the results. In the continuation of his work, Lee et al and Tashev et al substituted the neural network with RNN and LSTM and the input was a 32-dimensional vector consisting of pitch, voice probability, zero-crossing rate, MFCC with log energy and first-time derivatives of each frame. They used the probabilistic approach to learning and the aggregation of local probabilities into a global feature vector and ELM at the top. Chernykh et al's work uses the same probabilistic approach by using the Connectionist Temporal Classification loss function which allows

considering long utterances containing both emotional and neutral parts.CNN-LSTM hybrid architectures are also being studied in SER. The ability of CNN to extract high-level features, and the ability of LSTM to model long-term contextual dependencies, can be effective classification methods.Trigeorgis el al. used raw signals as input to a combination CNN and Bidirectional LSTM system to allow the model to acquire an intermediate representation of the raw input signaby omitting the feature extraction stage. Etienne et al. used a combination of CNN and BiLSTM with Log spectrograms as input data with data augmentation techniques such as vocal tract length perbutation.

OVERVIEW OF RECURRENT ARCHITECTURES

The central principle of RNN is based around the usage of sequential data. Aptly named 'recurrent' beacuse they execute the same process on every element in the series with output being reliant on previous computations. Another way to talk of RNNs is that they provide a ‘hidden unit’ that collects knowledge on what has been measured so far.They are networks with loops that allow information to persist.

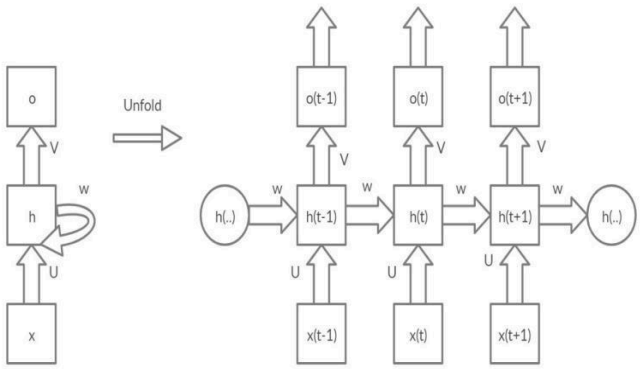

Fig. 1. A Recurrent Neural Network structure as it unfolds in time with its forward computation

Figure 1 is the description a basic RNN structure that unfolds in time.Here t denotes the time step and x(t) is the input. The hidden state h(t) acts as memory of the network.The value of h(t) is formulated by the current input and previous time step’s hidden state[11].

$$a(t) = U*x(t) + W*h(t-1) + b$$

$$h(t) = \tanh(a(t))$$

$$o(t) = V*h(t)$$

U,V,W are the weight parameters of the hidden layer, output layer and hidden state which are shared throughout the network.o(t) is the output of the network[12].

LSTM network are an upgrade from traditional RNN structures which use ‘memory cell’ instead of a hidden unit to learn long term dependencies which the RNN is incapable of. This cell consists of input(i_t),forget(f_t) and output(o_t) gates which regulate the information in the memory cell and help the network selectively remember and forget things.

$$f_t = \sigma (W_f[h_{t-1},x_t] + b_f)$$

$$i_t = \sigma (W_i[h_{t-1},x_t] + b_i)$$

$$o_t = \sigma (Wo[h_{t-1},x_t] + b_o)$$

W_x represents the weights and b_x are the bias terms of the respective gates(x).h_{t-1} is the value of the previous lstm cell.x_t is the input to the current timestep.

$$\hat{C}_t = \tanh(w_c[h_{t-1},x_t] + b_c)$$

$$C_t = f_t * C_{t-1} + (1-f_t) * \hat{C}_t$$

$$h_t = o_t * \tanh(c_t)$$

C_t is the current memory state at timestep t and $\hat{C}_t$ is candidate for cell state at t.Figure 2 denotes a single LSTM memory cell.

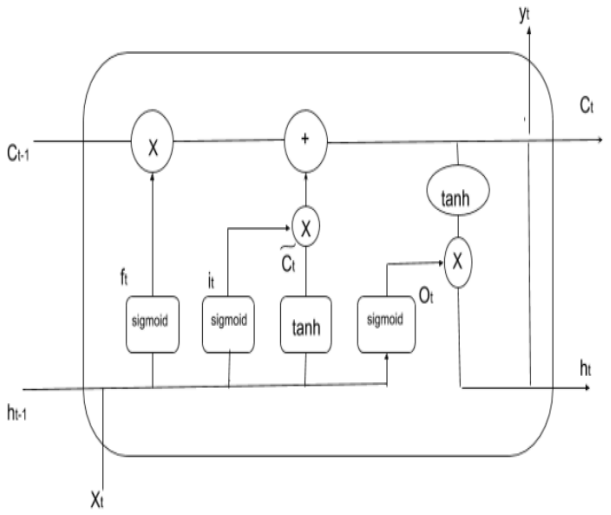

Fig. 2. A single Long Short Term Memory Block.

PROPOSED FRAMEWORK

Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is utilized for training both the models[14] . RAVDESS comprises of eight emotions (anger, calm, disgust, fear, happiness, sadness, surprise, neutral), all of which apart from neutral are expressed in normal and strong two intensities . There are 12 female speakers and 12 male speakers with 60 speech files each (4 files for each emotionintensity pair and 4 files for neutral). Speech files have annotated emotion and intensity labels, are less than fifteen seconds long, and consist of a single sentence expressing a single emotional state and read in North American English. All classes are balanced.

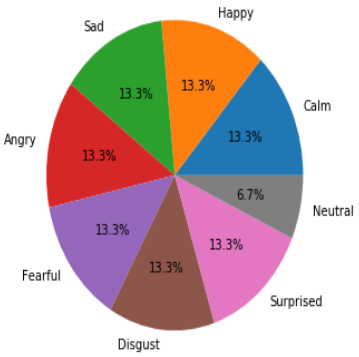

Fig. 3. The class distribution of emotions in RAVDESS dataset

Speech Representation

The first step in SER pipeline is the extraction of features from the speech emotion corpus. Mel Frequency Cepstrum Coefficient(MFCC) is one of the most popular acoustic characteristics for speech representation for they are modeled after the human auditory signal[15]. MFCC imitates the human auditory experience taking in a few factors- the human auditory perception does not obey a linear form; each tone has an real frequency determined by 'hertz;' each to ne has a subjective frequency 'pitch' calculated by a metric na med the mel scale, the main purpose of the subjective frequency is to capture the significant phonetic characteristics and the frequency scale of mel is linear below 1000 Hz and the logarithmic above 1000 Hz. (Tiwari, 2005; Shaneh and Taheri, 2009; Muda et al., 2010; Thakur and Sahayam,2013). The following flowchart describes the the process in which MFCC is extracted in the following manner-

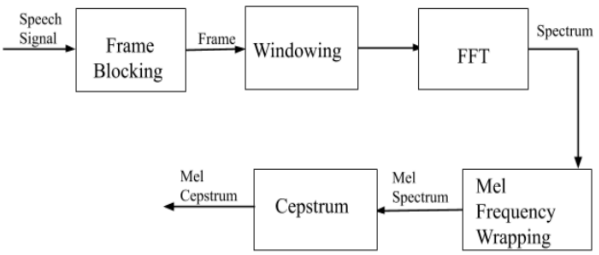

Fig. 4. A flowchart denoting the process of MFCC extraction

Data Augmentation

Audio based models often encounter the issue of overfitting likely due to less size of training data which can hinder the overall performance of the model. A way to resolve this is by increasing the size of the training data with the help of data augmentation. Corrupting the speech signal with external factors like noise help the model generalize in a better manner. The methods that we adopt for augmentation of the speech signals are as follows-

White Noise
It is generated by the following method-
Noise amplitude=0.005*random samples from uniform distribution*maximum value of the input data
Data = data + Noise amplitude * random sample from normal distribution

Random Shifting
It is shifting the input data over a random distribution of samples.The limit of distribution is assumed to between -5*500 to 5*500.

Stretching the sound
The signal is stretched at the rate of 0.8

Pitch tuning
Shifts the pitch of the waveform by n_semitones
The value of n_semitones here is assumed as 2

Random Value Change
We multiply the signal with any value from a distribution of random samples

Model Architectures

We use the first 40 MFCC coefficients as a input and perform the following data augmentations on the speech signal aas described above.Before computing the MFCC's we trim the speech to remove the silent parts.CNN-LSTM 's architecture consists of 4 blocks that learn from the input local features and an LSTM cell for long-term dependency computing.

Effectiveness of this (4 + 1) paradigm has been confirmed in several works[7][8][10].

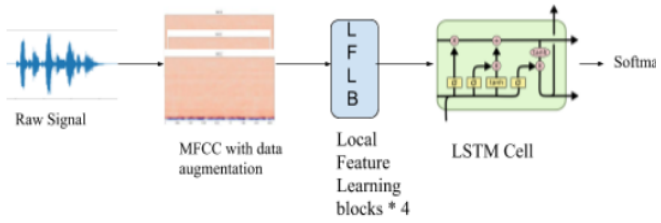

Fig. 5. Block diagram of the CNN-LSTM architecture.

The local feature learning blocks consist of a convolution layer, a normalization layer and a activation function is applied followed by pooling and dropout layers.We use 'Adam' as an optimizer with a learning rate of 1e-3 and a batch size of 64.The activation function used is Exponential Linear Unit(ELU).Unlike ReLu they have negative activation and it pushes the mean of the activations closer to zero. Having mean activations closer to zero also causes the faster learning and convergence.It performs identify operation on non negative inputs and exponential nonlinearity on negative inputs[].The alpha parameter is used to scale the negative.

$$\begin{array}{ll} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{array}$$

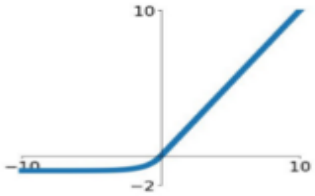

Fig. 6. Graph showing ELU activation

The LSTM architecture is simply formed by three LSTM cells followed by a Flatten layer. Both the architecures have a output layer with fully connected softmax activation that gives us the prediction of the emotion label of the speech.

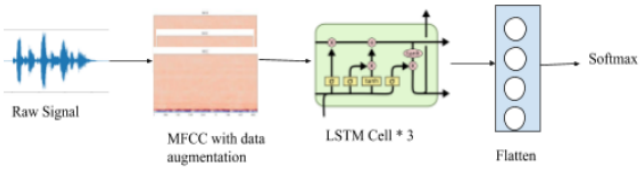

Fig. 7. Block diagram of the LSTM model

RESULTS

We analyze the performance of both of these models by computing the confusion matrix. The gender of the speakers is also taken into account in both of the proposed methods. So we get 16 classes of prediction.The CNN-LSTM archirecture gives us an accuracy of 95% on the test data whereas the LSTM architecture gives us an accuracy of 74% on the test data.

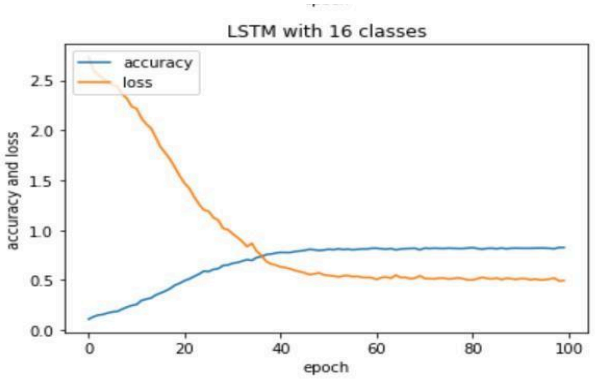

Fig. 8. Loss and accuracy plot of the LSTM model

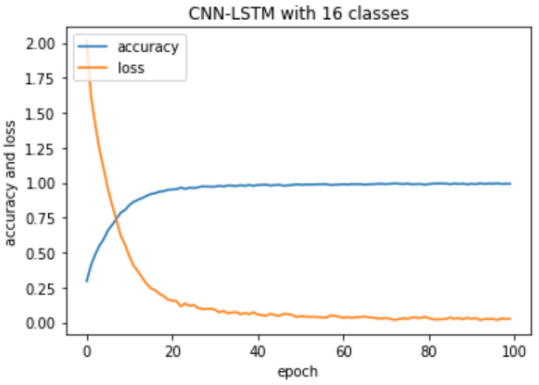

Fig. 9. Loss and accuracy plot of CNN-LSTM

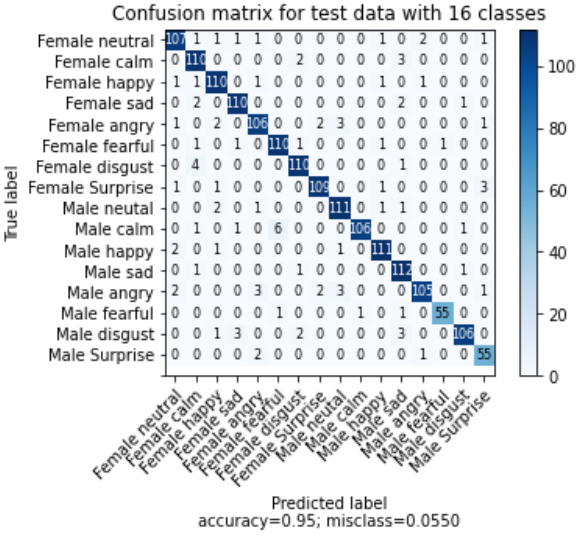

Fig. 10. Confusion matrix of CNN-LSTM

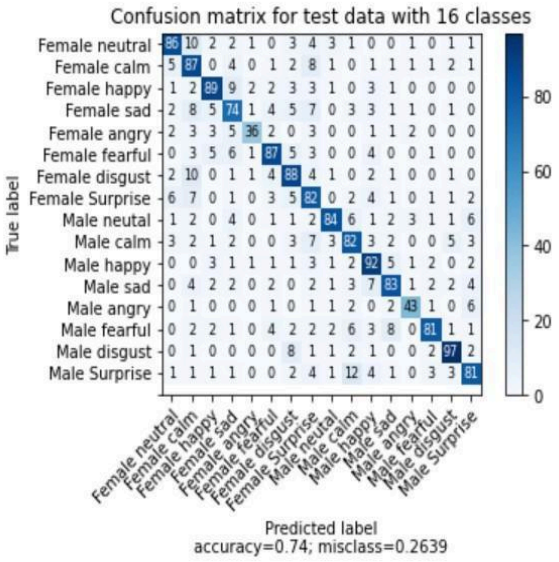

Fig. 11. Confusion matrix of LSTM

CONCLUSION

We observe that out of the two models CNN-LSTM performs better than the LSTM. The combined attributes of the CNN-LSTM architecture help better in classifying the speech signals into their respective gender-dependent emotion labels. We also discussed about the various data augmentation techniques to improve the performance of these algorithms. We can conclude that CNN based models provide us with better results. In the future, we seek to modify this existing architecture to create an end to end pipelines that requires less preprocessing of the input and provide us with better performance.

REFERENCE

Banse, Rainer & Scherer, Klaus. (1996). Acoustic Profiles in Vocal Emotion Expression. Journal of personality and social psychology. 70. 614-36. 10.1037/0022-3514.70.3.614.

Cowie, Roddy & Douglas-Cowie, Ellen & Tsapatsoulis, Nicolas & Votsis, George & Kollias, Stefanos & Fellenz, Winfried & Taylor, J.G.. (2001). Emotion recognition in human-computer interaction. Signal Processing Magazine, IEEE. 18. 32 - 80. 10.1109/79.911197.

Daantje Derks, Agneta H.Fischer, Arjan E.R.Bos. (2007). The role of emotion in computer-mediated communication: A review

K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In Interspeech 2014, 2014.

J. Lee and I. Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In Interspeech 2015, 2015

V. Chernykh, G. Sterling, and P. Prihodko. Emotion recognition from speech with recurrent neural networks. ArXiv e-prints, 2017.

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., Schmauch, B. (2018) CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. Proc. Workshop on Speech, Music and Mind 2018, 21-25, DOI: 10.21437/SMM.2018-5.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou , Björn Schuller1,Stefanos Zafeiriou(2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network

Zazo, R., Lozano-Diez, A., Gonzalez-Rodriguez, J. (2016) Evaluation of an LSTM-RNN System in Different NIST Language Recognition Frameworks. Proc. Odyssey 2016, 231-236.

Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and Recurrent Neural Networks. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). doi:10.1109/apsipa.2016.7820699

https://towardsdatascience.com/recurrent-neural-networks-rnns-3f06d7653a85

Bao, Wei & Yue, Jun & Rao, Yulei. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE. 12. 10.1371/journal.pone.0180944.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60–68. doi:10.1016/j.neunet.2017.02.013

Livingstone SR, Russo FA. 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. 2018;13(5):e0196391. Published 2018 May 16. doi:10.1371/journal.pone.0196391

Muaidi, Hasan et al. "Arabic Audio News Retrieval System Using Dependent Speaker Mode, Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques." Research Journal of Applied Sciences, Engineering and Technology 7 (2014): 5082-5097.

Tiwari, V., 2005. MFCC and its applications in speaker recognition. Int. J. Emerg. Technol., 1(1): 19-22.

Tiwari, Vibha. (2010). MFCC and its applications in speaker recognition. Int. J. Emerg. Technol.. 1.

Muaidi, Hasan et al. "Arabic Audio News Retrieval System Using Dependent Speaker Mode, Mel Frequency Cepstral Coefficient and Dynamic Time Warping Techniques." Research Journal of Applied Sciences, Engineering and Technology 7 (2014): 5082-5097.

Shaneh, M. and A. Taheri, 2009. Voice command recognition system based on MFCC and VQ algorithms. World Acad. Sci. Eng. Technol., 33: 534-538.

Thakur, A. and N. Sahayam, 2013. Speech recognition using Euclidean distance. Int. J. Emerg. Technol. Adv. Eng., 3(3).

Muda, L., M. Begam and I. Elamvazuthi, 2010. Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. J. Comput., 2(3): 138-143.

Liu, Gabrielle. (2018). Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech.

Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," CoRR, abs/1412.5567, 2014. [2] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust asr," in ASRU, 2009, pp. 205– 210.

Ko, Tom et al. "Audio augmentation for speech recognition." *INTERSPEECH* (2015).

Nguyen, Thai-Son & Stuker, Sebastian & Niehues, Jan & Waibel, Alex. (2020). Improving Sequence-To-Sequence Speech Recognition Training with On-The-Fly Data Augmentation. 7689-7693. 10.1109/ICASSP40776.2020.9054130

https://numpy.org/doc/.

https://ml-cheatsheet.readthedocs.io/en/latest/activation_functions.html

Clevert, Djork-Arné & Unterthiner, Thomas & Hochreiter, Sepp. (2016). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs).